Phase Space Attention: A Hairer Lift Circumvents the Single-Layer Induction Obstruction

Kingsuk Maitra **Shagun Sood** **Morteza Hosseini** **Suman Gunnala** **Vikram Gupta**
Qualcomm Cloud AI Division, Santa Clara, CA, USA
kmaitra@qti.qualcomm.com

Abstract

Sanford, Hsu and Telgarsky (SHT) prove that a single attention layer cannot solve the in-context induction task unless its per-head description size grows with the context length T. We *circumvent* this obstruction—exiting the premise of the SHT counting argument, not the bound itself—by lifting the post-RoPE query and key streams onto a phase space, in direct correspondence with Hairer's lift of Störmer–Verlet onto a one-step symplectic integrator. The reframing exhibits the obstruction as a *filter-order gap*: a standard bilinear score has joint z-transform order $(0, 0)$, whereas the induction discriminator needs key-side order ≥ 1. Phase Space Attention (PSA) closes the gap with the symplectic upper shear $M_\gamma : (q, p) \mapsto (q + \gamma p, p)$, p the first backward difference, adding no weights beyond one scalar γ. We prove that PSA is the unique per-stream high-pass, DC-neutral, swap-symmetric member of the factorised subclass of linear one-step causal bilinear augmentations (Theorem 4); that it is exactly symplectic at operator level (Theorem 6); and that it must be applied after RoPE (Corollary 7). In an explicit T_4-only Gaussian reduction of the score competition we derive a closed-form two-branch critical coupling, $\gamma_c^2 = 2L/(\sqrt{d_k} + \sqrt{d_k - 8L})$ for $d_k \geq 8L$, $L = \log(T-1)$ (Theorem 8), whose mean-field branch predicts held-out simulations at $r = 0.9876$ with zero fitted parameters, and which predicts that γ_c falls with head width, as trained models confirm in direction, though not in exponent. The SHT bound binds where the parameter budget approaches the bit budget of the two-layer circuit—negligible at frontier scale but potentially consequential for sub-100 M on-device models—and we work at 4–92 M by design. A one-layer, 54 k-parameter model reaches $69.1 \pm 8.2\%$ induction accuracy at $\gamma = 2$ (five seeds; chance 1.6%), and the pre-softmax spectrum of retrained single-layer models matches the closed-form transfer function at $r = 0.9966$; a separate 842 k-parameter multi-layer study locates the phase boundary across positional encodings. A matched control battery (seven variants plus a two-layer baseline) isolates the mechanism: key-side differencing is the operative ingredient, and a key-only shear (0.949) outperforms the symmetric operator (0.811), which we adopt for its structure rather than its accuracy. At first order the post-softmax response is a tangent correction of difference-of-attention character, related to Differential Transformer. All code is in the supplementary material as Jupyter notebooks with embedded outputs (for four multi-day sweeps, with their result files), including ten machine-checkable certificates, so every figure and number in the paper and appendices can be checked without GPU re-execution.

1 Introduction

Sanford et al. [16] (henceforth SHT) prove that single-layer attention cannot solve the in-context induction task—given a sequence of (token, target) pairs and a query token, retrieve the target paired with the earlier occurrence of the query—unless the per-head bit budget grows linearly in the context length T. The standard architecture realises induction with two stacked layers, a previous-token

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

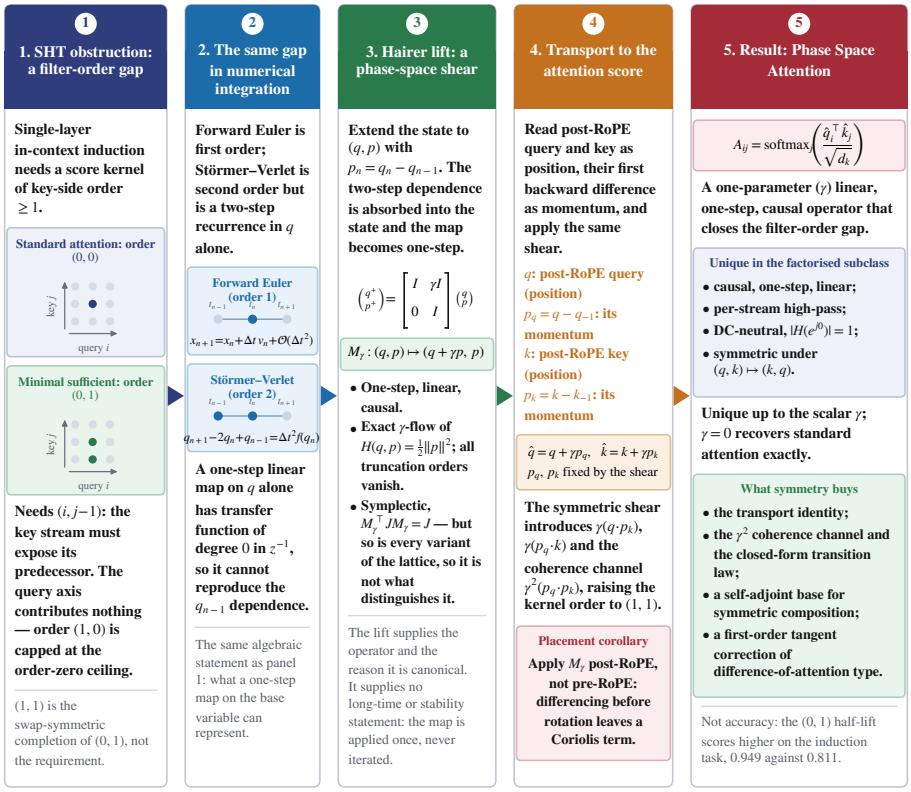


Figure 1: Roadmap from the SHT obstruction to PSA via Hairer's symplectic lift. **(1)** Standard attention is order $(0, 0)$; induction needs key-side order ≥ 1, whose minimal member is $(0, 1)$. **(2)** Forward Euler is one-step but first order; Störmer–Verlet attains order 2 but is two-step in q. **(3)** The lift extends the state to (q, p) and closes the gap with M_γ, the exact γ-flow of $H = \frac{1}{2}\|p\|^2$ (Lem. 2); symplecticity is shared by every member of the order lattice and does not distinguish it (Prop. 5). **(4)** Transporting M_γ to post-RoPE Q, K adds two single-ghost terms and the γ^2 coherence channel, raising the order to $(1, 1)$; placement must be post-RoPE (Cor. 7). **(5)** Within the factorised subclass M_γ is the unique causal, high-pass, DC-neutral, swap-symmetric augmentation (Thm. 4). What symmetry buys is structure, not accuracy (§7).

head followed by an induction head that copies the successor [21, 20, 23]; one layer cannot, in the SHT-precise sense, accomplish this with finite precision [16–19]. The bound is regime-specific. At frontier scale, where parameter budgets follow the usual scaling relations [44, 45], the extra layer is a rounding error; in the sub-100 M regime that ships in billions of phones, wearables and embedded controllers it is a material fraction of the budget. We therefore study 4–92 M models, without claiming a universal threshold: SHT's description size, parameter count, latency and cache cost are not interchangeable.

We exhibit the SHT obstruction as an algebraic statement about the linear time-invariant (LTI) class realised by a one-layer bilinear score (§2). The score $s_{t,j} = q_t^\top k_j$ has a factorised two-variable z-transform of degree zero in both variables, while the induction discriminator $\mathbf{1}\{\sigma_{j-1} = \sigma_t\}$ depends on the key-side *pair* (σ_{j-1}, σ_j), i.e. on a z_2^{-1} term. A polynomial of degree zero cannot represent one of degree one. This is a structural manifestation of the missing-history phenomenon that SHT quantify by a different argument, not an equivalence with their theorem. The same gap appears in numerical integration (§3): Störmer–Verlet is a two-step recurrence on configuration space with no one-step LTI realisation in q alone, and Hairer et al. [7, 8] circumvent this by passing to phase space, where the linear shear $M_\gamma : (q, p) \mapsto (q + \gamma p, p)$ is the exact γ-flow of $H = \frac{1}{2}\|p\|^2$. We perform the analogous lift on the attention substrate (§4): reading post-RoPE queries and keys as positions and their first backward differences as momenta, the lifted score $s_{t,j}^{\mathrm{PSA}} = (q_t + \gamma p_{q,t})^\top (k_j + \gamma p_{k,j})$ raises the joint order to $(1, 1)$ and makes single-layer induction representable. We call this *Phase Space Attention* (PSA). Figure 1 traces the correspondence.

Contributions. (i) A filter-order reading of the SHT obstruction, with an $O(1/T)$ ceiling for every key-side order-zero head (Prop. 1). (ii) The correspondence with the Verlet stencil and Hairer's lift

(§3), stated as a correspondence of algebra, not of long-time dynamics. (iii) Uniqueness of PSA within the factorised subclass under causality, per-stream high-pass, DC neutrality and swap symmetry (Thm. 4), with swap symmetry separated from symplecticity and characterised by a transport identity (Prop. 5). (iv) Operator-level symplectic invariance (Thm. 6) and a placement corollary with a closed-form pre-RoPE Coriolis term (Cor. 7). (v) A closed-form two-branch transition law for an explicit reduced score model (Thm. 8), validated held-out against simulation and in direction against trained models. (vi) An empirical programme at 4–92 M parameters, including a matched control battery that separates accuracy from structure (§7), and ten machine-checkable certificates.

What is claimed, and at which level. *Operator-level* statements concern the map M_γ itself and hold exactly, per application (Lem. 2, Thms. 4 and 6, Prop. 5, Cor. 7). *Transition-level* statements concern target softmax mass inside an explicit reduced model and are tested against simulation of that model (Thm. 8). *Empirical* statements are measurements on trained models, reported with protocols and seeds, including those that went against pre-registered predictions. No claim is promoted between levels, and we make no long-time, depth-wise or training-dynamical stability claim at any level: the operator is applied once per forward pass to a stream that is not its own output, so it generates no iterated dynamical system (§5). Accuracy is also not what the symmetric construction is for: on our induction benchmark the key-only half-lift is more accurate (§7), and we adopt the symmetric member for the structure it carries.

2 The single-layer induction obstruction as a filter-order gap

For a single head the unnormalised score is $s_{t,j} = Q_t^\top K_j$ with $Q_t = W_Q q_t$, $K_j = W_K k_j$. Since W_Q, W_K mix embedding coordinates within a position and never across positions, the joint z-transform in (t, j) factorises,

$$H_{\mathfrak{S}}(z_1, z_2) = H_q(z_1)\, H_k(z_2), \qquad \deg_{z_1^{-1}} H_q = \deg_{z_2^{-1}} H_k = 0. \tag{1}$$

We call $(\rho, \rho') = (\deg_{z_1^{-1}} H_q, \deg_{z_2^{-1}} H_k)$ the *joint filter order*. Summing heads, widening them or enlarging W_Q, W_K cannot raise it: order $(0, 0)$ is a property of the architecture, not of its parameter count. RoPE makes the score a function of $j - t$, but its coupling lives inside the substrate on which the order is counted, so no choice of RoPE frequency changes it.

Proposition 1 (Filter-order reading of SHT)**.** *Let token embeddings be i.i.d. zero-mean with covariance $\sigma^2 I_{d_k}$, $|\Sigma| \geq 2$. In the canonical induction model the target $j^* \in [1, T-1]$ is the unique index whose predecessor equals the query token A; distractors have own tokens drawn from Σ and predecessors from $\Sigma \setminus \{A\}$. If $s_{t,j} = \langle f(q_{\leq t}), g(k_j)\rangle$ has key-side order zero (f any function of the query stream), then for every f, g, every d_k and every precision,*

$$\Pr\big[\mathrm{softmax}_j(s_{t,\cdot}) \text{ peaks at } j^*\big] \;\leq\; \frac{1}{T-2}. \tag{2}$$

Proof sketch. Conditional on q_t, only position $j^* - 1$ has its token fixed (to A); the keys $\{k_j\}_{j \neq j^*-1}$ are i.i.d., so a score that depends on the key stream only through k_j is exchangeable on that sub-family. Since j^* is determined by the joint law of (σ_{j-1}, σ_j), it is uniform over the $T-2$ positions of the sub-family. Equality holds when the head suppresses $j^* - 1$; a head aligned with it falls below $1/(T-1)$. Full proof in Appendix T1.2. □

Relation to SHT. Proposition 1 is strictly weaker than SHT's theorem and internal to the bilinear class; it does not reproduce their counting argument. Its role is diagnostic: it locates the deficit on the key axis. Correspondingly, PSA does *not* escape SHT's bound. The PSA head's description size remains $O(d_k)$, independent of T, and its score depends jointly on (k_j, k_{j-1}), so it lies outside the class the counting argument enumerates: the premise is exited, the bound stands.

The order lattice. With the DC-neutral high-pass taps of Theorem 4, the four candidate orders are realised by $H_q(z_1) = 1 + \gamma_q(1 - z_1^{-1})$, $H_k(z_2) = 1 + \gamma_k(1 - z_2^{-1})$ with $(\rho, \rho') = (\mathbf{1}\{\gamma_q > 0\}, \mathbf{1}\{\gamma_k > 0\})$. $(0, 0)$ is standard attention; the query-only half-lift $(1, 0)$ leaves the key stream order-zero, so the ceiling (2) still binds for every γ_q (the discriminating negative control); the key-only half-lift $(0, 1)$ is the minimal representable member; and $(1, 1)$ is the swap-symmetric completion studied below. Order, symmetry and symplecticity are independent: key-side order ≥ 1 is what representability requires, and every member of the lattice is exactly symplectic (Prop. 5).

3 A parallel obstruction in numerical integration

For $\ddot{q} = f(q)$, forward Euler is first order, and on the quadratic potential $f(q) = -q$ its update is $\sqrt{1+\Delta t^2}$ times an orthogonal map, so the energy grows by exactly $(1+\Delta t^2)^n$ after n steps. The Störmer–Verlet stencil $q_{n+1} - 2q_n + q_{n-1} = \Delta t^2 f(q_n)$ is second order, symmetric and free of secular drift on quadratic potentials [12, 10, 11, 3], but it is *two-step* in q: a one-step linear map on configuration space has degree zero in z^{-1} and cannot reproduce the q_{n-1}-dependence—the stencil obstruction, an exact parallel of (1). Hairer et al. [7, 8] absorb the lag into the state: with $p_n = (q_n - q_{n-1})/\Delta t$ Verlet becomes the one-step kick–drift–kick palindrome, which preserves $\omega = dq \wedge dp$ exactly. In the free-particle limit $f \equiv 0$ it reduces to the drift alone, the shear

$$M_\gamma = \begin{pmatrix} I & \gamma I \\ 0 & I \end{pmatrix}, \qquad M_\gamma^\top J M_\gamma = J, \qquad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}. \tag{3}$$

Lemma 2 (Exact γ-flow)**.** *M_γ is the exact time-γ flow of $H(q,p) = \frac{1}{2}\|p\|^2$, with all truncation orders vanishing.*

Proof. $\dot{q} = \partial_p H = p$ and $\dot{p} = -\partial_q H = 0$; integrating from 0 to γ gives $(q + \gamma p, p)$. The flow is polynomial of degree one in time, so the shear is the flow, not an approximation. □

The correspondence—stencil order $0 \to 2$ in q, joint filter order $(0,0) \to (1,1)$ in (q,k), both closed by M_γ acting on position $\to$ (position, momentum)—is between the two *representability* arguments. The long-time results of geometric integration (backward error analysis, modified Hamiltonians, exponentially long energy bounds) are statements about a map iterated on its own output; §5 shows that no such iterate exists on the attention side, and none of those results is carried across.

4 Phase Space Attention as the Hairer lift

Let $\{q_t\}, \{k_j\}$ be the post-RoPE query and key streams, with kinematic momenta $p_{q,t} = q_t - q_{t-1}$, $p_{k,j} = k_j - k_{j-1}$ for $t, j \geq 2$ and $p_{q,1} = p_{k,1} = 0$ (no artificial first-token impulse).

Definition 3 (Phase Space Attention)**.** For $\gamma > 0$ the PSA score is

$$s^{\mathrm{PSA}}_{t,j} := (q_t + \gamma p_{q,t})^\top (k_j + \gamma p_{k,j}) = \underbrace{q_t^\top k_j}_{T_1} + \gamma \underbrace{p_{q,t}^\top k_j}_{T_2} + \gamma \underbrace{q_t^\top p_{k,j}}_{T_3} + \gamma^2 \underbrace{p_{q,t}^\top p_{k,j}}_{T_4}, \tag{4}$$

with weights $\mathrm{softmax}_j(s^{\mathrm{PSA}}_{t,j}/\sqrt{d_k})$ under the causal mask. Values are not augmented.

The augmented key $\hat{k}_j = (1+\gamma)k_j - \gamma k_{j-1}$ carries a predecessor component, the *ghost key*, inside the score-bearing key of position j: the predecessor copy that the previous-token layer of a standard induction circuit performs, and that the SHT bound prices, is replaced by a fixed linear functional of the stream the layer already has. Each stream is a causal FIR filter

$$H^u(z) = (1+\gamma) - \gamma z^{-1}, \qquad \left|H^u(e^{j\omega})\right|^2 = 1 + 4\gamma(1+\gamma)\sin^2(\omega/2), \tag{5}$$

with unit DC gain, Nyquist gain $1 + 2\gamma$, and strictly increasing magnitude on $(0, \pi]$; the score therefore has joint order exactly $(1,1)$. The implementation is two backward differences and two affine combinations before an unmodified attention kernel (Algorithm 1), costing $O(Td_k)$ extra time and memory; it composes with FlashAttention, and the sheared keys can be cached in place of the raw ones (Appendix T7).

Theorem 4 (Uniqueness within the factorised subclass)**.** *Let $\mathcal{C}_{\mathrm{PSA}}$ be the class of score augmentations $s_{t,j} = \Phi(q_t, q_{t-1}, k_j, k_{j-1})$ on post-RoPE streams with Φ bilinear separately in its q- and k-arguments, and consider its factorised subclass $\hat{q}_t = \alpha q_t + \beta q_{t-1}$, $\hat{k}_j = \alpha' k_j + \beta' k_{j-1}$, $s_{t,j} = \hat{q}_t^\top \hat{k}_j$. Impose* (C1) *causality and one-step memory;* (C2) *per-stream high-pass: $|H(e^{j0})| < |H(e^{j\pi})|$ with $|H(e^{j\omega})|$ non-decreasing on $[0,\pi]$;* (C3.1) *DC neutrality: $H(e^{j0}) = 1$ on each stream; and* (C3.2) *swap symmetry: the score is invariant under $(q,k) \mapsto (k,q)$. Then $\alpha = \alpha' = 1+\gamma$ and $\beta = \beta' = -\gamma$ for a single $\gamma > 0$: Definition 3 is the unique member satisfying* (C1)–(C3.2)*, up to the scalar γ.*

Algorithm 1 PSA forward pass (one head, post-RoPE). Lines 1–3 are the entire change; lines 4–6 are standard attention.

Require: post-RoPE $Q, K \in \mathbb{R}^{T\times d_k}$, values $V \in \mathbb{R}^{T\times d_v}$, $\gamma > 0$, causal mask $\mathbf{M}$

1: $P_q \leftarrow Q - \mathrm{shift}_{+1}(Q)$; $P_k \leftarrow K - \mathrm{shift}_{+1}(K)$ *# first differences; row 1 set to 0*

2: $\hat{Q} \leftarrow Q + \gamma P_q$; $\hat{K} \leftarrow K + \gamma P_k$ *# upper shear M_γ on each stream*

3: (cache $\hat{K}$ in place of K during decoding)

4: $S \leftarrow \hat{Q}\hat{K}^\top/\sqrt{d_k}$ *# order-$(1,1)$ score, Eq. (4)*

5: $A \leftarrow \mathrm{softmax}_j(S + \mathbf{M})$

6: **return** $Y \leftarrow AV$ *# values untouched*

Proof sketch. (C1) gives the two-tap form. On each stream $|H|^2 = \alpha^2 + \beta^2 + 2\alpha\beta\cos\omega$, so (C2) holds iff $\alpha\beta < 0$; (C3.1) gives $\alpha + \beta = 1$, hence $\beta = -\gamma < 0 < \alpha = 1 + \gamma$ and $\hat{q}_t = q_t + \gamma p_{q,t}$, $\hat{k}_j = k_j + \gamma' p_{k,j}$. Expanding the score with $(\gamma_q, \gamma_k) = (\gamma, \gamma')$, the swap exchanges the two single-ghost terms, so (C3.2) holds iff $\gamma = \gamma'$; the deviation for $\gamma_k = \gamma_q + \varepsilon$ is exactly $\varepsilon(q_t^\top p_{k,j} + \gamma_q p_{q,t}^\top p_{k,j})$. Full proof, the elimination of the lower shear, skew generators and RoPE reuse, and the non-linear extension in Appendix T3.4–T3.6. □

Of the four conditions, (C1) permits the order one that representability needs; (C2) selects the backward-difference orientation among order-one filters (a pure delay z^{-1}, as in KV-shifting [30], has order one but is all-pass); (C3.1) makes the augmentation a difference rather than a rescaling; and (C3.2) is a design commitment, *not* a consequence of symplecticity. The next proposition makes this precise and supplies the non-axiomatic content of (C3.2).

Proposition 5 (Symplecticity versus symmetry)**.** (i) *For all $\gamma_q, \gamma_k > 0$ the block map $\mathrm{diag}(M_{\gamma_q}, M_{\gamma_k})$ on (q, p_q, k, p_k) is exactly symplectic and conserves $H_{\mathrm{joint}} = \frac{1}{2}(\|p_q\|^2 + \|p_k\|^2)$; unequal couplings change the score by the bias $\varepsilon(q_t^\top p_{k,j} + \gamma_q p_{q,t}^\top p_{k,j})$, linear in $\varepsilon = \gamma_k - \gamma_q$, with no energy component.* (ii) (Transport identity.) *With $\Delta_{\mathrm{diag}} F(t,j) := F(t,j) - F(t-1, j-1)$, the symmetric score satisfies exactly*

$$s_{t,j}^{\mathrm{PSA}} = q_t^\top k_j + \gamma\, \Delta_{\mathrm{diag}}(q^\top k)(t,j) + \gamma(1+\gamma)\, p_{q,t}^\top p_{k,j}, \tag{6}$$

the base pairing transported one integrator step; for unequal couplings the score acquires in addition the antisymmetric current $\frac{\varepsilon}{2}(q_t^\top p_{k,j} - p_{q,t}^\top k_j)$, which is not a total difference of any local bilinear form.

Proof sketch. (i) Each block is unit upper triangular and $M_\gamma^\top J M_\gamma = J$ identically, with no compatibility condition between blocks; both momenta are fixed. (ii) Substitute the discrete Leibniz rule $\Delta_{\mathrm{diag}}(q^\top k) = p_q^\top k + q^\top p_k - p_q^\top p_k$ into (4); the non-existence claim is a coefficient comparison. Certificates C4, C5 and C9 verify (i), the bias and (ii) symbolically (Appendix T3.8). □

Proposition 5(i) retracts the claim of the reviewed version that unequal couplings incur energy drift: symplecticity does not single out the symmetric member. What does is (ii), the score-level image of self-adjointness of the symmetric Verlet composition.

5 Operator-level invariance and the placement corollary

Theorem 6 (Operator-level symplectic invariance)**.** *Let $\mathcal{M}_\gamma = \mathrm{diag}(M_\gamma, M_\gamma)$ act on the joint pre-softmax phase space $(q, p_q, k, p_k) \in \mathbb{R}^{4d_k}$ with canonical form Ω. Then, per application and for every $\gamma \in \mathbb{R}$:* (i) *$\mathcal{M}_\gamma^\top \Omega \mathcal{M}_\gamma = \Omega$ exactly;* (ii) *H_{joint} is exactly conserved; and* (iii) *if $\|\mathcal{M}' - \mathcal{M}_\gamma\|_2 \le \epsilon$ then $\|\mathcal{M}'^\top \Omega \mathcal{M}' - \Omega\|_2 \le 2\epsilon\|\mathcal{M}_\gamma\|_2 + \epsilon^2$.*

Proof sketch. Block computation gives $M_\gamma^\top J M_\gamma = \left(\begin{smallmatrix} 0 & I \\ -I & -\gamma I + \gamma I \end{smallmatrix}\right) = J$; the cancellation is exact, so the residual evaluates to 0.0 in floating point, not merely to machine precision (certificate C6). Both momenta are fixed points. (iii) is a first-order expansion with $\|\Omega\|_2 = 1$ (Appendix T4.1). □

What the theorem does not say. All three statements are per application; none is a stability guarantee for attention, training, depth or sequence. PSA forms $\hat{q}_t$ from momenta of the *untransformed* stream, and re-extracting a momentum from its output gives $\hat{q}_t - \hat{q}_{t-1} = (1+\gamma)p_{q,t} - \gamma p_{q,t-1} \neq p_{q,t}$: the transformed pair is not a phase state closed under the map. The shear is applied once per forward pass, layer $\ell + 1$ shears freshly projected queries and keys, and no composition $\mathcal{M}_\gamma \circ \mathcal{M}_\gamma$ occurs anywhere. Consequently we make no modified-Hamiltonian, backward-error, long-time or secular-drift statement on the attention side, and conservation of the auxiliary quantity H_{joint} is not evidence of stability (indeed $\|M_\gamma^n\|$ grows linearly in n, so iteration would not be power-bounded). This narrows the stability theorem of the reviewed version; Appendix T4.2 tabulates what transfers from geometric integration.

Corollary 7 (Placement post-RoPE)**.** *M_γ does not commute with RoPE. Applying it before the rotation R_t produces the Coriolis error*

$$\hat{q}_t^{\text{pre}} - \hat{q}_t^{\text{post}} = -\gamma(R_t - R_{t-1})q_{t-1}, \qquad \|\hat{q}_t^{\text{pre}} - \hat{q}_t^{\text{post}}\| = 2\gamma|\sin(\theta/2)|\,\|q_{t-1}\|, \tag{7}$$

with θ the per-band rotation angle, which destroys the per-stream high-pass character (5) in the rotating frame and forfeits the transport identity. It does not break symplecticity, since the composite $R_t M_\gamma R_{t-1}^{-1}$ is a product of symplectic maps.

Proof sketch. Pre-RoPE gives $R_t q_t + \gamma R_t q_t - \gamma R_t q_{t-1}$; post-RoPE gives $R_t q_t + \gamma R_t q_t - \gamma R_{t-1} q_{t-1}$; $R_t R_{t-1}^{-1}$ is a rotation by θ (Appendix T4.3). □

The penalty vanishes as $\theta \to 0$ and is largest in the high-frequency bands, which carry the least induction signal (§7). A single contrast of embedding-level against post-RoPE shear points the predicted way but is not a matched pair and does not isolate operator ordering (Appendix A30).

6 A closed-form transition in a reduced score model

Representability is not attainment. We now ask at which coupling the augmented head concentrates softmax mass on the target. Under isotropy T_1 is constant across candidates and T_2 is mean-zero, while two terms carry a target–distractor gap: T_3 at first order in γ and the AC–AC part of the coherence channel at second order, $\mathbb{E}[\gamma^2 T_4^{AA}(j^*)] = -\gamma^2\sigma^2 d_k$ (Appendix T5.2). The raw gap is negative; the learned projections W_Q, W_K fix its sign, and everything below uses only its magnitude. Normalising signal *and* fluctuation consistently by $\sqrt{d_k}$, the T_4 channel has target mean $a = \gamma^2\sigma^2\sqrt{d_k}$ and per-distractor scale $b = 2\gamma^2\sigma^2$, so

$$\text{SNR} = a/b = \sqrt{d_k}/2, \tag{8}$$

independent of γ and growing with head width: the coupling acts only as an inverse temperature, and any transition is a competition effect. (The reviewed version divided the signal by d_k but not the variance, which inverted the d_k dependence; Eq. (8) corrects it.)

Theorem 8 (Two-branch transition in the reduced model $\mathcal{G}$)**.** *Let $L = \log(T-1)$, $m = T-2$, and let $\mathcal{G}(d_k, T; \gamma)$ be the reduced score model defined by* (M1) *the gap is carried by T_4 alone (the γ-linear T_3 gap is omitted);* (M2) *normalised moments $a = \gamma^2\sqrt{d_k}$ and $s_j = bZ_j$, $b = 2\gamma^2$, at $\sigma = 1$;* (M3) *Z_j i.i.d. standard Gaussian;* (M4) *the characteristic coupling equates the target's weight with the* expected *distractor mass (annealed criterion). Then the expected target softmax mass $\mathcal{A} = \mathbb{E}[\text{softmax}_j(s)_{j^*}]$ rises from $1/(T-1)$ at $\gamma = 0$ to the γ-independent limit $p_\infty(d_k, T) = \int \varphi(z)\,\Phi(z + \sqrt{d_k}/2)^m\,dz$, through a transition at*

$$\gamma_c^2 = \begin{cases} \dfrac{2L}{\sqrt{d_k} + \sqrt{d_k - 8L}}, & d_k \geq d_k^* := 8L \quad \textit{(mean field)}, \\ \frac{1}{2}c^2\sqrt{2L}, & d_k < d_k^* \quad \textit{(condensed)}, \end{cases} \tag{9}$$

where $c = 0.741 \pm 0.015$ is a condensation constant calibrated on the condensed branch. On the mean-field branch $\gamma_c \sim (L^2/d_k)^{1/4}$, and the annealed criterion saturates its own validity bound $b \leq \sqrt{2L}$ exactly at d_k^*.

Proof sketch. By the Gaussian moment generating function the annealed balance is $e^a = m\,e^{b^2/2}$, i.e. $a = L + b^2/2$ (with $\log m \to L$, an $O(1/T)$ shift), which with (M2) is the quartic $2\gamma_c^4 - \sqrt{d_k}\gamma_c^2 + L =$

0; its smaller root, in conjugate form, is the first branch and is real iff $d_k \geq 8L$. At the crossing $b_c = \frac{1}{2}(\sqrt{d_k} - \sqrt{d_k - 8L})$ decreases in d_k and equals $\sqrt{2L}$ exactly at d_k^*: the freezing point of Derrida's random energy model [14, 15], below which the distractor sum is extreme-value dominated and the transition is located by freezing itself, $2\gamma_c^2 = \sqrt{2L}$, up to the $O(1)$ finite-m factor c. The limit p_∞ is the Gaussian argmax probability at SNR (8). Certificates C1–C3 verify the algebra; C7–C8 check p_∞ and the root against fresh Monte Carlo (Appendix T5.7). □

Theorem 8 replaces the erf-form law $\gamma_c = (d_k \log(T-1))^{1/4}$ of the reviewed version, which is retracted together with the small-signal and softmax-linearisation assumptions it rested on: the derivation above is non-perturbative in γ. It is a theorem about $\mathcal{G}$, not about trained transformers; transfer is tested separately in §7, and the argmax ordering inside $\mathcal{G}$ is γ-invariant, so the transitioning observable is softmax mass. The identification of attention scores with a random energy model is due to Giorlandino and Goldt [47] (see also 48); what is new here is the swept knob, a coupling inside the operator at fixed inverse temperature. We stress the most reduction-sensitive feature: retaining the omitted T_3 channel (Appendix T9) gives a quartic without closed form, moves γ_c at $d_k = 64$, $T = 512$ from 1.030 to 0.569, steepens the asymptotic exponent from -0.25 to -0.46, and removes the crossover jump. The branch structure is therefore a *discriminating* prediction that separates the two reductions, not a confirmed one.

Corollary 9 (Half-lifts)**.** (i) *The query-only half-lift* $(1, 0)$ *is capped at* $1/(T-2)$ *for every* γ_q, d_k *and training procedure, since the key stream remains order-zero (Prop. 1).* (ii) *The key-only half-lift* $(0, 1)$ *has first-order gap* $a' = \gamma\sqrt{d_k}$, *scale* $b'^2 = 1+2\gamma^2$, *transition* $\gamma_c = (2L+1)/(\sqrt{d_k}+\sqrt{d_k - 4L - 2})$ *for* $d_k \geq 4L + 2$, *and an asymptotic SNR* $\sqrt{d_k}/\sqrt{2}$, *a factor* $\sqrt{2}$ *above the symmetric operator. The transition law in* (ii) *is conditional and has not been tested against simulation.*

Both were pre-registered: (i) admits no free parameter, and any query-only accuracy above $1/(T-2)$ would falsify the filter-order reading; (ii) predicts that the minimal member should beat the symmetric completion on single-layer induction.

7 Empirical signatures

All experiments are synthetic induction tasks on small decoders (main-text experiments use at most 92 M parameters, most below 1 M). Every number is attached to its protocol and seeds; the programme comprises about 9,400 training runs and forty-one notebooks, thirty-seven with embedded outputs (Appendix T6.2). Where the final reruns refined a value quoted in the rebuttal (e.g. c, the spectral correlations), the value given here supersedes it.

Single-layer phase jump (Fig. 2a). Single-layer PSA reaches $69.1 \pm 8.2\%$ over five seeds at $\gamma^\star{=}2$ (best seed 79.4%), against $1.6 \pm 0.4\%$ for one standard layer and $1.6 \pm 0.3\%$ for two standard layers at matched budget, both at chance: a $+67.5$ pp mean jump, $43\times$ the baseline. This five-seed figure is the headline of this version; the 74.2% quoted in the reviewed version is the peak of the separately trained, three-seed-evaluated sweep plotted in the same panel. The multi-layer study of Fig. 2b is a separate claim: both positional-encoding families saturate together (98.5–99.6% for $\gamma \in [0.7, 3]$) and share a boundary to within $1.22\times$.

The order lattice, measured. Table 1 and Fig. 3 measure all four lattice members and three further mechanisms under one pre-registered protocol. Four readings follow. *(i)* The query-only half-lift never exceeds the $1/(T-2) = 0.077$ ceiling at any coupling, as Corollary 9(i) requires; it sits below the $1/(T-1) = 0.071$ benchmark because a head aligned with the previous-token position spends its argmax there. Leaving the $(0, 0)$ class is necessary but not sufficient. *(ii)* The key-only half-lift is the strongest fixed variant and dominates the symmetric operator at every coupling (0.949 vs. 0.811): *if single-layer induction accuracy is the objective, the key-side half-lift is the better construction*. Both are exactly symplectic, so the gap reflects the symmetric operator spending half its coupling on the query axis, which contributes nothing to representability. *(iii)* A low-pass two-tap injecting exactly one step of context reaches only 0.428: the ingredient is DC-neutral high-pass differencing, not local context as such. *(iv)* Asymmetric couplings train well, with a key-heavy ordering ($(1, 3)$: 0.875; $(3, 1)$: 0.630), consistent with the absorbable bias of Prop. 5. Two pre-registered predictions failed and are reported as such: a fully learned two-tap converged to the *minimal* member (identity on

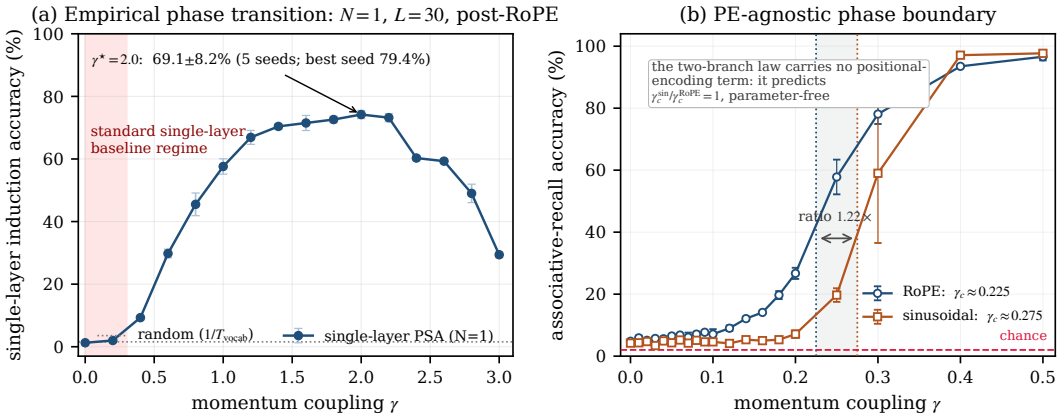


Figure 2: Two separate claims on deliberately different tasks; absolute accuracies are not comparable. **(a)** Single-layer induction under distractor noise: one-layer decoder, ∼54 k parameters, d_{model}=64, 4 heads, d_k=16, shared key/target alphabet of 64 tokens, 14 pairs padded to length 30. The plotted sweep is one trained model evaluated on three seeds (peak 74.2% at γ=2, 1.27% at γ=0); the annotated statistic is an independent five-seed run, $69.1 \pm 8.2\%$ at $\gamma^\star$=2 against $1.6 \pm 0.4\%$ (chance $1/64$). The shaded band marks couplings not measurably above baseline; it is not a region excluded by SHT, which constrains description size, not γ. **(b)** Phase boundary on a sanitised multi-layer task: four layers, ∼842 k parameters, d_k=32, disjoint key/target sub-alphabets, 12 pairs, three seeds; $\gamma_c \approx 0.225$ (RoPE) and 0.275 (sinusoidal), a ratio of 1.22; the reduced model of Theorem 8 has no positional-encoding term. In panel (a) the label box reports the independent five-seed statistic; its arrow marks $\gamma^\star$=2, not the peak of the plotted sweep.

Table 1: Matched control battery: $T-1$=14 pairs, d_k=16, d_{model}=64, three seeds × 3,000 steps, taps post-RoPE, values untouched, chance 0.071, ceiling for key-side order zero $1/(T-2) = 0.077$. Best accuracy over $\gamma \in \{0.5, 1, 2, 3\}$ (asymmetric: six (γ_q, γ_k) pairs), mean ± s.d. over seeds; grid maxima are best-of-k statistics, applied identically to every variant. Every variant except the low-pass and learned filters is exactly symplectic.

Variant	Order	Best accuracy	Reading
One-layer baseline	$(0, 0)$	0.033 ± 0.006	below the order-zero ceiling (Prop. 1)
Query-only shear	$(1, 0)$	0.033 ± 0.007	same class, same value at every γ (Cor. 9)
Key-only shear	$(0, 1)$	$\mathbf{0.949 \pm 0.006}$	minimal member; strongest fixed variant (γ=3)
PSA symmetric	$(1, 1)$	0.811 ± 0.006	rises from 0.505 (γ=0.5) to its peak at γ=3
Asymmetric (γ_q, γ_k)	$(1, 1)^*$	0.875 ± 0.008	at $(1, 3)$; bias absorbed (Prop. 5)
Low-pass two-tap	not DC-neutral	0.428 ± 0.018	one step of context is not the ingredient
Learned two-tap	learned	1.000 ± 0.000	taps b/a: −0.010 (q), −1.040 (k)
Two-layer baseline	$(0, 0)$, N=2	0.036 ± 0.001	0.024 ± 0.006 at 9,000 steps

queries, pure difference on keys) rather than the symmetric one, reaching 1.000; and the two-layer baseline did not learn the task at either budget, so we claim no impossibility for two standard layers. What the symmetric member carries instead is the structure: uniqueness (Thm. 4), the transport identity (Prop. 5), the γ^2 channel on which Theorem 8 is built, a self-adjoint base for symmetric composition, and the first-order response below.

Bode forensics for the score stream. On single-layer models retrained from scratch at each coupling (d_k=16, T=64 with 30 pairs, 3,000 steps, 200 held-out sequences) the measured pre-softmax AC power spectrum correlates with the closed form (5) at $r = 0.8951$, 0.9890, 0.9915, 0.9966 and 0.9924 for $\gamma = 0.6$, 1.2, 1.8, 2.0, 3.0, and after the softmax at $r = 0.94$–0.96; standard attention gives a flat spectrum and a degenerate correlation. This tests spectral *shape* only. These values replace those of the reviewed version after two diagnostic corrections: spectra are now normalised by total AC power rather than by the residual DC bin, which after per-row mean subtraction carries only 1.4×10^{-3} of the power (Pearson r is scale-invariant, certificate C10, so the change moves r by at most 2.2×10^{-16}); and symplecticity is reported by the residual $\|M_\gamma^\top J M_\gamma - J\|_{\max} = 0.0$ rather than by a determinant, which cannot distinguish symplectic from merely volume-preserving maps when $d_k > 1$. On a 2,000-run sweep over γ and the RoPE centre frequency θ, induction concentrates at low θ and the momentum gain falls with θ ($r = -0.679$, $p = 9.9 \times 10^{-4}$, $n = 20$): the score-axis high-pass detects the embedding-axis low-frequency structure in which RoPE concentrates energy [51] (Appendix A16).

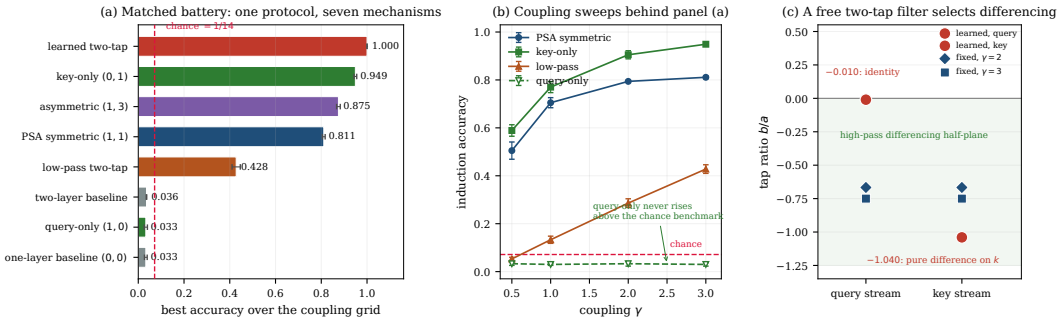


Figure 3: Matched control battery (protocol of Table 1). **(a)** Best accuracy per variant. **(b)** Coupling sweeps: the key-only curve dominates the symmetric one at every γ, and the query-only half-lift never rises above the chance benchmark. **(c)** Converged taps of a free two-tap filter against the fixed operator: identity on queries, a pure first difference on keys.

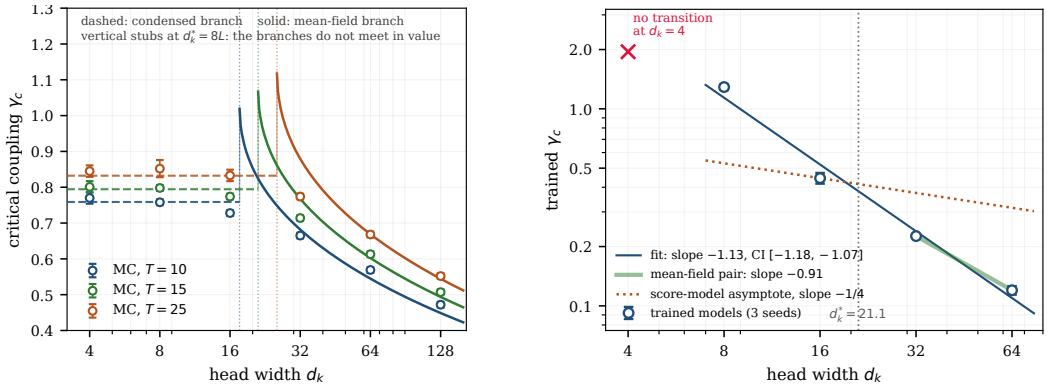


Figure 4: Left: Theorem 8 against direct simulation of $\mathcal{G}$ at three context lengths, five seeds and 6×10^4 trials per point (dashed: condensed branch; solid: mean field; stubs mark the value gap at $d_k^*{=}8L$). **Right:** trained γ_c against d_k (log–log; three seeds; no transition at $d_k{=}4$): direction and ordering agree with the law, the slope does not. Further panels in Appendix T6.9 and T6.11.

The transition law against simulation and trained models. On an 18-point (d_k, T) grid of direct simulations of $\mathcal{G}$ (five seeds, 6×10^4 trials per point, grid fixed before the law was evaluated), the mean-field branch of (9) predicts the nine points with $d_k \geq 8L$ at $r = 0.9876$, median error 2.6%, with zero fitted parameters (the condensed branch, which carries the calibrated c, is a consistency check only); the largest errors (+13.6%, +7.5%) sit next to the crossover, where the piecewise law has a jump that the data do not (Fig. 4, left). Rebuilding the ensemble with its true dependence structure moves γ_c by a median -0.6% (max 2.9%); the plateau p_∞ matches simulation to three decimals at all 18 points; and the condensed branch predicts the eight on-branch points of a ten-length T-axis scan at $r = 0.9995$, with $c = 0.717$ there against 0.741 on the grid. On trained models (Fig. 4, right; $T{-}1{=}14$, $d_{\text{model}}{=}4d_k$, three seeds), γ_c *falls* with head width as the law predicts—no transition at $d_k{=}4$, then 1.290, 0.445, 0.226 and 0.120 at $d_k = 8, 16, 32, 64$—resolving the opposite trend that the reviewed law implied, and the plateaus track p_∞ in shape ($r = 0.945$) but not in level. Transfer is approximate: the trained log–log slope is -1.13 (95% CI $[-1.18, -1.07]$) against the abstraction's $-1/4$, an open, measured gap. The operating-point ratio $\gamma\|p\|/\|q\|$ at $\gamma = 2$ is 1.86, 3.05, 3.12, 3.03 and 2.92 at $d_k = 4, \ldots, 64$, so the small-signal premise of the reviewed version does not hold there; Theorem 8 does not use it.

Coupling, depth and length: a characterisation. Measured, pre-registered, and not a theorem: across sixteen configurations (symmetric $\gamma \leq 8$ at depths 1 and 3, plus asymmetric and key-only variants; two seeds) no run diverged; the pre-softmax score RMS stays flat through depth for sheared models while the $\gamma{=}0$ control grows $5.8\times$ from layer 1 to 2; and a model trained at $\gamma{=}2$ keeps $\geq 90\%$ of its best accuracy when evaluated at any $\gamma \in [1.5, 4]$ (Appendix T6.12).

Proposition 10 (First-order response)**.** *With* $a_t = \text{softmax}_j(q_t^\top k_j/\sqrt{d_k})$ *and* $J_{a_t} = \text{diag}(a_t) - a_t a_t^\top$, $a_t^{\text{PSA}} = a_t + \gamma J_{a_t}\left[(p_{q,t}^\top k_\cdot + q_t^\top p_{k,\cdot})/\sqrt{d_k}\right] + O(\gamma^2)$*, a correction tangent to the simplex.*

The query-side part is the Jacobian acting on a difference of two score vectors, i.e. a tangent correction of difference-of-attention character. Differential Transformer [29] instead subtracts two separately normalised softmax maps; the two agree only under an additional smallness condition on $\|s_t - s_{t-1}\|$, which small γ does not supply, so we claim no architectural identity—a narrowing of the reviewed version (Appendix T8.2).

Certificates. Ten certificates (SymPy and float64 numerics; under a minute on a CPU) verify C1–C3 the transition algebra, C4–C5 the unequal-shear statements, C6 the operator residual, C7–C8 p_∞ and the root against fresh Monte Carlo, C9 the transport identity and C10 Pearson scale invariance. They check algebra conditional on the definitions, not modelling assumptions (Appendix T6.15).

8 Related work, limitations and discussion

Related work. Short causal convolutions on Q, K or V are known to unlock single-layer associative recall [4, 5], and KV-shifting [30] supplies a lagged key directly; these establish sufficiency for learned filters. Instead, we identify the DC-neutral first difference as the minimal fixed member, characterise its unique swap-symmetric completion, and derive the transition theory of its reduced model; the learned two-tap of §7 converges to that minimal member from data. This paper has two preprints. The first author's early preprint [52] introduced the same augmentation as "momentum attention" with spectral forensics, but none of the filter-order reading, the Hairer correspondence, the uniqueness, invariance and placement results or the transition law; the extended version of this paper [53] is the complete account, and Part T of the appendix reproduces it. Gahtan and Bronstein [49] couple Q and K through an integrator and find symplectic and non-symplectic integrators performing alike once the coupling is present, consistent with Prop. 5; their construction lies outside $\mathcal{C}_{\mathrm{PSA}}$, and Theorem 4 says nothing about it. Hamiltonian and symplectic networks [38, 39, 43, 41, 42, 40] learn and integrate a flow; ours is fixed and applied once. The transition here is in an architectural coupling at fixed training budget, not the training-time emergence of induction heads [20, 24].

Limitations. *The evidence is overwhelmingly synthetic*, and *practical gains at realistic language-model scale remain unestablished*: a $\sim$300 M parameter-efficiency study is a near-miss (Appendix A32), multi-hop reasoning with negations shows no effect ($\Delta = -0.4 \pm 1.2$ pp), the depth law fails at depth 30, and on randomised-offset induction at nano scale the best coupling is $\gamma^*{=}0$ at every depth tested (Appendix T6.14). Most claims rest on two to five seeds with seed standard deviations and best-of-k grid maxima rather than inferential statistics; the large separations we rely on ($+67.5$ pp, 13.8 pp) do not depend on marginal contrasts. The two headline scales (54 k single-layer, 842 k multi-layer) are separate claims. Theorem 8 concerns a reduced model whose branch structure is sensitive to the omitted T_3 channel, and its trained-model exponent is off by a factor of about four. Uniqueness is class-relative and swap symmetry is a design commitment; on the induction task alone the key-only member is the better engineering choice. Preserving the auxiliary pre-softmax quantity H_{joint} provides no demonstrated stability guarantee for a transformer, and stability is reported as evidence, not proved.

Corrections relative to the reviewed version. Following the reviews, we (i) replaced "resolves" by "circumvents" and state precisely that the SHT bound is not escaped; (ii) scoped uniqueness to the factorised subclass, added DC neutrality to the stated conditions, and separated swap symmetry from symplecticity, retracting the asymmetric energy-drift claim; (iii) conceded that the transformed stream is not a closed phase state and restated the stability theorem as operator-level invariance; (iv) retracted the erf-form transition law and its small-signal premise, correcting the event orientation, the chance level and the variance normalisation, and replaced it by Theorem 8; (v) sharpened the order-zero ceiling to $1/(T-2)$; (vi) corrected the spectral normaliser and the symplectic diagnostic; (vii) adopted the five-seed $69.1 \pm 8.2\%$ as the headline; and (viii) added the matched control battery and narrowed the Differential Transformer statement.

Outlook. Whether symmetric composition schemes [12, 11] yield a useful higher-order score is open. The falling $\gamma_c(d_k)$ is a concrete prediction for scale: at production widths the operative couplings are small ($\gamma_c\|p\|/\|q\| \approx 0.18$ at $d_k{=}64$), where the first-order response is quantitative (discriminating test: Appendix T9).

Acknowledgments and Disclosure of Funding

Funding. This work was funded by Qualcomm Technologies, Inc. No third-party funding or support was received in direct support of this work.

Competing interests. All authors are employees of Qualcomm, their listed primary affiliation. The authors declare no other competing interests.

References

[1] N. N. Čencov. *Statistical Decision Rules and Optimal Inference*, volume 53 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 1982.

[2] B. Chang, M. Chen, E. Haber, and E. H. Chi. AntisymmetricRNN: a dynamical system view on recurrent neural networks. In *International Conference on Learning Representations (ICLR)*, 2019.

[3] G. Benettin and A. Giorgilli. On the Hamiltonian interpolation of near-to-the-identity symplectic mappings with application to symplectic integration algorithms. *Journal of Statistical Physics*, 74(5):1117–1143, 1994.

[4] M. Li, X. Zhang, Y. Huang, and S. Oymak. On the power of convolution-augmented transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 39(17):18393–18402, 2025. arXiv:2407.05591.

[5] K. A. Wang, J. Shi, and E. B. Fox. Test-time regression: a unifying framework for designing sequence models with associative memory. *Journal of Machine Learning Research*, 27(169):1–41, 2026. arXiv:2501.12352.

[6] I. Lavie, K. Fischer, A. Lekov, F. Van Maele, Z. Ringel, and M. Helias. Phase transitions in attention: a Bayesian theory of copy head emergence. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2026. arXiv:2606.12058.

[7] E. Hairer, C. Lubich, and G. Wanner. Geometric numerical integration illustrated by the Störmer–Verlet method. *Acta Numerica*, 12:399–450, 2003.

[8] E. Hairer, C. Lubich, and G. Wanner. *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*. Springer Series in Computational Mathematics, vol. 31, 2nd edition, 2006.

[9] B. Leimkuhler and S. Reich. *Simulating Hamiltonian Dynamics*, volume 14 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, 2004.

[10] J. E. Marsden and M. West. Discrete mechanics and variational integrators. *Acta Numerica*, 10:357–514, 2001.

[11] R. I. McLachlan and G. R. W. Quispel. Splitting methods. *Acta Numerica*, 11:341–434, 2002.

[12] H. Yoshida. Construction of higher order symplectic integrators. *Physics Letters A*, 150(5–7):262–268, 1990.

[13] S. Reich. Backward error analysis for numerical integrators. *SIAM Journal on Numerical Analysis*, 36(5):1549–1570, 1999.

[14] B. Derrida. Random-energy model: limit of a family of disordered models. *Physical Review Letters*, 45:79–82, 1980.

[15] B. Derrida. Random-energy model: an exactly solvable model of disordered systems. *Physical Review B*, 24:2613–2626, 1981.

[16] C. Sanford, D. Hsu, and M. Telgarsky. One-layer transformers fail to solve the induction heads task. *arXiv:2408.14332*, 2024.

[17] C. Sanford, D. Hsu, and M. Telgarsky. Representational strengths and limitations of transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[18] B. L. Edelman, S. Goel, S. Kakade, and C. Zhang. Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning (ICML)*, 2022.

[19] B. Liu, J. T. Ash, S. Goel, A. Krishnamurthy, and C. Zhang. Transformers learn shortcuts to automata. In *International Conference on Learning Representations (ICLR)*, 2023.

[20] C. Olsson, N. Elhage, N. Nanda, et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022.

[21] N. Elhage, N. Nanda, C. Olsson, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

[22] A. Conmy, A. Mavor-Parker, A. Lynch, S. Heimersheim, and A. Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *NeurIPS*, 2023.

[23] A. Bietti, V. Cabannes, D. Bouchacourt, H. Jegou, and L. Bottou. Birth of a transformer: a memory viewpoint. In *NeurIPS*, 2023.

[24] G. Reddy. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. In *ICLR*, 2024.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[26] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*, 2015.

[27] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[28] O. Press, N. A. Smith, and M. Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *ICLR*, 2022.

[29] T. Ye, L. Dong, Y. Xia, Y. Sun, Y. Zhu, G. Huang, and F. Wei. Differential transformer. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.05258.

[30] M. Xu, B. Wang, and W. Chen. KV shifting attention enhances language modeling. In *International Conference on Machine Learning (ICML)*, PMLR 267:69725–69748, 2025. arXiv:2411.19574.

[31] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *ICML*, 2020.

[32] A. Gu and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Conference on Language Modeling (COLM)*, 2024.

[33] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020.

[34] S. Garg, D. Tsipras, P. Liang, and G. Valiant. What can transformers learn in-context? a case study of simple function classes. In *NeurIPS*, 2022.

[35] E. Akyürek, D. Schuurmans, J. Andreas, T. Ma, and D. Zhou. What learning algorithm is in-context learning? investigations with linear models. In *ICLR*, 2023.

[36] J. von Oswald, E. Niklasson, E. Randazzo, J. Sacramento, A. Mordvintsev, A. Zhmoginov, and M. Vladymyrov. Transformers learn in-context by gradient descent. In *ICML*, 2023.

[37] Y. Bai, F. Chen, H. Wang, C. Xiong, and S. Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *NeurIPS*, 2023.

[38] S. Greydanus, M. Dzamba, and J. Yosinski. Hamiltonian neural networks. In *NeurIPS*, 2019.

[39] M. Cranmer, S. Greydanus, S. Hoyer, P. Battaglia, D. Spergel, and S. Ho. Lagrangian neural networks. In *ICLR Workshop on Integration of Deep Neural Models and Differential Equations*, 2020.

[40] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud. Neural ordinary differential equations. In *NeurIPS*, 2018.

[41] Z. Chen, J. Zhang, M. Arjovsky, and L. Bottou. Symplectic recurrent neural networks. In *ICLR*, 2020.

[42] S. Saemundsson, A. Terenin, K. Hofmann, and M. P. Deisenroth. Variational integrator networks for physically structured embeddings. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

[43] P. Jin, Z. Zhang, A. Zhu, Y. Tang, and G. E. Karniadakis. SympNets: Intrinsic structure-preserving symplectic networks for identifying Hamiltonian systems. *Neural Networks*, 132:166–179, 2020.

[44] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv:2001.08361*, 2020.

[45] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, et al. An empirical analysis of compute-optimal large language model training. In *NeurIPS*, 2022. arXiv:2203.15556.

[46] A. Power, Y. Burda, H. Edwards, I. Babuschkin, and V. Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv:2201.02177*, 2022.

[47] A. Giorlandino and S. Goldt. Two failure modes of deep transformers and how to avoid them: a unified theory of signal propagation at initialisation. In *ICLR*, 2026. *arXiv:2505.24333*.

[48] S. Chen, Z. Lin, Y. Polyanskiy, and P. Rigollet. Critical attention scaling in long-context transformers. In *ICLR*, 2026. *arXiv:2510.05554*.

[49] B. Gahtan and A. M. Bronstein. Coupled query-key dynamics for attention. *arXiv:2604.01683v1*, 2026. (Version 2 is retitled "Invertible query-key coupling composes with attention mechanisms".)

[50] V. Stein, W. Li, and G. Steidl. SympFormer: Accelerated attention blocks via inertial dynamics on density manifolds. *arXiv:2603.16535*, 2026.

[51] J. Xiong, L. Fan, H. Shen, Z. Su, M. Yang, L. Kong, and N. Wong. DoPE: Denoising rotary position embedding. *arXiv:2511.09146*, 2025.

[52] K. Maitra. Momentum attention: the physics of in-context learning and spectral forensics for mechanistic interpretability. *arXiv:2602.04902*, 2026. Early preprint of this work. https://arxiv.org/abs/2602.04902.

[53] K. Maitra, S. Sood, M. Hosseini, S. Gunnala, and V. Gupta. Phase space attention: a Hairer lift circumvents the single-layer induction obstruction. *arXiv:2609.32319*, 2026. Extended version of this paper. https://arxiv.org/abs/2609.32319.

Technical Appendices

Organisation. The appendices make the paper self-contained. **Part T** (Sections T1–T9) is the extended version of the theory in Sections 2–6: every statement of the main text is restated there with its full proof, together with the supporting results the page limit excludes (the design-space eliminations, the spectral and Fisher-geometric analysis of the softmax, the reduced score model in full, and the corollaries delimiting the mechanism), the complete empirical section, the related work in detail, and the derivations referenced from the main text. Results in Part T carry their own numbering (Theorem T$n.m$); the correspondence with the main text is: Proposition 1 ↔ Proposition T1.1; Lemma 2 ↔ Lemma T2.1; Theorem 4 ↔ Theorem T3.18; Proposition 5 ↔ Propositions T3.27 and T3.25; Theorem 6 ↔ Theorem T4.1; Corollary 7 ↔ Corollary T4.4; Theorem 8 ↔ Definition T5.20 and Theorem T5.22; Corollary 9 ↔ Corollaries T5.26 and T5.27; Proposition 10 ↔ Proposition T8.3. **Part A** (Sections A1–A36) is the empirical programme in full: every measurement summarised in Section 7 is reported there with its protocol, tables and figures, and each section names the notebook that produced it. The notebooks themselves, with outputs embedded, and the certificate notebook are provided as supplementary material.

Material added after review. Part T reproduces the extended version of this paper, posted as arXiv:2609.32319 [53]; the first author's earlier preprint arXiv:2602.04902 [52] is an early, partial version of the work. Besides the corrections and controls developed during the review period and summarised in Section 8, it contains three items that post-date the review and are not used by any claim of the main text: the deployability derivation (Section T7), the depth-and-coupling mechanism study on randomised-offset induction (Section T6.14), and the addendum that retains the first-order T_3 channel in the reduced model (Section T9), together with the 91.3 M-parameter sweeps of Sections A31.6–A31.11. They are included for completeness and are labelled as such where they appear; the main text refers to them only to state limitations.

The trade-off between accuracy and structure. Table 2 collects, for the four members of the order lattice, the measured accuracy and the structure each carries.

Table 2: The trade-off, stated once. Accuracies are the maximum over the coupling grid, so each is a best-of-k statistic and is upward biased relative to a value selected on held-out data; the comparison between variants is fair because every variant is scored by the same rule on the same grid. Protocol as in §T6.5 (three seeds, chance 0.071; Fig. 7 and Table 8). Every entry in the left column is exactly symplectic; symplecticity does not separate them (Prop. T3.27).

Operator	Order	Accuracy	What it carries
Query-only $(\gamma, 0)$	$(1, 0)$	0.033	Nothing: wrong axis; capped at $1/(T-2) = 0.077$, and below chance because the previous-token position takes the argmax (Cor. T5.26, Rem. T1.2)
Key-only $(0, \gamma)$	$(0, 1)$	**0.949**	Representability, and the best accuracy in the battery. Its own transition law (Cor. T5.27)
Asymmetric (γ_q, γ_k)	$(1, 1)^*$	0.875	Symplectic and energy-preserving, but an exact score bias and no transport identity (Prop. T3.27)
Symmetric (γ, γ)	$(1, 1)$	0.811	Uniqueness (Thm. T3.18); transport identity (Prop. T3.25); the γ^2 channel and the two-branch law of the reduced model (Thm. T5.22); self-adjoint higher-order completion; Differential Transformer at first order (Prop. T8.3)

Part T: Extended theory, with full proofs

T1 The single-layer induction obstruction as a filter-order gap

This section restates the SHT obstruction as a statement about attainable polynomial degree in a two-variable transfer function. The restatement is strictly weaker than SHT's theorem and internal to the bilinear class: its purpose is diagnostic. It identifies *which* structural feature the class lacks, and that feature is then exactly what the augmentation of §T3 supplies.

T1.1 The two-variable transfer function

Fix a transformer in the SHT hypothesis class $\mathcal{T}(h, m, p)$: h heads, embedding precision m bits, parameter precision p. For a single representative head, the unnormalised attention score is $s_{t,j} = Q_t^\top K_j$ with $Q_t = W_Q\, q_t$ and $K_j = W_K\, k_j$. The score is a function of two position indices, so the natural object is the joint z-transform in the pair (t, j), taken along the query axis with variable z_1 and the key axis with variable z_2. Because W_Q and W_K act pointwise in the position index — they mix embedding coordinates, never positions — the transform factorises:

$$H_{\mathfrak{S}}(z_1, z_2) = H_q(z_1)\, H_k(z_2), \qquad \deg_{z_1^{-1}} H_q = \deg_{z_2^{-1}} H_k = 0. \tag{T1.1}$$

We call the pair $(\deg_{z_1^{-1}} H_q,\, \deg_{z_2^{-1}} H_k)$ the *joint filter order* of the head, and write it (ρ, ρ'). Standard attention is order $(0, 0)$.

The degree cannot be raised by widening the head or by summing heads: a finite sum of products of degree-zero factors is again degree zero in each variable, polynomial degrees being subadditive under multiplication and stable under summation. Nor can it be raised by enlarging W_Q or W_K, since those act within a position. Order $(0, 0)$ is a property of the architecture, not of its parameter count, and this is the sense in which the obstruction is structural.

What the filter order is a property of. One clarification is needed before the count is used, because two descriptions of the same substrate are in play. The factorisation (T1.1) is a statement about maps acting *pointwise in the position index* on the post-RoPE streams: W_Q and W_K mix embedding coordinates within a position and move nothing between positions, so the lag structure of the score is exactly the lag structure of the two per-stream pre-filters. RoPE itself is not such a map — it makes the score a genuine function of $j - t$, which is why the shear and the rotation fail to commute (Cor. T4.4) — but its relative-position coupling lives *inside* the substrate on which the order is counted, and is orthogonal to the lag structure being counted. The joint order is therefore a property of the hypothesis class of pointwise-in-position maps of the post-RoPE stream, and not of the positional encoding: standard attention is order $(0, 0)$ in that class, and no choice of RoPE frequency changes it. This is also what gives Corollary T4.4 something to attach to. The order count is blind to the rotation; the Coriolis term is not.

T1.2 What the induction discriminator requires

The induction task supplies a sequence containing (token, target) pairs and a query token A; the target is the token following the earlier occurrence of A. The discriminating indicator is $\mathbf{1}\{\sigma_{j-1} = \sigma_t\}$, which depends on the key-side *pair* (σ_{j-1}, σ_j) and on the query-side token σ_t. Under the embedding maps $\sigma \mapsto k_\sigma$ and $\sigma \mapsto q_\sigma$, dependence on the pair is dependence on k_j and its first backward difference Δk_j, that is, a z_2^{-1} term in H_k. The following proposition makes the consequence of its absence precise.

Proposition T1.1 (Filter-order reading of SHT)**.** *Let Σ be a token alphabet with $|\Sigma| \geq 2$, embeddings drawn i.i.d. zero-mean with covariance $\sigma^2 I_{d_k}$. Consider the canonical induction sequence model: the induction target $j^* \in [1, T-1]$ has predecessor token equal to the query token A; distractors $j \neq j^*$ have own token drawn independently from Σ and predecessor token drawn independently from $\Sigma \setminus \{A\}$, so that j^* is the unique index whose predecessor is the query token. If $s_{t,j} = \langle f(q_t), g(k_j)\rangle$ has joint order $(0, 0)$ — that is, g depends on the key stream only through k_j — then conditional on q_t the sub-family $\{s_{t,j}\}_{j \neq j^*-1}$ is exchangeable, and consequently*

$$\Pr\big[\mathrm{softmax}_j(s_{t,\cdot}) \text{ peaks at } j^*\big] \;\leq\; \tfrac{1}{T-2} \tag{T1.2}$$

for every f and g, at every embedding dimension, and at every precision. The bound is attained exactly when the head suppresses the previous-token position $j^* - 1$, and the probability falls below $1/(T-1)$ whenever the head instead aligns with it.

Remark T1.2 (Why the constant is $1/(T-2)$ and not $1/(T-1)$). The position $j^* - 1$ is not exchangeable with the rest of the candidate set: its token is pinned to the query token A by the definition of the induction target, so $k_{j^*-1} = k_A$ is a deterministic function of σ_t while every other k_j is a free draw. A key-side order-zero filter cannot see the pair (σ_{j-1}, σ_j), but it *can* see that one candidate carries the query's own token, and it may exploit that either way. A head anti-aligned with the previous-token direction suppresses $j^* - 1$ outright and leaves j^* competing against $T-3$ distractors, attaining $1/(T-2)$; a head aligned with it hands the argmax to $j^* - 1$ and drives the peak probability towards zero. Neither reaches induction, and the ceiling is $O(1/T)$ with no dependence on d_k or precision — which is the content of the proposition. An earlier statement of this result asserted equality at $1/(T-1)$; that value is the interior case in which the previous-token channel happens to be neutral, not a universal bound. §T6.5 measures the aligned case.

Proof. Condition on q_t. Among the candidates $j \in [1, T-1]$, exactly one — the position $j^* - 1$ — has its token fixed by the model, since $\sigma_{j^*-1} = A = \sigma_t$; every other token, including σ_{j^*} itself, is drawn independently. The keys $\{k_j\}_{j \neq j^*-1}$ are therefore an i.i.d. family, and a score depending on the key stream only through a pointwise function of k_j is a measurable function of it, so $\{s_{t,j}\}_{j \neq j^*-1}$ is exchangeable. The index j^* is determined by the predecessor relation, a property of the *joint* law of (σ_{j-1}, σ_j), and is independent of that sub-family; hence j^* is equally likely to occupy any of its $T-2$ sorted positions and $\Pr[j^* \text{ maximises the sub-family}] = 1/(T-2)$. Since peaking at j^* requires in addition that $s_{t,j^*} > s_{t,j^*-1}$, the stated bound follows, with equality precisely when that second event has conditional probability one. The discriminating information $\sigma_{j^*-1} = A$ remains invisible to a filter of key-side order zero, because that filter never forms a function of two consecutive key positions: the previous-token position can be identified and suppressed, but the token it precedes cannot be read. □

Remark T1.3 (Relation to SHT). Proposition T1.1 is weaker than SHT's theorem and does not reproduce it. SHT establish a lower bound on per-head description size by a counting and communication argument over a class of scores; Proposition T1.1 establishes an $O(1/T)$ ceiling on induction accuracy within the bilinear class under a specific generative model. Its role is to locate the deficit, and it locates it on the key axis. Correspondingly, the augmentation we introduce does not escape SHT's bound: the PSA head's own bit budget remains $\mathcal{O}(d_k)$, independent of T, and the score it forms is not of the form the counting argument enumerates. The premise is exited, not the bound.

T1.3 The order lattice

The minimum increment that makes the joint law of (σ_{j-1}, σ_j) accessible is a single z_2^{-1} term on the key side. This gives four candidate orders and, with the DC-neutrality and high-pass requirements of Theorem T3.18, a corresponding lattice of augmentations:

$$(\rho, \rho') \in \{(0,0),\ (1,0),\ (0,1),\ (1,1)\}, \quad H_q(z_1) = 1 + \gamma_q(1 - z_1^{-1}),\ \ H_k(z_2) = 1 + \gamma_k(1 - z_2^{-1}), \tag{T1.3}$$

with $\rho = \mathbf{1}\{\gamma_q > 0\}$ and $\rho' = \mathbf{1}\{\gamma_k > 0\}$. Each member has a distinct and falsifiable prediction, and §T6 measures all four under a single matched protocol.

- $(0,0)$ is standard attention: capped at the order-zero ceiling $1/(T-2)$, by Proposition T1.1.
- $(1,0)$, the query-only half-lift, leaves the key stream order-zero, so the exchangeability argument of Proposition T1.1 applies unchanged: the $1/(T-2)$ ceiling binds for every γ_q, and §T6.5 measures the aligned case below chance. It is a first-order rule applied to the wrong axis, and it is the discriminating negative control — leaving the $(0,0)$ class is necessary but not sufficient.
- $(0,1)$, the key-only half-lift, meets the minimum. It is DC-neutral and high-pass, and it makes the task representable. Corollary T5.27 gives its transition law. It is also, measured under a matched protocol, the most accurate member of the lattice on the induction task — more accurate than the symmetric completion (§T6.5, Fig. 7, Table 2). Order, not symmetry, is what representability requires.
- $(1,1)$ is the swap-symmetric completion, and the subject of the rest of the paper.

Two features of this lattice deserve emphasis because they are easily conflated. Order and *symmetry* are independent: the minimal representable member is $(0,1)$, and $(1,1)$ is the unique member of the lattice invariant under exchanging the two streams (Theorem T3.18). And order and *symplecticity* are also independent: every member of (T1.3) is exactly symplectic on the joint phase space, including the asymmetric ones (Proposition T3.27). Neither symmetry nor symplecticity is what makes the task representable; key-side order ≥ 1 is. What symmetry buys is the transport identity of Proposition T3.25 and the closed-form theory that follows from it, and we are explicit throughout that this is a structural rather than a performance argument.

Corollary T3.5 verifies that the $(1,1)$ member attains joint order exactly $(1,1)$, closing the gap Proposition T1.1 identifies. The $(1,1)$ member is realised by the upper shear $(q,p) \mapsto (q+\gamma p, p)$ with p the first backward difference, applied identically to both streams. §T3 establishes that within the factorised subclass this is the only way to close the gap while preserving swap symmetry, and §T2 shows that the same shear arises canonically from a parallel obstruction in numerical integration.

T2 A parallel obstruction in numerical integration

The obstruction of §T1 has an exact counterpart in the numerical integration of Newton's law, and the counterpart comes with a resolution that has been standard for four decades. This section sets out the classical statements in the form we need. They are statements about iterated maps; §T4.2 determines which of them may be carried across, and the answer is: the algebra, and not the long-time dynamics.

T2.1 The stencil obstruction

Consider $\ddot{q}(t) = f(q(t))$ with $q \in \mathbb{R}^{d_k}$, discretised at step Δt. Forward Euler,

$$q_{n+1} = q_n + \Delta t\, v_n, \qquad v_{n+1} = v_n + \Delta t\, f(q_n), \tag{T2.1}$$

is first order: the local truncation error is $\mathcal{O}(\Delta t^2)$ and the global configuration error $\mathcal{O}(\Delta t)$.

Secular energy growth on a quadratic potential. The failure of (T2.1) is sharper than its order suggests, and the sharpest instance is the linear oscillator — the quadratic potential $V(q) = \frac{1}{2}\|q\|^2$, $f(q) = -q$ — where everything can be computed in closed form. Writing $x_n = (q_n, v_n)$, the update is $x_{n+1} = A_{\mathrm{E}} x_n$ with

$$A_{\mathrm{E}} = \begin{pmatrix} I & \Delta t\, I \\ -\Delta t\, I & I \end{pmatrix}, \qquad A_{\mathrm{E}}^{\top} A_{\mathrm{E}} = (1+\Delta t^2)\, I_{2d_k}, \tag{T2.2}$$

so A_{E} is exactly $\sqrt{1+\Delta t^2}$ times an orthogonal matrix. The energy $E_n = \frac{1}{2}\|v_n\|^2 + \frac{1}{2}\|q_n\|^2 = \frac{1}{2}\|x_n\|^2$ therefore satisfies the identity

$$\frac{E_n}{E_0} = \big(1+\Delta t^2\big)^n, \qquad \text{equivalently} \qquad \frac{\|x_n\|}{\|x_0\|} = \big(1+\Delta t^2\big)^{n/2}, \tag{T2.3}$$

for every initial condition and with no error term. The growth is *secular*: it is not an oscillation about the correct value but a monotone drift, exponential in the number of steps at fixed Δt, and the secular drift persists at every fixed non-zero step, although at a fixed physical horizon it decreases linearly in Δt — reducing Δt by a factor of two while integrating to the same physical horizon doubles n, and substituting $n = t_{\mathrm{end}}/\Delta t$ gives

$$\frac{E_n}{E_0} = \big(1+\Delta t^2\big)^{t_{\mathrm{end}}/\Delta t} = \exp\Big(\tfrac{t_{\mathrm{end}}}{\Delta t}\log\big(1+\Delta t^2\big)\Big) \;=\; \exp\big(\Delta t\, t_{\mathrm{end}} + \mathcal{O}(\Delta t^3)\big), \tag{T2.4}$$

using $\log(1+x) = x + \mathcal{O}(x^2)$ and, as above, $K = I$ so that the oscillator frequency is unity and $\Delta t\, t_{\mathrm{end}}$ is dimensionless. The energy error at a fixed horizon is therefore $e^{\Delta t\, t_{\mathrm{end}}} - 1 = \Delta t\, t_{\mathrm{end}} + \mathcal{O}(\Delta t^2)$: first order in the step size, which is the order of the method, and no better. Note that $\det A_{\mathrm{E}} = (1+\Delta t^2)^{d_k} \neq 1$: forward Euler is not even volume-preserving on this problem, let alone symplectic.

The Verlet stencil and why it resists one-step realisation. The Störmer–Verlet stencil

$$q_{n+1} - 2q_n + q_{n-1} = \Delta t^2 f(q_n) \tag{T2.5}$$

is second order and symmetric (invariant under $\Delta t \mapsto -\Delta t$ with the time direction reversed), and on quadratic potentials it conserves a modified energy exactly and the true energy to within $\mathcal{O}(\Delta t^2)$ with no drift over exponentially long times [12, 10, 11, 3]. But as a recurrence in the configuration variable alone it is *two-step*: computing q_{n+1} requires both q_n and q_{n-1}. This is the obstruction. A one-step linear map on configuration space, $q_{n+1} = \Lambda q_n$, has a transfer function of degree zero in z^{-1} and cannot reproduce the q_{n-1}-dependence of (T2.5); the parallel with (T1.1) is exact, and the two obstructions are parallel algebraic statements about attainable degree — instances of one state-augmentation motif rather than a single theorem.

T2.2 The Hairer lift

Hairer et al. [7, 8] circumvent the stencil obstruction by extending the state to phase space (q, p) with $p_n = (q_n - q_{n-1})/\Delta t$, on which Verlet has the one-step symplectic decomposition

$$p_{n+1/2} = p_n + \tfrac{\Delta t}{2} f(q_n), \qquad q_{n+1} = q_n + \Delta t\, p_{n+1/2}, \qquad p_{n+1} = p_{n+1/2} + \tfrac{\Delta t}{2} f(q_{n+1}), \tag{T2.6}$$

the "kick–drift–kick" palindrome. The symplectic two-form $\omega = dq \wedge dp$ is exactly preserved at each step. The two-step dependence has not disappeared; it has been *absorbed into the state*. The lagged information that (T2.5) carried in q_{n-1} is carried in p_n, and the map on the extended state is one-step.

In the free-particle limit $f \equiv 0$ the two kicks vanish and (T2.6) reduces to the drift step alone, the linear shear M_γ with $\gamma = \Delta t$:

$$M_\gamma = \begin{pmatrix} I & \gamma I \\ 0 & I \end{pmatrix}, \qquad M_\gamma^\top J M_\gamma = J, \quad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}. \tag{T2.7}$$

Lemma T2.1 (Exact γ-flow). *M_γ is the exact time-γ flow of the free-particle Hamiltonian $H(q,p) = \frac{1}{2}\|p\|^2$, with all truncation orders vanishing identically.*

Proof. $\dot q = \partial_p H = p$, $\dot p = -\partial_q H = 0$; integrating from 0 to γ gives $(q + \gamma p, p) = M_\gamma(q,p)$. Since $\dot p \equiv 0$ the solution is a polynomial of degree one in the time variable and the Taylor expansion of the flow terminates: the shear is not a first-order approximation to the flow, it is the flow. □

Remark T2.2 (Why M_γ and not some other one-step map). Three properties canonicalise the shear among linear one-step maps on the extended state, and all three are used later. It is the exact flow of a Hamiltonian rather than an approximation (Lemma T2.1), so its symplecticity is exact at every γ and not asymptotic in small γ. It raises the attainable order in the base variable from zero to one — the composite map has a z^{-1} term in q that no one-step map on q alone possesses — which is precisely the deficit identified in §T1. And it is the drift half of a self-adjoint palindrome, which is what distinguishes the symmetric composition from the Euler-class alternatives and reappears on the attention side as Proposition T3.25.

T2.3 Backward error analysis and the quadratic potential

This subsection develops the classical long-time theory in the detail the later scoping argument requires. Everything in it is a statement about *iterated integrators*; §T4.2 determines which parts have an attention-side analogue, and the answer will be: the algebraic collapse below, and none of the long-time consequences.

Poisson brackets and the modified Hamiltonian. For $F, G \in C^\infty(\mathcal{M})$ on a symplectic manifold $(\mathcal{M}, \Omega)$ with coordinates (q, p), the Poisson bracket is

$$\{F, G\} := \sum_{i=1}^{d_k} \Big(\frac{\partial F}{\partial q_i}\frac{\partial G}{\partial p_i} - \frac{\partial F}{\partial p_i}\frac{\partial G}{\partial q_i} \Big), \tag{T2.8}$$

and the Hamiltonian vector field of F is $X_F = \Omega^{-1}\nabla F = (\partial_p F, -\partial_q F)$. Hamiltonian vector fields and Poisson brackets form isomorphic Lie algebras up to sign,

$$[X_F, X_G] = -X_{\{F,G\}}, \tag{T2.9}$$

so composition of Hamiltonian flows obeys the Baker–Campbell–Hausdorff formula

$$\phi_F^h \circ \phi_G^h = \phi_{Z(h)}^1,$$
$$Z(h) = hX_F + hX_G + \tfrac{h^2}{2}[X_F, X_G] + \tfrac{h^3}{12}\big([X_F,[X_F,X_G]] - [X_G,[X_G,X_F]]\big) + \cdots \tag{T2.10}$$

For a one-step integrator Φ_h applied *repeatedly to its own output*, the modified Hamiltonian $\widetilde{H}_h$ is the formal power series in h whose exact flow interpolates the numerical trajectory to all orders; for a symplectic near-identity integrator the formal modified vector field can be chosen Hamiltonian, and under standard analyticity and compactness hypotheses this yields the long-time near-conservation result [7, 8, 3]. Near-conservation of $\widetilde{H}_h$ over times exponentially long in $1/h$ is what delivers bounded energy error for symplectic methods, and every hypothesis in that chain refers to the iterate.

Worked example: BCH corrections for a quadratic potential. Take $H(q,p) = \frac{1}{2}p^\top p + \frac{1}{2}q^\top Kq$ with $K \succ 0$ — the multidimensional harmonic oscillator — and split it as $T(q,p) = \frac{1}{2}p^\top p$ and $V(q,p) = \frac{1}{2}q^\top Kq$. The drift flow $\Phi_h^T : (q,p) \mapsto (q + hp,\, p)$ is the upper shear, generated by $X_T = (p, 0)$; the kick flow $\Phi_h^V : (q,p) \mapsto (q,\, p - hKq)$ is the lower shear, generated by $X_V = (0, -Kq)$. Störmer–Verlet is the palindromic composition $\Phi_h^{\mathrm{Verlet}} = \Phi_{h/2}^V \circ \Phi_h^T \circ \Phi_{h/2}^V$. The relevant brackets are

$$\{T,V\} = -p^\top Kq, \qquad \{T,\{T,V\}\} = p^\top Kp, \qquad \{V,\{V,T\}\} = q^\top K^2 q, \tag{T2.11}$$

and (T2.10) gives, for the symmetric splitting, a modified Hamiltonian even in h,

$$\begin{aligned} \widetilde{H}_h &= T + V + h^2 H_2 + h^4 H_4 + \cdots, \\ H_2 &= -\tfrac{1}{8}\{V,\{V,T\}\} = -\tfrac{1}{8}\, q^\top K^2 q. \end{aligned} \tag{T2.12}$$

The normalisation needs a word, because $\widetilde{H}_h$ is not unique. On a quadratic potential the map is linear and its conserved quadratic form is unique only up to scale, so H_2 is fixed only modulo an added multiple of H: the admissible family is $H_2 + \mu(T+V)$ for arbitrary μ. Equation (T2.12) takes $\mu = 0$, the representative with no $p^\top Kp$ correction. With that choice the conserved form is

$$\widetilde{H}_h = \tfrac{1}{2}\|p\|^2 + \tfrac{1}{2}\, q^\top K\Big(I - \tfrac{h^2}{4}K\Big)q, \tag{T2.13}$$

which the Störmer–Verlet map preserves *exactly* on the quadratic potential — not merely to $\mathcal{O}(h^4)$ — as the 2×2 eigenvalue computation confirms. The corrections are therefore genuinely non-zero: the symmetric integrator conserves $\widetilde{H}_h$ rather than H, and the true energy oscillates about its initial value with amplitude $\mathcal{O}(h^2)$ and no drift. Bounded oscillation additionally requires the linear stability condition $h^2\lambda_{\max}(K) < 4$; beyond it (T2.13) ceases to be positive definite and the trajectory diverges.

Proposition T2.3 (Collapse of the expansion in the free case)**.** *If $V \equiv 0$ then every bracket in (T2.11) vanishes identically, every correction H_k with $k \geq 1$ vanishes, and the expansion collapses to $\widetilde{H}_h = T = H$ exactly, at every order and for every h.*

Proof. With $V \equiv 0$ the Hamiltonian vector field X_V is identically zero, so every iterated bracket containing V vanishes. Every term of (T2.10) beyond hX_T contains at least one factor X_V, hence $Z(h) = hX_T$ and $\widetilde{H}_h = T$. Equivalently and without any appeal to BCH: by Lemma T2.1 the drift step *is* the exact flow of T, so the numerical and exact trajectories coincide and the interpolating Hamiltonian is T itself. □

Remark T2.4 (What Proposition T2.3 does and does not give). The collapse is an algebraic fact about the shear, and it is the reason Lemma T2.1 can be stated as an exactness rather than an approximation. It is *not* the source of any claim made later about attention. On the attention side there is no iterate to interpolate, so there is nothing for a modified Hamiltonian to be the interpolant of; §T4.2 makes this precise and Table 4 records it. The reason we present the quadratic-potential expansion at all is that it is the sharpest way to see what the free case is free *of*, and therefore the sharpest way to see what does not carry across.

Eigenvalue structure of the shear. One further algebraic property of M_γ is used in §T4.1 and worth recording here, because it settles what the spectrum does and does not say about iterating the map.

Proposition T2.5 (Spectrum of M_γ)**.** *Let $\gamma \neq 0$ and $d_k \geq 1$. The matrix $M_\gamma \in \mathbb{R}^{2d_k \times 2d_k}$ has the single eigenvalue $\lambda = 1$ with algebraic multiplicity $2d_k$ and geometric multiplicity d_k; it is non-diagonalisable, consisting of d_k Jordan blocks of size 2. At $\gamma = 0$ the map is the identity and is diagonalisable.*

Proof. $M_\gamma - I = \left(\begin{smallmatrix} 0 & \gamma I \\ 0 & 0 \end{smallmatrix}\right)$ is nilpotent of index two for $\gamma \neq 0$, so the characteristic polynomial is $(\lambda - 1)^{2d_k}$ and the algebraic multiplicity is $2d_k$. The eigenspace is $\ker(M_\gamma - I) = \{(q, 0)\}$, of dimension d_k, giving geometric multiplicity d_k. A matrix whose sole eigenvalue has algebraic multiplicity $2d_k$ and geometric multiplicity d_k with $(M_\gamma - I)^2 = 0$ has Jordan form consisting of d_k blocks of size 2. At $\gamma = 0$ the map is the identity and the statement degenerates to $2d_k$ blocks of size one. □

Remark T2.6 (The shear is not power-bounded). It is worth being exact about what Proposition T2.5 implies, because the spectrum invites a misreading. The spectral radius is $\rho(M_\gamma) = 1$, so there is no exponential growth. But the sole eigenvalue is *defective* — geometric multiplicity d_k against algebraic multiplicity $2d_k$ — and the standard criterion for a linear map to be power-bounded requires every eigenvalue on the unit circle to be semisimple. That criterion fails here. Explicitly,

$$M_\gamma^n = \begin{pmatrix} I & n\gamma I \\ 0 & I \end{pmatrix}, \qquad \|M_\gamma^n\|_2 \;\sim\; n|\gamma| \;\to\; \infty \quad (\gamma \neq 0), \tag{T2.14}$$

so $\sup_n \|M_\gamma^n\| = \infty$: in the Lyapunov sense the iterated map is *marginally unstable*, with growth linear in n rather than exponential. This is the classical picture of a free particle, whose position grows linearly in time at constant momentum, and it is exactly what one should expect of the exact flow of $H = \frac{1}{2}\|p\|^2$ (Lem. T2.1).

We record it because it forecloses an inference that Theorem T4.1 might otherwise seem to invite: exact symplecticity is not boundedness, and preserving Ω and H_{joint} at every application says nothing about the norm of an iterate. The theorem claims the former only, per application, and (T2.14) is one more reason the paper declines the long-time questions rather than answering them — the answer, had we iterated, would not have been favourable.

T2.4 The correspondence

The two obstructions and their resolutions align term by term: stencil order $(0) \to (2)$ in q on the numerical-integration side, and joint filter order $(0, 0) \to (1, 1)$ in (q, k) on the transformer side, both closed by the same operator M_γ applied to position $\to$ (position, momentum). In both cases the obstruction is a statement about what a one-step map on the base variable can represent, and in both cases the resolution is to change what the state is rather than to strengthen the map. Figure 1 traces the five steps of the correspondence.

Two remarks are needed before the transport is performed, and both are honoured in the sections that follow. First, the correspondence is between the *representability* arguments on the two sides. The numerical-integration literature also contains a large body of long-time dynamical results — backward error analysis, modified Hamiltonians, exponentially long energy bounds — which depend on hypotheses that the attention setting does not supply; §T4.2 states which is which, and Table 4 records the boundary explicitly. Second, the correspondence identifies an operator, not an accuracy ordering. The Euler-class member on the key stream solves the induction task, as §T6 measures; what distinguishes the symmetric Verlet-class member is structure, and we do not argue otherwise.

T3 Phase Space Attention as the Hairer lift

Let $\{q_t\}_{t=1}^T$, $\{k_j\}_{j=1}^T$ be the post-RoPE query and key streams. Define the kinematic momenta $p_{q,t} = q_t - q_{t-1}$, $p_{k,j} = k_j - k_{j-1}$ (with $p_{q,1} = p_{k,1} = 0$; see below).

Boundary convention. We set $p_{q,1} = p_{k,1} = 0$ and $p_{q,t} = q_t - q_{t-1}$, $p_{k,j} = k_j - k_{j-1}$ for $t, j \geq 2$. This is the convention Algorithm 2 implements, and it is adopted in preference to $q_0 = k_0 = 0$ — which would give $p_{q,1} = q_1$ — because the latter injects an artificial impulse at the first token equal to the token itself. The two agree for $t, j \geq 2$, so every transfer-function statement below, which concerns the interior of the stream, is unaffected; statements quantified over all $t, j \geq 1$ should be read with the first position carrying zero momentum.

Definition T3.1 (Phase Space Attention)**.** The Phase Space Attention score with momentum coupling $\gamma > 0$ is

$$s^{\mathrm{PSA}}_{t,j} := (q_t + \gamma p_{q,t})^\top (k_j + \gamma p_{k,j}) = q_t^\top k_j + \gamma\, p_{q,t}^\top k_j + \gamma\, q_t^\top p_{k,j} + \gamma^2\, p_{q,t}^\top p_{k,j}. \tag{T3.1}$$

The PSA attention weights are $\mathrm{softmax}_j(s^{\mathrm{PSA}}_{t,j}/\sqrt{d_k})$. Values V are *not* momentum-augmented (§T3.7).

The four-term decomposition isolates the standard score T_1, two single-ghost cross terms T_2 and T_3 each measuring alignment of one stream's position with the other stream's first difference, and the double-ghost term $T_4 = p_{q,t}^\top p_{k,j}$, which is large exactly when query and key are changing in the same direction in embedding space. Proposition T5.4 establishes that under the canonical induction sequence model T_3 carries a discriminative gap at first order in γ and T_4^{AA} carries one at second order; the reduced model of §T5.6 deliberately retains only the latter.

Algorithmic statement. PSA is a drop-in replacement for the score head: given post-RoPE query/key tensors and $\gamma > 0$, Algorithm 2 computes the output in $O(Td_k)$ extra time and memory, with no extra learnable parameters beyond the single scalar γ, which is swept as a hyperparameter in every experiment reported here and may equally be optimised (Rem. T4.3). The new operations are two backward differences and two shifted differences and two affine combinations.

Algorithm 2 Phase Space Attention forward pass (single head, post-RoPE).

Require: post-RoPE queries $Q \in \mathbb{R}^{T\times d_k}$, post-RoPE keys $K \in \mathbb{R}^{T\times d_k}$, values $V \in \mathbb{R}^{T\times d_v}$, coupling $\gamma \in \mathbb{R}_{>0}$, causal mask $\mathbf{M}$
Ensure: attention output $Y \in \mathbb{R}^{T\times d_v}$

1: $P_q \leftarrow Q - \mathrm{shift}_{+1}(Q)$ — *# first backward difference; $P_q[0,:] \leftarrow 0$*
2: $P_k \leftarrow K - \mathrm{shift}_{+1}(K)$ — *# first backward difference; $P_k[0,:] \leftarrow 0$*
3: $\hat{Q} \leftarrow Q + \gamma P_q$; $\hat{K} \leftarrow K + \gamma P_k$ — *# symplectic upper shear M_γ on each stream*
4: $S \leftarrow \hat{Q}\hat{K}^\top/\sqrt{d_k}$ — *# order-$(1,1)$ score, four-term decomp. (T3.1)*
5: $A \leftarrow \mathrm{softmax}_j(S + \mathbf{M})$ — *# $\mathbf{M}_{tj} = -\infty$ if $j > t$, else 0*
6: $Y \leftarrow AV$ — *# values are* not *momentum-augmented (§T3.7)*
7: **return** Y

The post-RoPE placement (lines 1–2) is essential and not interchangeable with pre-RoPE placement: Corollary T4.4 establishes the closed-form Coriolis penalty incurred by applying the shift before RoPE. Lines 1–3 are the entire architectural change relative to standard attention; lines 4–7 are unmodified.

T3.1 The four-term decomposition and the Ghost Key

Theorem T3.2 (Four-term score decomposition)**.** *For every $t, j \geq 1$ the PSA score of Definition T3.1 decomposes exactly as*

$$s^{\mathrm{PSA}}_{t,j} = \underbrace{q_t^\top k_j}_{T_1} + \gamma \underbrace{p_{q,t}^\top k_j}_{T_2} + \gamma \underbrace{q_t^\top p_{k,j}}_{T_3} + \gamma^2 \underbrace{p_{q,t}^\top p_{k,j}}_{T_4}. \tag{T3.2}$$

Proof. Bilinearity of the inner product gives $(a + \gamma b)^\top (c + \gamma d) = a^\top c + \gamma b^\top c + \gamma a^\top d + \gamma^2 b^\top d$; substitute $a = q_t$, $b = p_{q,t}$, $c = k_j$, $d = p_{k,j}$. No approximation is involved and the identity holds for every γ. □

The four terms have distinct roles, and the distinction is what the empirical programme of §T6 is built to test. T_1 is the standard attention score, oblivious to momentum. T_2 and T_3 are single-ghost

cross terms, each pairing one stream's position against the other stream's first difference. T_4 is the double-ghost coherence channel, large exactly when query and key are changing in the same direction in embedding space, which is the structural signature of pattern completion. Proposition T5.4 below establishes that under isotropy the discriminative gap is carried by T_3 at order γ and by the AC–AC component of T_4 at order γ^2.

Definition T3.3 (Ghost Key)**.** The augmented key satisfies

$$\hat{k}_j = (1+\gamma)k_j - \gamma k_{j-1}, \tag{T3.3}$$

and the term $-\gamma k_{j-1}$ is the *Ghost Key*: a representation of the predecessor token carried inside the key at position j.

The Ghost Key is the mechanism by which the premise of the counting argument is exited. The induction discriminator requires access to the pair (σ_{j-1}, σ_j); in a standard transformer the second attention layer exists to copy k_{j-1} into the residual stream at position j, and it is that copy which the SHT bound prices. Equation (T3.3) injects a fixed predecessor-dependent component into the score-bearing key without a second attention layer, as a linear functional of the stream the layer already has. It is a mixture, not an injective encoding: two d_k-dimensional vectors are not separately recoverable from one, and nothing below requires that they be. Nothing about the bound is contradicted: the head's own description length remains $\mathcal{O}(d_k)$, independent of T.

T3.2 Filter-order audit

Proposition T3.4 (Per-stream transfer function)**.** *The map $k_j \mapsto \hat{k}_j = (1+\gamma)k_j - \gamma k_{j-1}$ is a causal FIR filter with transfer function and squared magnitude response*

$$H^u(z) = (1+\gamma) - \gamma z^{-1}, \qquad \left|H^u(e^{j\omega})\right|^2 = 1 + 4\gamma(1+\gamma)\sin^2(\omega/2). \tag{T3.4}$$

Consequently (i) $|H^u(e^{j0})| = 1$*, unit DC gain;* (ii) $|H^u(e^{j\pi})| = 1 + 2\gamma$*, the Nyquist gain;* (iii) $|H^u(e^{j\omega})|$ *is strictly increasing on* $(0, \pi]$ *for every* $\gamma > 0$*, so* H^u *is high-pass.*

Proof. The z-transform of the recursion gives $Z[\hat{k}](z) = \big((1+\gamma) - \gamma z^{-1}\big) Z[k](z)$, which is H^u. On the unit circle,

$$\begin{aligned}\left|H^u(e^{j\omega})\right|^2 &= \left|(1+\gamma) - \gamma e^{-j\omega}\right|^2 = (1+\gamma)^2 - 2\gamma(1+\gamma)\cos\omega + \gamma^2 \\ &= 1 + 2\gamma + 2\gamma^2 - 2\gamma(1+\gamma)\cos\omega = 1 + 4\gamma(1+\gamma)\sin^2(\omega/2),\end{aligned}$$

using $(1+\gamma)^2 + \gamma^2 = 1 + 2\gamma + 2\gamma^2$ and $1 - \cos\omega = 2\sin^2(\omega/2)$. Setting $\omega = 0$ gives $|H^u|^2 = 1$; setting $\omega = \pi$ gives $1 + 4\gamma(1+\gamma) = (1+2\gamma)^2$. Monotonicity is immediate since $\gamma(1+\gamma) > 0$ and $\sin^2(\omega/2)$ is increasing on $(0, \pi]$. □

Corollary T3.5 (The gap is closed)**.** *For every $\gamma > 0$ the augmentation is a causal FIR filter of order exactly* 1 *on each stream, so the bilinear score has joint filter order* $(1, 1)$ *in* (t, j)*, up from* $(0, 0)$ *for standard attention.*

Proof. $H^u(z) = (1+\gamma) - \gamma z^{-1}$ has degree exactly one in z^{-1} whenever $\gamma \neq 0$, so each per-stream order is one; by bilinearity of the pairing the joint order is the pair of the two per-stream orders. Together with Proposition T1.1, which caps order $(0, 0)$ at $1/(T-2)$, this identifies precisely what the augmentation supplies. □

T3.3 Spectral consequences of the shear

The transfer function of Proposition T3.4 has three consequences that are used later and are worth isolating: a frequency-domain form of the score, a quantitative statement of how much energy the augmentation moves into the AC band, and an exact invariance of the momentum coordinate that is often misquoted as a Parseval identity for M_γ and is not one.

Proposition T3.6 (Fourier-domain form of the score)**.** *Fix the two-dimensional transform convention* $\tilde{s}(\omega_1, \omega_2) := \sum_{t,j} s_{t,j}\, e^{-j(\omega_1 t + \omega_2 j)}$ *and the per-stream transforms* $\tilde{q}(\omega) = \sum_t q_t e^{-j\omega t}$*,* $\tilde{k}(\omega) = \sum_j k_j e^{-j\omega j}$*. Under the Hermitian pairing* $s_{t,j} = q_t^{\mathsf{H}} k_j$ *the score transform is the* outer *bilinear*

product $\tilde{s}(\omega_1, \omega_2) = \tilde{q}(\omega_1)^{\mathsf{H}}\, \tilde{k}(\omega_2)$, with no convolution between the streams: the two indices are independent. Filtering each stream by H^u then gives

$$\tilde{\hat{s}}(\omega_1, \omega_2) = \overline{H^u(e^{j\omega_1})}\; H^u(e^{j\omega_2})\, \tilde{q}(\omega_1)^{\mathsf{H}}\, \tilde{k}(\omega_2), \tag{T3.5}$$

the conjugation falling on the query response because that stream enters the pairing conjugated, so the score spectrum is reweighted by the product of the two per-stream responses, amplifying the joint high-frequency band most.

Proof. Convolution in the index domain is multiplication in the transform domain, so the per-stream statement is immediate from Proposition T3.4. Taking the transform of $\hat{s}_{t,j} = \hat{q}_t^{\top} \hat{k}_j$ separately in each index and using bilinearity gives (T3.5); the conjugate on the query factor is the usual convention for the cross-spectrum. □

Theorem T3.7 (Band-averaged power gain)**.** *Let H^u be the per-stream transfer function of Proposition T3.4. Then*

(i) (Exact mean-square gain.) *The band-averaged power gain over $[0, \pi]$ is*

$$\frac{1}{\pi} \int_0^{\pi} \left| H^u(e^{j\omega}) \right|^2 d\omega \;=\; 1 + 2\gamma(1+\gamma), \tag{T3.6}$$

exactly and for every γ: the augmentation multiplies the band-averaged power by a factor quadratic in the coupling.

(ii) (All of the excess is above DC.) *The gain at $\omega = 0$ is exactly 1 for every γ* (Prop. T3.4(i))*, so the entire excess $2\gamma(1+\gamma)$ is carried at non-zero frequency.*

(iii) (The excess is high-weighted.) *For every $\omega_0 \in (0, \pi)$ the fraction of the excess carried on $[\omega_0, \pi]$ strictly exceeds the fraction of the band it occupies, $(\pi - \omega_0)/\pi$, and the fraction is independent of γ.*

Proof. **(i)** By Proposition T3.4, $|H^u(e^{j\omega})|^2 = 1 + 4\gamma(1+\gamma)\sin^2(\omega/2)$. Since $\frac{1}{\pi}\int_0^{\pi} \sin^2(\omega/2)\, d\omega = \frac{1}{2}$, the average is $1 + 4\gamma(1+\gamma)\cdot\frac{1}{2}$, which is (T3.6). **(ii)** is Proposition T3.4(i). **(iii)** The excess density is $4\gamma(1+\gamma)\sin^2(\omega/2)$, whose γ-dependence is a common positive factor and therefore cancels from any ratio of its integrals. The claim reduces to $\int_{\omega_0}^{\pi} \sin^2(\omega/2)\, d\omega \big/ \int_0^{\pi} \sin^2(\omega/2)\, d\omega > (\pi - \omega_0)/\pi$, which holds because $\sin^2(\omega/2)$ is strictly increasing on $[0, \pi]$, so its mean over $[\omega_0, \pi]$ strictly exceeds its mean over $[0, \pi]$. □

Remark T3.8 (Why the statement is framed by averages rather than a cutoff). A band-split formulation — fixing a 3 dB cutoff ω_c and comparing energy above and below it — is the more familiar way to state a high-pass claim, and it is worth recording why it is not used here. The cutoff is convention-dependent: measured 3 dB above the DC gain it exists only for $\gamma \geq (\sqrt{2} - 1)/2 \approx 0.207$, since below that the response never reaches twice its DC value, while measured 3 dB below the passband maximum it exists for all γ but moves with the coupling. Worse, the resulting ratio does not grow quadratically: at a fixed split the AC-to-DC ratio saturates as $\gamma \to \infty$, because both bands scale with the same factor $4\gamma(1+\gamma)$. The quadratic growth is a statement about *total* added power, (T3.6), not about a ratio between bands, and (iii) is the correct sense in which that power is concentrated at high frequency.

Proposition T3.9 (Pointwise invariance of the momentum coordinate)**.** *Let $\{(q_n, p_n)\}$ be the orbit of (q_0, p_0) under iteration of M_γ. Then $p_n = p_0$ pointwise for every n, hence $\tilde{p}_n = \tilde{p}_0$ as functions of ω and*

$$\int_{-\pi}^{\pi} \left| \tilde{p}_n(\omega) \right|^2 d\omega = \int_{-\pi}^{\pi} \left| \tilde{p}_0(\omega) \right|^2 d\omega \qquad \text{for all } n \geq 0, \tag{T3.7}$$

with no redistribution of energy between frequency bands.

Proof. The shear fixes the momentum block, $M_\gamma(q, p) = (q + \gamma p,\, p)$, so $p_n = p_0$ pointwise in the index. Transforming term by term gives $\tilde{p}_n = \tilde{p}_0$ identically in ω, and squaring and integrating gives (T3.7). Since the integrand itself is unchanged, the identity holds band by band and not merely in aggregate. □

Remark T3.10 (This is not a Parseval identity for M_γ). It would be a misreading to take (T3.7) as saying that M_γ acts as an isometry. It does not: in the Euclidean norm

$$\|M_\gamma(q,p)\|^2 = \|q\|^2 + 2\gamma\, q^\top p + (1+\gamma^2)\|p\|^2,$$

which exceeds $\|q\|^2 + \|p\|^2$ by $2\gamma q^\top p + \gamma^2\|p\|^2$, generically non-zero. What Proposition T3.9 asserts is the invariance of one coordinate block, of which the spectral statement is a tautological consequence. Reading it as an isometry would smuggle back in exactly the boundedness claim that §T4.2 declines to make; Proposition T2.5 says the opposite, that iteration would grow the position block linearly.

T3.4 The design space, and why the upper shear

Definition T3.1 was introduced constructively. We now show it is forced, by enumerating the admissible local design space and eliminating every alternative on independent grounds. Each elimination below is a separate argument, and each is proved rather than asserted.

Three elementary families. Linear symplectic maps form the group $\mathrm{Sp}(2d_k, \mathbb{R})$. Near the identity each one-parameter subgroup $\{\exp(\gamma X)\}$ is generated by an element of the Lie algebra $\mathfrak{sp}(2d_k, \mathbb{R})$, characterised by $X^\top\Omega + \Omega X = 0$, which in block form reads

$$X = \begin{pmatrix} A & B \\ C & -A^\top \end{pmatrix}, \qquad B = B^\top, \quad C = C^\top, \tag{T3.8}$$

with $A \in \mathbb{R}^{d_k \times d_k}$ arbitrary. Three elementary sub-families capture the alternatives relevant to the design conditions below — they are not an exhaustive classification, and nothing downstream requires one — namely the qualitatively distinct possibilities. **(F1) Shears**, $A = 0$: the *upper* shear takes $B \neq 0$, $C = 0$, and since $X^2 = 0$ the exponential terminates, $\exp(\gamma X) = I + \gamma X$, giving M_γ at $B = I$; the *lower* shear takes $B = 0$, $C \neq 0$, giving $M^l_\gamma : (q,p) \mapsto (q,\, p + \gamma q)$. **(F2) Rotations**, $A = 0$, $C = -B$ with B positive definite: these generate the maximal compact subgroup $\mathrm{Sp}(2d_k, \mathbb{R}) \cap \mathrm{O}(2d_k) \cong \mathrm{U}(d_k)$, of which RoPE is the instance $U = \cos(t\theta)I$, $V = \sin(t\theta)I$. **(F3) Skew generators**, A skew-symmetric: purely imaginary spectrum, the family underlying antisymmetric recurrent networks [2].

Proposition T3.11 (The lower shear is null on the score)**.** *For every $\gamma \in \mathbb{R}$ and all streams, $(\hat{q}^l_t)^\top \hat{k}^l_j = q_t^\top k_j$.*

Proof. The lower shear leaves the position block fixed: $M^l_\gamma(q,p) = (q,\, p + \gamma q)$, so $\hat{q}^l_t = q_t$ and $\hat{k}^l_j = k_j$. The bilinear score reads only the position block, so it is unchanged, and in particular independent of γ. □

Corollary T3.12 (The lower shear is all-pass)**.** *The transfer function induced on the score-bearing stream by the lower shear is $H^l(z) \equiv 1$, of unit magnitude at every frequency; it violates* (C2) *identically, whereas Proposition T3.4 gives $|H^u(e^{j\pi})| = 1 + 2\gamma > 1$ strictly.*

Proof. Immediate from Proposition T3.11: the position stream is mapped to itself, so the ratio of transforms is the constant 1. □

Proposition T3.13 (Hamiltonian generators of the two shears)**.** *The upper shear is the γ-flow of $H^u(q,p) = \frac{1}{2}\|p\|^2$; the lower shear is the γ-flow of $H^l(q,p) = -\frac{1}{2}\|q\|^2$.*

Proof. For H^u, Hamilton's equations are $\dot{q} = \partial_p H = p$ and $\dot{p} = -\partial_q H = 0$, integrating to $(q_0 + \gamma p_0,\, p_0)$. For H^l they are $\dot{q} = 0$ and $\dot{p} = q$, integrating to $(q_0,\, p_0 + \gamma q_0)$. □

Remark T3.14 (Why the generator matters). In Proposition T3.13, $H^u \geq 0$ is the free particle: energy bounded below, straight-line trajectories. $H^l \leq 0$ is a negative quadratic-potential shear: energy bounded above and momentum growing without bound in γ. Even setting aside Proposition T3.11, the lower shear assigns the roles backwards — it shifts momentum by position, inserting content into a channel the score never reads, while leaving untouched the channel it does read. The upper shear shifts position by momentum, which is what matches the first-difference structure $p_{q,t} = q_t - q_{t-1}$.

Proposition T3.15 (Skew generators do not act as a per-stream shear)**.** *Let $X^\top = -X$, so that $\exp(\gamma X)$ is orthogonal. Then $\exp(\gamma X)$ does not have the shear form $(q, p) \mapsto (q + \gamma Bp,\ p)$ for any $\gamma \neq 0$: it moves the momentum block, and the induced map on the score-bearing stream is not a per-stream two-tap filter of the class of Theorem T3.18. The compact family is therefore outside the factorised subclass in which the uniqueness theorem is stated, rather than being eliminated within it.*

Proof. A shear fixes the momentum block, $\exp(\gamma X)|_p = \mathrm{id}$, so X would have to satisfy $Xe_p = 0$ on the momentum coordinates; for skew-symmetric X this forces the corresponding rows to vanish as well, hence $X = 0$. For $\gamma \neq 0$ and $X \neq 0$ the momentum block is moved, and the momenta entering the next position are no longer the first differences of the transformed positions, so the construction falls outside the factorised subclass in which Theorem T3.18 is stated. Conditions (C1) and (C2) are not what fails. □

Remark T3.16 (What this does and does not establish). It would be convenient to argue that orthogonality of $\exp(\gamma X)$ on the joint phase space makes the induced filter all-pass, hence in violation of (C2). That argument is invalid, and we record why, since it is the natural first attempt. Norm preservation on (q, p) jointly does not constrain the transfer function of the *projected* score-bearing coordinate. A planar phase-space rotation gives $\hat{q} = \cos\alpha\, q + \sin\alpha\, p$, and substituting $p = (1 - z^{-1})q$ yields

$$H_q(z) = \cos\alpha + \sin\alpha\,(1 - z^{-1}), \tag{T3.9}$$

whose magnitude is frequency-dependent for every $\alpha \notin \pi\mathbb{Z}$: at $\alpha = 0.8$, for instance, $|H_q(e^{j0})| = 0.697$ against $|H_q(e^{j\pi})| = 2.131$. The compact family is excluded from the classification by the structural argument above, not by an all-pass property it does not have.

Proposition T3.17 (Rotations cannot be repurposed: the positional encoding catastrophe)**.** *RoPE owes its usefulness to relative-position factorisation, $q_t^\top k_j = Q_t^\top R((j-t)\theta)K_j$. Suppose momentum were encoded inside the rotation phase, $\tilde{q}_t = R(t\theta + \gamma\psi_t)Q_t$ and $\tilde{k}_j = R(j\theta + \gamma\psi_j)K_j$. Then*

$$\tilde{q}_t^\top \tilde{k}_j = Q_t^\top\, R\big((j-t)\theta + \gamma(\psi_j - \psi_t)\big)\, K_j, \tag{T3.10}$$

which is not a function of $j - t$ alone unless ψ_t is affine in t — in which case the construction collapses to a frequency shift and carries no momentum information.

Proof. Planar rotations commute and satisfy $R(a)^\top R(b) = R(b - a)$; applying this to the two dual-use streams gives (T3.10) directly. The result depends on $\psi_j - \psi_t$. If this is to be a function of $j - t$ alone for all pairs, then ψ satisfies the Cauchy equation $\psi_j - \psi_t = f(j - t)$, whose solutions on the integers are $\psi_t = at + b$. Substituting back, the phase becomes $(j - t)(\theta + \gamma a)$: a rescaled rotation frequency, with the momentum dependence gone. □

Table 3 collects the four verdicts. Each row is settled by a separate proved statement rather than by a preference, and the elimination of the lower shear is settled twice over: it is both operationally null on the score and all-pass in the filter sense.

Table 3: The local design space. Each family is eliminated by a separate proved statement, not by a preference. "Operational" asks whether the map changes the score at all.

Family	(C1)	(C2)	Operational	Verdict
Upper shear M_γ	yes	yes (Prop. T3.4)	yes	**admissible**
Lower shear M_γ^l	yes	no, all-pass (Cor. T3.12)	no (Prop. T3.11)	dismissed
Skew (F3)	yes	—	moves the momentum block (Prop. T3.15)	outside the class
Rotation (RoPE reuse)	yes	yes	breaks factorisation (Prop. T3.17)	dismissed

Within the admissible family the coefficient matrix B in (T3.8) remains free. The canonical choice $B = I$ is what makes the generator the isotropic free-particle Hamiltonian rather than an anisotropic variant; any positive-definite B would be symplectic and high-pass, at the cost of introducing $d_k(d_k + 1)/2$ parameters and a preferred set of directions in embedding space that the construction otherwise does not need.

T3.5 Uniqueness within the factorised subclass

We now ask what freedom remains once the augmentation is required to be linear, one-step, causal, order-raising in the sense of §T1, and compatible with the phase-space reading. The answer is a single scalar. We state the conditions separately, because they are not of equal standing: three are forced by the filter-order analysis, and the fourth — symmetry between the two streams — is an independent design commitment which Prop. T3.25 then characterises algebraically rather than leaving as a bare axiom.

Theorem T3.18 (Uniqueness of PSA within the factorised subclass)**.** *Let $\mathcal{C}_{\mathrm{PSA}}$ be the class of score augmentations $s_{t,j} = \Phi(q_t, q_{t-1}, k_j, k_{j-1})$ acting on post-RoPE streams, with Φ bilinear separately in its q- and k-arguments, and consider the factorised subclass in which the augmentation acts as $\hat{q}_t = \alpha q_t + \beta q_{t-1}$, $\hat{k}_j = \alpha' k_j + \beta' k_{j-1}$ and $s_{t,j} = \hat{q}_t^\top \hat{k}_j$. Impose*

(C1) Causality and one-step memory.
$\hat{q}_t$ depends only on q_t, q_{t-1}, and $\hat{k}_j$ only on k_j, k_{j-1}.

(C2) Per-stream high-pass.
The per-stream transfer function $H(z) = \alpha + \beta z^{-1}$ satisfies $|H(e^{j0})| < |H(e^{j\pi})|$ with $|H(e^{j\omega})|$ monotone non-decreasing on $[0, \pi]$, and likewise for the key stream.

(C3.1) DC neutrality.
$H(e^{j0}) = 1$ on each stream: the augmentation leaves a constant stream unchanged. The stronger equality rather than $|H(e^{j0})| = 1$ is what the words require: magnitude alone would admit $H(e^{j0}) = -1$, which negates a constant stream. At score level a simultaneous sign flip of both factors is a gauge that leaves the score invariant, so this is a choice of representative within that gauge and costs no generality.

(C3.2) Swap symmetry.
The score is invariant under the relabelling $(q, k) \mapsto (k, q)$ of the two streams.

Then $\alpha = \alpha' = 1 + \gamma$ and $\beta = \beta' = -\gamma$ for a single $\gamma \in \mathbb{R}_{>0}$; that is, the augmentation of Definition T3.1 is the unique member of the factorised subclass satisfying (C1)–(C3.2)*, up to the scalar coupling γ.*

Proof. **Step 1 (C1).** Linearity, causality and one-step memory give $\hat{q}_t = \alpha q_t + \beta q_{t-1}$ with $\alpha, \beta \in \mathbb{R}$, and analogously $\hat{k}_j = \alpha' k_j + \beta' k_{j-1}$. This is the general form; no generality is lost.

Step 2 (C2). On the query stream, $|H(e^{j\omega})|^2 = \alpha^2 + \beta^2 + 2\alpha\beta\cos\omega$, which is monotone non-decreasing on $[0, \pi]$ if and only if $\alpha\beta \le 0$, and satisfies $|H(e^{j0})| < |H(e^{j\pi})|$, i.e. $(\alpha+\beta)^2 < (\alpha-\beta)^2$, if and only if $\alpha\beta < 0$. Hence α and β have strictly opposite signs.

Step 3 (C3.1). DC neutrality gives $\alpha + \beta = 1$ directly; had we imposed only $|\alpha + \beta| = 1$, the alternative $\alpha + \beta = -1$ would survive and is the image of the present solution under the sign gauge noted in (C3.1). Fixing the orientation so that the augmentation acts as the identity plus a correction (equivalently, taking $\alpha > 0$, which is a choice of sign for the readout and not a restriction), we obtain $\alpha + \beta = 1$. Together with $\alpha\beta < 0$ this forces $\beta < 0 < \alpha$, and writing $\beta = -\gamma$ gives $\alpha = 1 + \gamma$ with $\gamma > 0$. Thus $\hat{q}_t = q_t + \gamma(q_t - q_{t-1}) = q_t + \gamma p_{q,t}$, which is the upper shear M_γ acting on $(q_t, p_{q,t})$. The same argument on the key stream gives $\hat{k}_j = k_j + \gamma' p_{k,j}$ for some $\gamma' > 0$, a priori distinct.

Step 4 (C3.2). Expanding with $\gamma_q := \gamma$, $\gamma_k := \gamma'$,

$$s_{t,j} = q_t^\top k_j + \gamma_q\, p_{q,t}^\top k_j + \gamma_k\, q_t^\top p_{k,j} + \gamma_q\gamma_k\, p_{q,t}^\top p_{k,j}. \tag{T3.11}$$

Under the relabelling $(q, k) \mapsto (k, q)$ the first and fourth terms are fixed while the second and third exchange, so the score is swap-invariant if and only if $\gamma_q = \gamma_k$. Writing $\gamma_k = \gamma_q + \varepsilon$, the deviation from the symmetric score is exactly $\varepsilon\,(q_t^\top p_{k,j} + \gamma_q\, p_{q,t}^\top p_{k,j})$, which vanishes identically for all streams only when $\varepsilon = 0$. Hence $\gamma_q = \gamma_k =: \gamma$, and the augmentation is that of Definition T3.1. □

Remark T3.19 (The standing of each condition). (C1) is the hypothesis-class restriction inherited from the SHT setting, and it is (C1) — specifically a non-zero lag coefficient — that supplies the key-side order §T1 identifies as the representability requirement. (C2) does something different, and the two should not be conflated: a pure delay $H(z) = z^{-1}$ has order one and is all-pass, so

order alone selects no spectral orientation. (C2) is the additional design restriction that picks the backward-difference realisation among order-one filters, motivated by the construction and supported by the low-pass control of §T6.5 rather than forced by representability. (C3.1) is what makes the augmentation a *difference* rather than a rescaling: without it, $\alpha + \beta$ is free and the augmentation silently changes the gain on constant components of the stream, which interacts with normalisation layers in a way the filter-order analysis does not control. (C3.2) is a design commitment, and we mark it as such: it is *not* implied by symplecticity, as Prop. T3.27 makes precise. Its justification is Prop. T3.25 — the symmetric member is the one whose score is the base pairing transported one integrator step — and the extensibility argument of §T8.

Remark T3.20 (Scope of the uniqueness claim). Theorem T3.18 is a uniqueness statement within the factorised subclass of $\mathcal{C}_{\mathrm{PSA}}$: bilinear, one-step, linear, acting stream-wise on the post-RoPE substrate. Non-linear shears, multi-step memories, augmentations that couple the two streams before the pairing, and score augmentations outside the bilinear form are not covered, and we make no claim about them. In particular the theorem is silent about the value stream, about augmentations applied pre-RoPE (Cor. T4.4 addresses these separately), and about non-symplectic couplings outside the class — where, as Gahtan and Bronstein [49] report, other choices can and do help. Nothing in Theorem T3.18 contradicts that finding; the theorem constrains a class, and their construction lies outside it.

T3.6 Beyond the linear class

Theorem T3.18 is a statement within the linear one-step class. The obvious question is whether non-linearity would buy anything. Two results bound the answer: an algebraic constraint that any symplectic non-linear shear must satisfy pointwise, and an observation about whether a learned module could satisfy it.

Proposition T3.21 (Symplectic non-linear shears have symmetric Jacobian)**.** *Let $g : \mathbb{R}^{d_k} \to \mathbb{R}^{d_k}$ be C^1 and consider the shear-type map $\Phi^g_\gamma : (q, p) \mapsto (q + \gamma\, g(p),\, p)$. Then Φ^g_γ is, for $\gamma \neq 0$, symplectic at every point if and only if*

$$Dg(p)^\top = Dg(p) \qquad \text{for every } p \in \mathbb{R}^{d_k}. \tag{T3.12}$$

Proof. **Step 1 (Jacobian).** Writing $(Q, P) = \Phi^g_\gamma(q, p)$ with $Q = q + \gamma g(p)$ and $P = p$, the block derivatives are $\partial Q/\partial q = I$, $\partial Q/\partial p = \gamma\, Dg(p)$, $\partial P/\partial q = 0$, $\partial P/\partial p = I$, so

$$\mathcal{J}(q, p) = \begin{pmatrix} I & \gamma\, Dg(p) \\ 0 & I \end{pmatrix}. \tag{T3.13}$$

Step 2 (the condition). Symplecticity at (q, p) means $\mathcal{J}^\top \Omega\, \mathcal{J} = \Omega$ with $\Omega = \left(\begin{smallmatrix} 0 & I \\ -I & 0 \end{smallmatrix}\right)$. Computing $\Omega\mathcal{J}$ first,

$$\Omega\mathcal{J} = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \begin{pmatrix} I & \gamma Dg(p) \\ 0 & I \end{pmatrix} = \begin{pmatrix} 0 & I \\ -I & -\gamma Dg(p) \end{pmatrix},$$

and then

$$\mathcal{J}^\top \Omega \mathcal{J} = \begin{pmatrix} I & 0 \\ \gamma Dg(p)^\top & I \end{pmatrix} \begin{pmatrix} 0 & I \\ -I & -\gamma Dg(p) \end{pmatrix} = \begin{pmatrix} 0 & I \\ -I & \gamma\big(Dg(p)^\top - Dg(p)\big) \end{pmatrix}.$$

This equals Ω if and only if the $(2, 2)$ block vanishes, which for $\gamma \neq 0$ is exactly (T3.12). **Step 3 (the linear case).** If $g(p) = Ap$ then $Dg \equiv A$ and the condition is $A = A^\top$; the canonical choice $A = I$ satisfies it with no constraint on anything learned. □

Remark T3.22 (The constraint is pointwise and of high codimension). Condition (T3.12) must hold at *every* p, not on average and not at a stationary point. Since $Dg(p)$ has d_k^2 free entries while symmetric matrices form a subspace of dimension $d_k(d_k + 1)/2$, the requirement cuts out a subspace of codimension $d_k(d_k - 1)/2$ at each point: at $d_k = 64$, codimension 2016, imposed everywhere on $\mathbb{R}^{64}$.

Remark T3.23 (Whether a learned module could satisfy it). Proposition T3.21 is permissive in principle: symmetric-Jacobian non-linearities exist. The operational question is what would actually be implemented. The natural candidate is a position-wise MLP, $g(p) = W_2\, \sigma(W_1 p + b_1) + b_2$, whose Jacobian is

$$Dg(p) = W_2\, \mathrm{diag}\big(\sigma'(W_1 p + b_1)\big)\, W_1. \tag{T3.14}$$

For generic W_1, W_2 this product is not symmetric at any p, and gradient descent on a downstream loss imposes no symmetry-preserving constraint on the pair. The condition can of course be imposed — by penalising $\|Dg - Dg^\top\|_F^2$, or by parameterising g as a gradient field — but it is not preserved automatically by a generic unconstrained parameterisation and update. The consequence of violating it is stated at the level this paper claims: the map ceases to be symplectic at the points where symmetry fails, hence ceases to be the exact flow of any Hamiltonian, and Lemma T2.1 and Theorem T4.1 no longer apply to it. We make no claim about what such a violation would do to training. The linear shear corresponds to $g = \mathrm{id}$, $Dg \equiv I$, which is symmetric automatically and without a learned parameter: linearity is not a restriction here so much as what makes the constraint free.

Remark T3.24 (A drift coincidence). Even granting an admissible non-linear g, the position increment is $\gamma\, g(p)$ rather than γp, so the high-pass gain acquires a dependence on the embedding magnitude: the filter would respond differently to the same frequency content at different scales. The only g for which the increment is scale-uniform across the trajectory is $g(p) = cp$, whose scalar is absorbed into γ: the identity up to coupling is the unique isotropic, direction-neutral linear choice.

T3.7 Why the value stream is untouched

The construction leaves V alone, and this is structural rather than a tuning decision. In the phase-space reading, q carries position and p carries its rate of change; the shear transforms the position coordinate by the momentum. The value stream is a measurement channel, not a dynamical coordinate: it is what the attention weights read out, not what they are computed from. Applying M_γ to V would shift values by value-differences, which has no phase-space interpretation and, more concretely, changes what the layer outputs rather than what it attends to. The filter-order analysis of §T1 is entirely about the score, and the score does not read V; there is therefore no order-theoretic argument for augmenting it, and Theorem T3.18 says nothing about such an augmentation either way. §T6 holds V fixed in every variant of the control battery for the same reason, so that the battery isolates the score-side mechanism.

T3.8 Two structural propositions

The two propositions of this subsection separate what symplecticity does from what symmetry does. They are stated here because they are the content of condition (C3.2) and because they determine which of the neighbouring augmentations the empirical battery of §T6 should be expected to discriminate.

Proposition T3.25 (Transport identity)**.** *Let Δ denote the first backward difference in the position index, $\Delta x_t := x_t - x_{t-1}$. Because the score carries two independent indices, the difference acting on it must be named: we use the* diagonal *backward difference*

$$\Delta_{\mathrm{diag}} F(t,j) := F(t,j) - F(t-1, j-1), \tag{T3.15}$$

which we use because the discrete Leibniz rule below holds for it; we do not classify every one-step difference operator. With equal couplings $\gamma_q = \gamma_k = \gamma$, the PSA score satisfies exactly

$$s^{\mathrm{PSA}}_{t,j} \;=\; q_t^\top k_j \;+\; \gamma\, \Delta_{\mathrm{diag}}\big(q^\top k\big)(t,j) \;+\; \gamma(1+\gamma)\, p_{q,t}^\top p_{k,j}, \tag{T3.16}$$

$$\Delta_{\mathrm{diag}}\big(q^\top k\big)(t,j) = p_{q,t}^\top k_j + q_t^\top p_{k,j} - p_{q,t}^\top p_{k,j}, \tag{T3.17}$$

where the second expression is the discrete Leibniz rule for Δ_{diag} applied to the base pairing. With unequal couplings $\gamma_k = \gamma_q + \varepsilon$ the score acquires, in addition, the antisymmetric current $\frac{\varepsilon}{2}\big(q_t^\top p_{k,j} - p_{q,t}^\top k_j\big)$ together with a symmetric remainder, and the antisymmetric part is not a total difference in the position index: there is no bilinear G with $\Delta G = q^\top \Delta k - \Delta q^\top k$.

Proof. Expanding $\hat q_t^\top \hat k_j$ with $\gamma_q = \gamma_k$ gives (T3.1); substituting the discrete Leibniz rule $\Delta(q^\top k) = p_q^\top k + q^\top p_k - p_q^\top p_k$ and collecting the γ^2 terms yields (T3.16), the coefficient $\gamma(1+\gamma)$ arising as $\gamma^2 + \gamma$ from the double-ghost term and the Leibniz correction. For the second claim, symmetrise and antisymmetrise (T3.11) in the stream label; the antisymmetric part is proportional to $q_t^\top k_{t-1} - q_{t-1}^\top k_t$. Suppose this were ΔG for a G bilinear and local in $\{q_t, q_{t-1}\} \times \{k_t, k_{t-1}\}$, say $G_t = a\, q_t^\top k_t + b\, q_t^\top k_{t-1} + c\, q_{t-1}^\top k_t + d\, q_{t-1}^\top k_{t-1}$. Then $\Delta G_t = G_{t+1} - G_t$ contains the monomials $q_{t+1}^\top k_{t+1}$, $q_{t+1}^\top k_t$ and $q_t^\top k_{t+1}$ with coefficients a, b and c, none of which appears in the target, forcing $a = b = c = 0$;

the surviving $\Delta G_t = -d\, q_{t-1}^\top k_{t-1}$ then cannot reproduce $q_t^\top k_{t-1} - q_{t-1}^\top k_t$ for any d. No such G exists in that class; we do not claim it for a wider one. Certificate C9 verifies both statements symbolically. □

Remark T3.26 (Reading the identity). Equation (T3.16) says that the symmetric PSA score is the base bilinear pairing, plus γ times its own transported increment, plus the pure coherence channel. This is the score-level image of self-adjointness of the symmetric composition in the integrator setting: the symmetric member is the unique one whose score is expressible using only the base pairing and its own difference, without a current term. It is this identity, rather than symplecticity, that distinguishes the symmetric member — a distinction that the ablations of §T6 are designed to detect and that Prop. T3.27 makes unavoidable.

Proposition T3.27 (Asymmetric shears are symplectic; asymmetry is a score bias)**.** *Let $\gamma_q, \gamma_k > 0$ be arbitrary and let $\mathcal{M}_{\gamma_q,\gamma_k} := \mathrm{diag}(M_{\gamma_q}, M_{\gamma_k})$ act on (q, p_q, k, p_k). Then:*

(i) $\mathcal{M}_{\gamma_q,\gamma_k}$ is exactly symplectic, $\mathcal{M}^\top \Omega\, \mathcal{M} = \Omega$, for every pair (γ_q, γ_k);

(ii) $H_{\mathrm{joint}} = \frac{1}{2}(\|p_q\|^2 + \|p_k\|^2)$ is exactly conserved for every pair (γ_q, γ_k);

(iii) writing $\gamma_k = \gamma_q + \varepsilon$, the score deviates from the symmetric score by exactly

$$s^{\mathrm{asym}}_{t,j} - s^{\mathrm{sym}}_{t,j} = \varepsilon\big(q_t^\top p_{k,j} + \gamma_q\, p_{q,t}^\top p_{k,j}\big), \tag{T3.18}$$

a score bias linear in ε with no energy component.

Consequently symplecticity does not single out the symmetric member, and unequal coupling costs a score bias rather than a loss of structure.

Proof. (i) and (ii): each block M_γ is unit upper triangular, and

$$M_\gamma^\top J M_\gamma = \begin{pmatrix} I & 0 \\ \gamma I & I \end{pmatrix} \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \begin{pmatrix} I & \gamma I \\ 0 & I \end{pmatrix} = \begin{pmatrix} 0 & I \\ -I & -\gamma I + \gamma I \end{pmatrix} = J,$$

the γ terms cancelling identically, so M_γ is symplectic for every γ independently; a direct sum of symplectic maps on complementary summands is symplectic on the sum, with no compatibility condition between the two couplings. Both shears fix their own momentum block, so both $\|p_q\|$ and $\|p_k\|$ are individually invariant, hence so is H_{joint}. (iii) is the difference of (T3.11) at $(\gamma_q, \gamma_q + \varepsilon)$ and at (γ_q, γ_q), computed term by term. Certificates C4 and C5 verify (i)–(ii) and (iii) respectively, the C4 residual evaluating to exactly 0.0. □

Remark T3.28 (What this implies for the empirical programme). Proposition T3.27 predicts that an asymmetrically sheared model should train successfully, the bias (T3.18) being absorbable by the learned readout, and it forbids any account of the mechanism that rests on symplecticity as the operative discriminator *within the two-tap class*. The operative discriminators the filter-order analysis leaves are (a) DC-neutral high-pass structure and (b) stream placement. §T6 tests exactly this: the asymmetric variant trains to 0.875, the low-pass member of matched parameter count collapses to 0.428, the query-only half-lift stays at 0.033 against a ceiling of 0.077, and the key-only half-lift reaches 0.949. Symplecticity is a structural commitment, not a performance claim, and we do not present it as one.

T4 Operator-level symplectic invariance and the placement corollary

Lemma T2.1 states that M_γ is the exact γ-flow of $H(q,p) = \frac{1}{2}\|p\|^2$. This section records what that fact does, and does not, license once the operator is transported onto the attention substrate. We state the invariance first, then delimit it, then derive the placement constraint it implies. The delimitation in §T4.2 is not a caveat appended to a stronger claim; it is part of the claim, and the theorem below is stated so that nothing outside it is asserted anywhere in the paper.

T4.1 The invariance

Theorem T4.1 (Operator-level symplectic invariance of M_γ)**.** *Let Ω denote the canonical symplectic form on the joint pre-softmax phase space $(q, p_q, k, p_k) \in \mathbb{R}^{4d_k}$ and let $\mathcal{M}_\gamma := \mathrm{diag}(M_\gamma, M_\gamma)$ act on it, M_γ as in (T2.7). Then:*

(i) (Exactness.) *Each application of $\mathcal{M}_\gamma$ preserves Ω exactly, $\mathcal{M}_\gamma^\top \Omega \mathcal{M}_\gamma = \Omega$, with no truncation term at any order in γ.*

(ii) (Invariant.) *The quadratic form $H_{\text{joint}}(q, p_q, k, p_k) = \frac{1}{2}\big(\|p_q\|^2 + \|p_k\|^2\big)$ is exactly conserved by $\mathcal{M}_\gamma$, both momenta being fixed points of the shear.*

(iii) (Perturbation estimate, per application.) *If $\mathcal{M}'$ satisfies $\|\mathcal{M}' - \mathcal{M}_\gamma\|_2 \le \epsilon$, then $\|\mathcal{M}'^\top \Omega \mathcal{M}' - \Omega\|_2 \le 2\epsilon\|\mathcal{M}_\gamma\|_2 + \epsilon^2$ and, for any phase point x, $|H_{\text{joint}}(\mathcal{M}'x) - H_{\text{joint}}(x)| \le \big(\epsilon\|\mathcal{M}_\gamma\|_2 + \frac{1}{2}\epsilon^2\big)\|x\|^2$.*

All three statements are per application*, and none of them is a stability guarantee. Part (iii) in particular is a continuity estimate — it bounds how far a map* near *$\mathcal{M}_\gamma$ departs from preserving Ω and H_{joint} in a single application — and it says nothing about what happens when any map is applied repeatedly, which is what a stability statement would have to address. The construction defines no iterated dynamical system on its own output, and no claim about stability of any kind, whether of training dynamics, optimisation, depth-wise propagation or long-horizon behaviour of the attention map, is made or used anywhere in this paper. The trained behaviour under coupling and depth is reported as measurement in §T6.12, and is labelled Empirical there. See §T4.2 for the boundary in full.*

Proof. **(i)** Write $\mathcal{M}_\gamma$ in the block ordering (q, p_q, k, p_k). It is block diagonal with two identical $2d_k \times 2d_k$ blocks $M_\gamma = \left(\begin{smallmatrix} I & \gamma I \\ 0 & I \end{smallmatrix}\right)$, and Ω is correspondingly block diagonal with two copies of J. Direct computation of the four sub-blocks of $M_\gamma^\top J M_\gamma$ gives

$$\begin{pmatrix} I & 0 \\ \gamma I & I \end{pmatrix}\begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}\begin{pmatrix} I & \gamma I \\ 0 & I \end{pmatrix} = \begin{pmatrix} 0 & I \\ -I & -\gamma I + \gamma I \end{pmatrix} = J,$$

the γ terms cancelling identically rather than to leading order. A block-diagonal map preserving each summand's form preserves their direct sum, so $\mathcal{M}_\gamma^\top \Omega \mathcal{M}_\gamma = \Omega$. Because M_γ is unit upper triangular, the identity is exact in exact arithmetic and, evaluated numerically, the residual is identically zero in floating point rather than at machine precision (certificate C6). **(ii)** The shear fixes the momentum block: $\mathcal{M}_\gamma(q, p_q, k, p_k) = (q + \gamma p_q,\ p_q,\ k + \gamma p_k,\ p_k)$, so both $\|p_q\|$ and $\|p_k\|$ are unchanged and H_{joint} is invariant. This holds for every γ, with no smallness condition. **(iii)** Write $\mathcal{M}' = \mathcal{M}_\gamma + E$ with $\|E\|_2 \le \epsilon$. Then $\mathcal{M}'^\top \Omega \mathcal{M}' - \Omega = \mathcal{M}_\gamma^\top \Omega E + E^\top \Omega \mathcal{M}_\gamma + E^\top \Omega E$, and $\|\Omega\|_2 = 1$ gives the stated bound; the energy bound follows from $H_{\text{joint}}(y) - H_{\text{joint}}(x) = \frac{1}{2}(\|\Pi y\|^2 - \|\Pi x\|^2)$ with Π the momentum projection, $\|\Pi\|_2 = 1$, and the same expansion. □

Remark T4.2 (Why the residual is exactly zero, not near zero). The relevant diagnostic is the residual $\|\mathcal{M}_\gamma^\top \Omega \mathcal{M}_\gamma - \Omega\|$ and not the determinant. Unit determinant is necessary but not sufficient for symplecticity in dimension $2d_k > 2$: the group $\mathrm{Sp}(2d_k, \mathbb{R})$ is a proper subgroup of $\mathrm{SL}(2d_k, \mathbb{R})$ for $d_k > 1$, so a determinant check cannot distinguish a symplectic map from a volume-preserving non-symplectic one. The residual is the faithful test, and for a unit-triangular block it evaluates to exactly 0.0 because the cancelling terms $-\gamma I + \gamma I$ are represented exactly in binary floating point. §T6 reports it as measured on trained models.

T4.2 What transfers from geometric integration, and what does not

The correspondence of §T2 is a correspondence of *algebra*, not of *dynamics*. Making the distinction explicit matters here more than in most transplanted analogies, because the geometric-integration literature's headline results are long-time statements about iterated maps, and the attention-side construction does not iterate.

The transformed stream is not a phase state. PSA forms $\hat{q}_t = q_t + \gamma p_{q,t}$ from the momenta $p_{q,t} = q_t - q_{t-1}$ computed on the *untransformed* stream. If one were to re-extract a momentum from the transformed stream by the same backward difference, one would obtain

$$\hat{q}_t - \hat{q}_{t-1} \;=\; p_{q,t} + \gamma\,(p_{q,t} - p_{q,t-1}) \;=\; (1+\gamma)\,p_{q,t} - \gamma\,p_{q,t-1} \;\neq\; p_{q,t}, \tag{T4.1}$$

so the pair $(\hat{q}_t, p_{q,t})$ is not the image of $(q_t, p_{q,t})$ under a map that closes on its own output. The transformation is applied once per forward pass, to a stream supplied by the embedding and RoPE stages, and its output is consumed by the softmax. There is no discrete trajectory $x_0 \mapsto x_1 \mapsto x_2 \mapsto \cdots$ generated by $\mathcal{M}_\gamma$ anywhere in the architecture.

Consequences for what may be claimed. Backward error analysis constructs, for a one-step integrator $\Phi_{\Delta t}$ applied repeatedly to its own output, a modified Hamiltonian $\widetilde{H}$ whose exact flow interpolates the numerical trajectory to all orders in Δt; the near-conservation of $\widetilde{H}$ over times exponentially long in $1/\Delta t$ is what yields bounded energy error for symplectic methods [7, 8, 13, 3]. Every hypothesis of that construction concerns the iterate. Since the attention-side map does not iterate, we make no modified-Hamiltonian statement, no backward-error expansion, no long-time energy bound, and no secular-drift statement of any kind on the attention side. Table 4 states the boundary line by line.

Table 4: The boundary of the correspondence. The left column lists the classical statements of §T2; the right column records whether an analogue is asserted for the attention operator. The distinguishing hypothesis in every declined row is iteration of the map on its own output, which (T4.1) shows does not occur.

Classical statement (integrator)	Requires	Attention-side analogue
M_γ is the exact flow of a quadratic Hamiltonian; symplectic at each application	one application	**Asserted.** Theorem T4.1(i)–(ii); certificate C6
Order of the stencil is raised from 1 to 2 by passing to phase space	algebra only	**Asserted** in the filter-order image: joint order $(0,0) \to (1,1)$ (§T1, Thm. T3.18)
Self-adjointness of the symmetric composition	algebra only	**Asserted** as the transport identity, Prop. T3.25
Existence of a modified Hamiltonian $\widetilde{H}$ conserved to all orders	iterated map	**Not asserted.** No iterate exists; see (T4.1)
Energy error bounded over times $e^{c/\Delta t}$; secular drift absent	iterated map, small step, analytic field	**Not asserted.** No long-time or depth-wise stability claim is made
Secular energy growth $E_n/E_0 = (1+\Delta t^2)^n$ for the non-symplectic alternative	iterated map	**Not asserted.** The perturbation statement we make is per application, Thm. T4.1(iii)
Symplecticity discriminates task performance	—	**Not asserted.** Measured to be false within the two-tap class: §T6 and Prop. T3.27

Remark T4.3 (Learning γ does not create an iterate). The claim that $\mathcal{M}_\gamma$ is applied once is a statement about the data flow, not about γ being held constant, and it is worth separating the two because the coupling may reasonably be optimised rather than swept.

Along the sequence. The momenta are formed from the untransformed stream, and the output $(\hat{q}, \hat{k})$ is consumed by the softmax; it never returns as the position variable from which a momentum is taken. Equation (T4.1) is the algebraic form of this, and nothing in it involves the value of γ.

Through depth. At layer $\ell + 1$ the shear acts on queries and keys freshly projected from the residual stream, not on the sheared tensors of layer ℓ, which never enter the residual stream at all. The composition $\mathcal{M}_\gamma \circ \mathcal{M}_\gamma$ occurs nowhere in the architecture, with per-layer couplings or without.

Through training. If γ is optimised then $\gamma_{n+1} = \gamma_n - \eta\, \partial_\gamma \mathcal{L}$ is indeed an iterated map. It iterates on the *parameter*, however, not on the phase space (q, p) that carries Ω; these are different spaces, and we make no symplectic claim about the optimiser trajectory, nor need one. What matters is that parts (i) and (ii) of Theorem T4.1 hold for *every* $\gamma \in \mathbb{R}$, with no smallness condition and no requirement that γ be fixed. They therefore hold at every step of any such trajectory and at every value it visits. Concretely, $\{\mathcal{M}_\gamma\}_{\gamma \in \mathbb{R}}$ is a one-parameter subgroup of $\mathrm{Sp}(2d_k, \mathbb{R})$, closed under composition,

$$M_a M_b = \begin{pmatrix} I & aI \\ 0 & I \end{pmatrix} \begin{pmatrix} I & bI \\ 0 & I \end{pmatrix} = \begin{pmatrix} I & (a+b)I \\ 0 & I \end{pmatrix} = M_{a+b}, \tag{T4.2}$$

so a trajectory through the family never leaves the group.

What would break the claim. An architecture that fed $\hat{q}$ back as the position variable of a second shear on the same stream would compose the operator with itself, and the backward-error apparatus of §T2.3 would become relevant to it. We do not do this. By (T4.2) even that composite would remain exactly symplectic — what it would forfeit is the "applied once" statement, and with it the licence to decline the long-time questions.

What the analogy is for. Stripped of the dynamical content, the correspondence still carries the paper's argument, because that argument is about *representability*, not about stability. Both sides face the same obstruction — a one-step map on the base variable cannot reach the required order — and both resolve it by the same move, an extension of the state by a first difference and a linear

shear on the extended state. The integrator literature supplies the operator, the reason it is canonical among one-step linear maps (it is the exact flow of the free-particle Hamiltonian, Lemma T2.1), and the reason the symmetric member is distinguished within its class (self-adjointness, Prop. T3.25). It supplies nothing about long-time behaviour to the attention side, and we take nothing.

Empirical status of the dynamical question. That the dynamical question is not answered by theorem does not make it uninteresting, and §T6 reports a pre-registered characterisation of the trained behaviour — divergence, depth-wise score growth, and coupling-perturbation sensitivity — as measurement. It is labelled Empirical throughout and is not a substitute for the theorem we do not prove.

T4.3 The placement corollary

Corollary T4.4 (Placement post-RoPE). *M_γ does not commute with RoPE: applying M_γ* before *the RoPE rotation R produces a Coriolis error*

$$\hat{q}_t^{\mathrm{pre}} - \hat{q}_t^{\mathrm{post}} = -\gamma(R_t - R_{t-1})q_{t-1}, \qquad \|\hat{q}_t^{\mathrm{pre}} - \hat{q}_t^{\mathrm{post}}\| = 2\gamma|\sin(\theta/2)|\,\|q_{t-1}\|, \tag{T4.3}$$

where θ is the per-band RoPE rotation angle. Pre-RoPE placement therefore destroys the per-stream high-pass character of the augmentation in the rotating frame. It does not *break symplecticity: RoPE rotations are themselves symplectic, so the pre-RoPE composite $R_t M_\gamma R_{t-1}^{-1}$ is a product of symplectic maps and satisfies $\mathcal{M}^\top \Omega\, \mathcal{M} = \Omega$ exactly at every θ. The cost of the wrong placement is the Coriolis term (T4.3), not a loss of symplecticity or of the momentum invariant, though the swap-symmetric transport identity is forfeited.*

Proof. RoPE acts on q_t as $R_t q_t$ and on q_{t-1} as $R_{t-1}q_{t-1}$. Pre-RoPE shear gives $R_t\big(q_t + \gamma(q_t - q_{t-1})\big) = R_t q_t + \gamma R_t q_t - \gamma R_t q_{t-1}$, whereas the post-RoPE shear gives $R_t q_t + \gamma R_t q_t - \gamma R_{t-1} q_{t-1}$. The difference is $-\gamma(R_t - R_{t-1})q_{t-1}$, and since $R_t R_{t-1}^{-1}$ is a rotation by the per-band angle θ, $\|(R_t - R_{t-1})q_{t-1}\| = 2|\sin(\theta/2)|\,\|q_{t-1}\|$ by direct trigonometric computation. For the symplecticity claim, the pre-RoPE composite acting on the post-RoPE phase variables is $R_t M_\gamma R_{t-1}^{-1}$ rather than M_γ. Each R is a block rotation, hence orthogonal and symplectic, $R^\top \Omega R = \Omega$; a product of symplectic maps is symplectic, so the composite satisfies the identity exactly and its residual is zero at every θ. Appendix A8 gives the band-by-band accounting; Certificate C6 evaluates the post-RoPE residual $\|M_\gamma^\top J M_\gamma - J\|_F$ on the operator itself, where it is exactly zero and independent of the surrounding model, and §T6 reports it as exactly zero under the residual diagnostic at $d_k = 64$. □

Remark T4.5 (Interpretation of the Coriolis term). The error is a genuine frame effect rather than a bookkeeping artefact. Differencing in the unrotated frame and then rotating computes the difference of two vectors expressed in *different* frames; the discrepancy is the rotation increment applied to the lagged vector, which is exactly the discrete Coriolis term of a rotating reference frame. Its magnitude vanishes as $\theta \to 0$ — the low-frequency RoPE bands, where the two placements nearly agree — and is largest in the high-frequency bands. Since §T6 finds induction efficacy concentrated at low θ, the penalty is largest precisely in the bands that carry the least induction signal, which is why the pre-RoPE variant degrades gracefully rather than catastrophically in the A/B sweep of Appendix A30 while still losing consistently.

T5 Closed-form single-layer phase transition

Representability is not attainment. Theorem T5.7 says the augmented head *can* express the induction discriminator; this section derives *where* on the coupling axis it does so, in closed form and with no fitted parameters. The derivation is assumption-light at the score level: it uses the exact second moments of the four-term decomposition (T3.1) and an annealed evaluation of the softmax competition, and nothing else. No small-signal condition, no softmax linearisation, and no subspace-separation hypothesis enters at any point; the criterion states and saturates its own domain of validity exactly at the branch crossover, which is the sense in which it is self-delimiting.

One point of nomenclature. The transition derived here is on the *coupling* axis at a fixed training budget: γ_c is a property of the operator, not of the optimiser. It should not be identified with the abrupt emergence of induction heads *during training* reported by Olsson et al. [20] and modelled

in closed form by Reddy [24], which is a transition in training time and in data distribution at fixed architecture. The two are measured on different axes, and neither is an estimator of the other. Whether the shape derived here nonetheless accounts for the shape they observe is a separate question, and it is argued separately — at the level of qualitative shape and mechanism, not of fitted prefactors — in Appendix A35.

The lineage of the argument. Figure 5 sets the three strands side by side. Panel (a) is Derrida's random energy model [14, 15]: the annealed entropy $s(b) = L - b^2/2$ counts the states carrying the partition sum, and it vanishes at $b_c = \sqrt{2L}$, beyond which one extremal state carries the sum and the annealed evaluation over-counts. Panel (b) is the consequence in the variables of an attention score, where $b = 2\gamma^2$: the two branches of (T5.14) are the annealed and frozen phases of that same model, and the crossover sits at $d_k^* = 8L$ because that is exactly where b_c is reached — not at an imposed cut. The step at the crossover is worth reading carefully: it is the factor c and nothing else, since the mean-field branch evaluates to $(2L)^{1/4}/\sqrt{2}$ there and the condensed branch to c times that. Without the calibrated constant the two would meet.

Panel (c) is the distinction of the preceding paragraph, drawn. The horizontal line is γ_c, a property of the operator at a fixed training budget; the vertical line is the emergence of induction heads during training [20, 24], a property of the optimiser at fixed architecture. Only the upper-right quadrant has the circuit, and neither line predicts the other. The strip beneath records the route by which the freezing scale reached attention: Derrida in 1980–81, the training-time transition in Olsson et al. [20] and Reddy [24], the identification of attention scores with the random energy model in Giorlandino and Goldt [47] — and the disagreement over its scaling raised by Chen et al. [48] — and, here, the same scale reached along a different axis.

T5.1 The softmax: Fréchet linearisation and the DC/AC split

The pre-softmax score has been characterised exactly. The softmax is the non-linearity that converts scores to weights, and it destroys the Hamiltonian structure globally. This subsection derives the linearisation used throughout the Layer-2 analysis, bounds its remainder, and introduces the decomposition on which the induction driver is identified. The results here are conditional — they hold in a regime that §T6 measures rather than assumes — and they are marked as such. The two-branch theorem of §T5.7 does *not* depend on any of them.

Definition T5.1 (DC/AC split)**.** Decompose each post-RoPE stream into a slowly varying component with frequency content concentrated near $\omega = 0$ and a rapidly varying complement, $q_t = q_t^{\mathrm{DC}} + q_t^{\mathrm{AC}}$ and $k_j = k_j^{\mathrm{DC}} + k_j^{\mathrm{AC}}$. At score level write $S_{t,j} = S_{t,j}^{\mathrm{DC}} + S_{t,j}^{\mathrm{AC}}$ with $S_{t,j}^{\mathrm{DC}} := (q_t^{\mathrm{DC}})^\top k_j^{\mathrm{DC}}$, and at weight level $a_t^{\mathrm{DC}} := \mathrm{softmax}(S_t^{\mathrm{DC}})$.

Proposition T5.2 (Fréchet linearisation of the softmax)**.** *Let $S = S^{\mathrm{DC}} + S^{\mathrm{AC}} \in \mathbb{R}^T$ and set $a := \mathrm{softmax}(S^{\mathrm{DC}})$ and $J_a := \mathrm{diag}(a) - aa^\top$. Then*

$$\mathrm{softmax}(S) = a + J_a S^{\mathrm{AC}} + R(S^{\mathrm{AC}}), \qquad \|R(S^{\mathrm{AC}})\|_1 \le C\,\|S^{\mathrm{AC}}\|_\infty^2, \tag{T5.1}$$

with C depending only on $\|a\|_\infty$ and not on T; the choice of ℓ_1 on the output and ℓ_∞ on the perturbation is what keeps C dimension-free, since softmax *maps into the simplex.*

Proof. **Step 1.** Componentwise, $e^{S_i} = e^{S_i^{\mathrm{DC}}}\big(1 + S_i^{\mathrm{AC}} + \mathcal{O}((S_i^{\mathrm{AC}})^2)\big)$ by the scalar exponential series. **Step 2 (partition function).** With $Z^{\mathrm{DC}} := \sum_j e^{S_j^{\mathrm{DC}}}$,

$$Z = \sum_j e^{S_j} = Z^{\mathrm{DC}}\Big(1 + \sum_j a_j S_j^{\mathrm{AC}} + \mathcal{O}(\|S^{\mathrm{AC}}\|^2)\Big) = Z^{\mathrm{DC}}\big(1 + a^\top S^{\mathrm{AC}} + \mathcal{O}(\|S^{\mathrm{AC}}\|^2)\big),$$

so, by $(1+x)^{-1} = 1 - x + \mathcal{O}(x^2)$ applied to $x = a^\top S^{\mathrm{AC}}$, $Z^{-1} = (Z^{\mathrm{DC}})^{-1}\big(1 - a^\top S^{\mathrm{AC}} + \mathcal{O}(\|S^{\mathrm{AC}}\|^2)\big)$. **Step 3 (combine).** For each i,

$$\begin{aligned}[\mathrm{softmax}(S)]_i &= a_i(1 + S_i^{\mathrm{AC}})(1 - a^\top S^{\mathrm{AC}}) + \mathcal{O}(\|S^{\mathrm{AC}}\|^2)\\ &= a_i + a_i S_i^{\mathrm{AC}} - a_i\, a^\top S^{\mathrm{AC}} + \mathcal{O}(\|S^{\mathrm{AC}}\|^2),\end{aligned}$$

and the linear term is exactly $[\mathrm{diag}(a)S^{\mathrm{AC}}]_i - [aa^\top S^{\mathrm{AC}}]_i = [J_a S^{\mathrm{AC}}]_i$. The remainder collects quadratic and higher terms whose coefficients are bounded by $\|a\|_\infty$, giving the stated bound. □

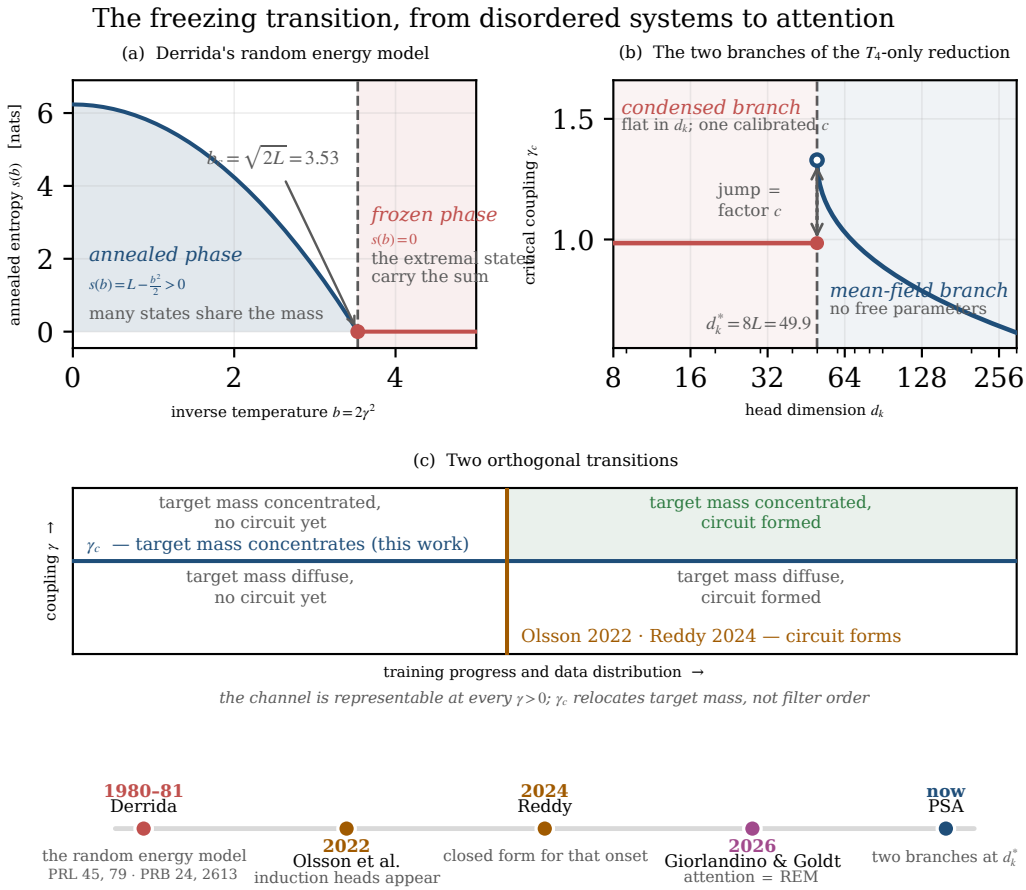


Figure 5: The freezing transition, from disordered systems to attention. (a) The annealed entropy of the random energy model, $s(b) = L - b^2/2$, at $L = \log 511$. It vanishes at $b_c = \sqrt{2L} = 3.53$: to the left the partition sum is shared among $e^{s(b)}$ states; to the right the sum is extreme-value dominated, carried by the leading extremal states, and the annealed evaluation is no longer valid. (b) The same two phases as the two branches of (T5.14), in the same two colours. The panel concerns the T_4-only reduction $\mathcal{G}$ and nothing wider: §T9 finds the branch structure to be the most reduction-sensitive feature of the theory, not surviving the restoration of T_3. The condensed branch is $\gamma_c = c(2L)^{1/4}/\sqrt{2}$, flat in d_k and carrying the one calibrated constant; the mean-field branch is $\gamma_c^2 = 2L/(\sqrt{d_k} + \sqrt{d_k - 8L})$, with no free parameter. The crossover is $d_k^* = 8L = 49.9$, where the annealed bound saturates, and the step across it is exactly the factor c. (c) The coupling transition and the training transition are orthogonal: the first is a property of the operator, the second of the optimiser, and neither is an estimator of the other. The horizontal line is γ_c, where the reduced model concentrates target mass; it is not where the shear switches on or the channel becomes representable, both of which hold at every $\gamma > 0$ (Thm. T3.18, Cor. T3.5). §T6.14 measures trained models taking the single-layer route between $0.17\gamma_c$ and $0.45\gamma_c$ — inside the lower half of the panel.

Proposition T5.3 (Score-level from embedding-level perturbation size)**.** *Let* $\varepsilon_q := \|q^{\mathrm{AC}}\|/\|q^{\mathrm{DC}}\|$, $\varepsilon_k := \|k^{\mathrm{AC}}\|/\|k^{\mathrm{DC}}\|$, *and let* $c^{\mathrm{DC}} := (q^{\mathrm{DC}})^\top k^{\mathrm{DC}}/(\|q^{\mathrm{DC}}\|\|k^{\mathrm{DC}}\|) > 0$ *be the DC cosine similarity. Then the score-level ratio* $\varepsilon_S := \|S^{\mathrm{AC}}\|/\|S^{\mathrm{DC}}\|$ *obeys*

$$\varepsilon_S \le \frac{\varepsilon_q + \varepsilon_k + \varepsilon_q\varepsilon_k}{c^{\mathrm{DC}}}. \tag{T5.2}$$

Proof. Expanding, $S^{\mathrm{AC}} = (q^{\mathrm{AC}})^\top k^{\mathrm{DC}} + (q^{\mathrm{DC}})^\top k^{\mathrm{AC}} + (q^{\mathrm{AC}})^\top k^{\mathrm{AC}}$. Cauchy–Schwarz bounds each term by the product of the norms; dividing by $\|S^{\mathrm{DC}}\| = c^{\mathrm{DC}}\|q^{\mathrm{DC}}\|\|k^{\mathrm{DC}}\|$ and substituting the definitions gives (T5.2). The bound degrades as the DC cosine similarity approaches zero, which is the regime in which the split ceases to be informative. □

T5.2 The sixteen-term decomposition and the induction driver

Applying Definition T5.1 inside each of the four terms of Theorem T3.2 refines the decomposition into sixteen sub-terms, labelled T_i^{XY} with $X, Y \in \{D, A\}$ recording the DC/AC type of the query-side and key-side factors. Writing $p^{\mathrm{AC}}_{q,t} := q^{\mathrm{AC}}_t - q^{\mathrm{AC}}_{t-1}$ and similarly on the key side, each T_i splits into four.

Proposition T5.4 (The discriminative sub-terms)**.** *Consider the canonical induction pattern with query token* A *at position* t*, induction target at* j^* *whose predecessor is* A*, and distractors elsewhere. Under isotropy — token embeddings i.i.d. zero-mean with covariance* $\sigma^2 I_{d_k}$*, positions independent of content — the base term* T_1 *is constant in* j *and* T_2 *is mean-zero, while* two *sub-terms carry a non-zero expected gap between* j^* *and the distractors. The single-ghost term* T_3 *carries one at first order in* γ *(eq.* (T5.4) *in the proof below); at second order the carrier is*

$$T_4^{AA}(t,j) = \gamma^2 \big(p^{\mathrm{AC}}_{q,t}\big)^\top \big(p^{\mathrm{AC}}_{k,j}\big), \qquad \mathbb{E}\big[T_4^{AA}(j^*)\big] = -\gamma^2\sigma^2 d_k, \qquad \mathbb{E}\big[T_4^{AA}(j)\big] = 0 \;\; (j \ne j^*). \tag{T5.3}$$

Within block T_4*,* T_4^{AA} *is the unique sub-term with a non-zero gap; every remaining sub-term of* T_1*,* T_2 *and* T_4 *is either constant in* j *or mean-zero.*

Proof. Two facts are used throughout: under isotropy any inner product between independently sampled token embeddings is mean-zero; and the target j^* is distinguished only by its *predecessor* token being A, distractors having i.i.d. predecessors.

Block T_1**.** $T_1 = q_t^\top k_j$ does not reference the predecessor at all, so its expectation cannot depend on whether $j = j^*$. Its four sub-terms are therefore each constant in j: $\mathbb{E}[T_1^{DD}] = \sigma^2$ when the token identities coincide and the remaining three vanish by isotropy. Constant in j means non-discriminative, whatever the value.

Block T_2**.** $T_2 = p_{q,t}^\top k_j = q_t^\top k_j - q_{t-1}^\top k_j$ is a first difference on the *query* side. Both summands are expectations of inner products between a query-side embedding and k_j; neither references k_{j-1}, so again the predecessor plays no role and $\mathbb{E}[T_2(j^*)] = \mathbb{E}[T_2(j \ne j^*)]$. The common value is zero because the two contributions cancel: $\mathbb{E}[q_t^\top k_j]$ and $\mathbb{E}[q_{t-1}^\top k_j]$ are equal under exchangeability of the distractor positions. All four sub-terms inherit this.

Block T_3**.** $T_3 = q_t^\top p_{k,j} = q_t^\top k_j - q_t^\top k_{j-1}$ references k_{j-1}, and at $j = j^*$ the predecessor is the query token A, so $\mathbb{E}[q_t^\top k_{j^*-1}] = \sigma^2 d_k$ while $\mathbb{E}[q_t^\top k_j] = 0$ for every j. Hence

$$\mathbb{E}\big[T_3(j^*)\big] = -\sigma^2 d_k, \qquad \mathbb{E}\big[T_3(j)\big] = 0 \quad (j \ne j^*), \tag{T5.4}$$

so T_3 is discriminative. The two gaps differ in their order in the coupling: T_3 enters the score as γT_3 and so contributes $\gamma\sigma^2 d_k$, while T_4^{AA} as written in (T5.3) already carries its coefficient and contributes $\gamma^2\sigma^2 d_k$. The T_3 contribution is therefore the larger for $\gamma < 1$ and the smaller for $\gamma > 1$. Remark T5.5 records what this implies for the reduced model of §T5.6.

Block T_4**.** Here the four sub-terms behave differently. $T_4^{DD} = \gamma^2 (p_q^{\mathrm{DC}})^\top p_k^{\mathrm{DC}}$ involves only slowly varying components, which by construction do not resolve token identity at a specific position, so $\mathbb{E}[T_4^{DD}(j)] \approx 0$ uniformly in j. The cross terms T_4^{DA} and T_4^{AD} pair a DC factor with an AC first difference; under isotropy these are independent at a given position and both expectations vanish. That leaves T_4^{AA}. At $j = j^*$ the key-side difference is $p^{\mathrm{AC}}_{k,j^*} = k^{\mathrm{AC}}_{j^*} - k^{\mathrm{AC}}_{j^*-1}$, whose second term is the AC part of token A; the query-side difference $p^{\mathrm{AC}}_{q,t}$ references the same token. The inner product

therefore has expectation $-\gamma^2\sigma^2 d_k$, the sum running over the d_k coordinates and the sign arising from the subtraction in the key-side difference. At $j \neq j^*$ the predecessor is independent of the query token and the expectation is zero. Hence T_4^{AA} is the unique discriminative sub-term *within block* T_4; T_3 carries its own gap at first order by (T5.4). □

Remark T5.5 (Scope of the reduced model, and an open transfer question). Equation (T5.4) bears on what follows, and we state the consequence rather than leave it to be discovered. The score model of §T5.6, and hence Theorem T5.22 built on it, retains the γ^2 coherence channel and omits the γ-linear T_3 gap. That is a modelling choice and not a consequence of the decomposition: below $\gamma = 1$ the omitted term is the larger of the two. Three things follow, none of them concealed elsewhere. Theorem T5.22 is a prediction of the reduced model that retains T_4 alone, validated against simulation of that same model (§T6.9); it is not derived from the four-term decomposition without omission. The key-only corollary (Cor. T5.27) is built on precisely the T_3 channel omitted here, so the two laws do not share a common reduction — Remark T5.28 records a variance convention that also differs. And restoring T_3 would add signal at small γ, the direction that would move the score-model exponent towards the steeper one measured on trained models (§T6.11); whether it accounts for that gap is open and we do not claim that it does. §T9 carries that calculation out: it gives the explicit dual-channel relation obtained when T_3 is retained, and reports what changes.

Remark T5.6 (The sign, and why it does not matter). The gap is negative before any sign convention absorbed into W_Q, W_K: the head assigns the target a *lower* raw coherence score than the distractors, and it is the trained projections that orient the momentum vectors so the discriminative signal appears with usable sign. Every statement downstream depends only on the magnitude $|\Delta(\gamma)| = \gamma^2\sigma^2 d_k$ of the gap the reduced model retains.

T5.3 Signal-to-noise and the coupling-invariance of the ratio

Theorem T5.7 (Single-layer induction signal-to-noise)**.** *Under isotropy, a single-layer PSA head realises the induction discriminator without key composition, and the signal-to-noise ratio of the score gap between the target and a distractor is — with the numerator taken as the* magnitude *of the expected gap, since the raw gap is negative before learned orientation (Remark T5.6) and every statement downstream uses* $|\Delta(\gamma)|$ *—*

$$\mathrm{SNR} := \frac{|\mathbb{E}[s_{t,j^*}] - \mathbb{E}[s_{t,j}]|}{\mathrm{std}(s_{t,j^*} - s_{t,j})} = \frac{\gamma^2\sigma^2\sqrt{d_k}}{2\gamma^2\sigma^2} = \frac{\sqrt{d_k}}{2}, \tag{T5.5}$$

independent of γ *and growing as the square root of head width. The factor* 2 *in the denominator is the variance of a difference of two independent embeddings relative to one, and it is carried consistently through §T5.6:* (T5.5) *and* (T5.12) *are the same quantity.*

Proof. **Step 1.** This is the T_4-only signal-to-noise ratio: as in (M1) of Definition T5.20 we omit the first-order T_3 gap that Proposition T5.4 establishes, and retain the coherence channel alone. By that proposition $\mathbb{E}[T_1(j)]$ is constant in j and $\mathbb{E}[T_2(j)] = 0$, while the retained channel has $\mathbb{E}[T_4(j^*)] = -\gamma^2\sigma^2 d_k$ and $\mathbb{E}[T_4(j)] = 0$ otherwise, by (T5.3). **Step 2.** Summing, the expected gap is $\mathbb{E}[s_{t,j^*}] - \mathbb{E}[s_{t,j}] = -\gamma^2\sigma^2 d_k$, of magnitude $|\Delta(\gamma)| = \gamma^2\sigma^2 d_k$: the coherence channel is a sum of d_k independent coordinate products, so its mean is extensive in the head width. **Step 3.** Normalise the score consistently by $1/\sqrt{d_k}$, so that both the signal and the fluctuation are divided by the same factor:

$$a = \frac{\gamma^2\sigma^2 d_k}{\sqrt{d_k}} = \gamma^2\sigma^2\sqrt{d_k}, \qquad b = \frac{\gamma^2\sqrt{4\sigma^4 d_k}}{\sqrt{d_k}} = 2\gamma^2\sigma^2, \tag{T5.6}$$

the variance $4\sigma^4 d_k$ of $p_q^\top p_k$ arising because each of its d_k coordinate products pairs two first differences, each of variance $2\sigma^2$. **Step 4.** Forming the ratio, $\mathrm{SNR} = a/b = \sqrt{d_k}/2$, in which both γ^2 and σ^2 cancel, and it is essential that the same normalisation be applied to both: dividing the signal by d_k while leaving the variance undivided inverts the direction of the d_k dependence, and with it the sign of the exponent in the transition law derived below (Thm. T5.22); the present ratio does not depend on that law. □

Remark T5.8 (Coupling-invariance and what it forces). The cancellation in (T5.5) is the structural fact behind the whole of §T5.7: the coupling cannot improve the separation between target and distractor, because it scales signal and noise identically. This is the same statement as (T5.12) below,

reached by a second route. It follows that γ can only act as an inverse temperature in the softmax competition, and therefore that whatever transition exists must be a competition effect — a change in how the field of candidates is resolved, not a change in what the candidates look like. A construction in which γ improved the SNR would have no transition of this kind, and the observed sharpness would need a different explanation.

T5.4 The tangent simplex and the Fisher geometry

The linearisation of §T5.1 places the attention response in a geometric setting that is worth making explicit, because it determines the sense in which the DC operating point and the AC perturbation are independent, and because the naive Euclidean answer to that question is wrong.

Write Δ^{T-1} for the probability simplex, $\Delta_\circ^{T-1}$ for its interior, and $T_a\Delta^{T-1} = \{v \in \mathbb{R}^T : \mathbf{1}^\top v = 0\}$ for the tangent hyperplane at a: the zero-sum vectors.

Lemma T5.9 (The softmax Jacobian maps into the tangent hyperplane)**.** *Let* $a = \mathrm{softmax}(S^{\mathrm{DC}}) \in \Delta_\circ^{T-1}$ *and* $J_a = \mathrm{diag}(a) - aa^\top$. *Then* $J_a v \in T_a\Delta^{T-1}$ *for every* $v \in \mathbb{R}^T$*; moreover* $\ker(J_a) = \mathrm{span}(\mathbf{1})$ *and* $\mathrm{rank}(J_a) = T - 1$.

Proof. For the range: $\mathbf{1}^\top J_a = \mathbf{1}^\top \mathrm{diag}(a) - (\mathbf{1}^\top a)a^\top = a^\top - a^\top = 0$, using $\mathbf{1}^\top a = 1$; hence $\mathbf{1}^\top J_a v = 0$ for all v. For the kernel: $J_a\mathbf{1} = a - a(a^\top\mathbf{1}) = a - a = 0$, so $\mathbf{1} \in \ker(J_a)$. Conversely if $J_a v = 0$ then $a_i v_i = a_i(a^\top v)$ for each i, and since $a_i > 0$ on the interior this forces $v_i = a^\top v$, constant in i, so $v \in \mathrm{span}(\mathbf{1})$. The kernel is therefore one-dimensional and the rank is $T - 1$. □

Corollary T5.10 (The AC response is tangent)**.** *In the linearisation of Proposition T5.2, the correction* $a^{\mathrm{AC}} := J_a S^{\mathrm{AC}}$ *satisfies* $\mathbf{1}^\top a^{\mathrm{AC}} = 0$*: it redistributes attention mass without creating or destroying any.*

Proof. Immediate from Lemma T5.9, whose range statement gives $\mathbf{1}^\top J_a v = 0$ for every $v \in \mathbb{R}^T$; apply it with $v = S^{\mathrm{AC}}$. □

Definition T5.11 (Fisher–Shahshahani–Rao metric)**.** For $a \in \Delta_\circ^{T-1}$ define the *ambient* weighted bilinear form on the whole of $\mathbb{R}^T$,

$$\langle u, v\rangle_a^{\mathrm{amb}} := \sum_i \frac{u_i v_i}{a_i}, \qquad u, v \in \mathbb{R}^T, \tag{T5.7}$$

and let $\langle\cdot,\cdot\rangle_{F,a}$ denote its restriction to the tangent space $T_a\Delta^{T-1}$, which is the Fisher–Shahshahani–Rao metric.

Proposition T5.12 (Properties of the Fisher metric)**.** *The metric of Definition T5.11 is permutation-invariant; it is the Hessian of the Kullback–Leibler divergence, in that* $D_{\mathrm{KL}}(a + v \,\|\, a) = \frac{1}{2}\langle v, v\rangle_{F,a} + \mathcal{O}(\|v\|^3)$ *for tangent* v*; it agrees with the Euclidean metric, up to the factor* T*, only at the centroid* $a = \mathbf{1}/T$*; and by Čencov's theorem it is, up to scale, the unique metric monotone under Markov morphisms (coarse-grainings by sufficient statistics) [1].*

Proof. Permutation invariance is immediate from the componentwise definition. For the Hessian property, expand $D_{\mathrm{KL}}(a + v\|a) = \sum_i (a_i + v_i)\log\big(1 + v_i/a_i\big)$ in v: the term linear in v is $\sum_i v_i$, which vanishes by tangency, and the quadratic term is $\frac{1}{2}\sum_i v_i^2/a_i$, which is $\frac{1}{2}\langle v, v\rangle_{F,a}$. At the centroid $a_i = 1/T$ gives $\langle u, v\rangle_{F,a} = T\, u^\top v$. The uniqueness statement is Čencov's theorem. □

Proposition T5.13 (The tangent correction is ambient-orthogonal to the operating point)**.** *With* a *the DC operating point and* $a^{\mathrm{AC}} = J_a S^{\mathrm{AC}}$ *the linearised correction,*

$$\langle a, a^{\mathrm{AC}}\rangle_a^{\mathrm{amb}} = \sum_i \frac{a_i\, a_i^{\mathrm{AC}}}{a_i} = \mathbf{1}^\top a^{\mathrm{AC}} = 0. \tag{T5.8}$$

The statement is about the ambient form (T5.7), not the Fisher metric itself: a *is a point of the simplex and satisfies* $\mathbf{1}^\top a = 1$*, so it is not a tangent vector and* $\langle a, \cdot\rangle_{F,a}$ *is not defined. The orthogonality is exact and unconditional.*

Proof. The a_i cancel in the ambient form (T5.7), leaving the coordinate sum, which vanishes by Corollary T5.10. □

Remark T5.14 (Why the Euclidean answer is the wrong one). The Euclidean inner product $a^\top a^{\mathrm{AC}}$ is *not* zero in general: a has strictly positive entries while a^{AC} has zero sum and mixed signs, and the products $a_i a_i^{\mathrm{AC}}$ need not cancel. A statement of the form "the DC and AC parts of the attention response are orthogonal" is therefore false in the Euclidean inner product and true in the a-weighted ambient form (T5.7). The precision matters twice over. It matters because the weighted version is a geometric fact about the simplex — it needs no hypothesis about signal sizes — whereas a Euclidean version would require a subspace-separation assumption and would hold only approximately. And it matters because the weighted statement is *not* a statement in the Fisher metric: that metric is defined on tangent vectors, and a is not one (Prop. T5.13). What the Fisher metric does supply is the geometry of the tangent correction itself, once the base point is set aside.

Proposition T5.15 (The DC point is a fixed point, and PSA acts tangentially)**.** *Define the linearised attention map* $\mathcal{L}(S) := a + J_a(S - S^{\mathrm{DC}})$. *Then* (i) $\mathcal{L}(S^{\mathrm{DC}}) = a$*;* (ii) *the differential of* $\mathcal{L}$ *is* J_a *everywhere, symmetric, with eigenvalue* 0 *on* $\mathrm{span}(\mathbf{1})$ *and* $T-1$ *non-zero eigenvalues bounded by* $\max_i a_i \le 1$*; and* (iii) *replacing* S *by the PSA-augmented score leaves the DC component unchanged,* $\hat{S}^{\mathrm{DC}} = S^{\mathrm{DC}}$*, and alters only the tangent component, which is driven by the filtered score perturbation,* $\hat{a}^{\mathrm{AC}} = J_a\big(H^u \cdot S^{\mathrm{AC}}\big)$.

Proof. (i) is immediate. (ii): J_a is symmetric by inspection, its kernel is $\mathrm{span}(\mathbf{1})$ by Lemma T5.9, and $\|J_a\|_{\mathrm{op}} \le \|\mathrm{diag}(a)\|_{\mathrm{op}} = \max_i a_i$ since $aa^\top$ is positive semi-definite and subtracting it cannot increase the operator norm on the tangent space. (iii): the DC component is unchanged because the augmentation has unit DC gain, $H^u(e^{j0}) = 1$ exactly, by Proposition T3.4(i); magnitude alone would permit a sign flip; it is exactly this normalisation that condition (C3.1) of Theorem T3.18 imposes. The AC component is the filtered response by Proposition T3.6, and it is tangent by Corollary T5.10. □

Remark T5.16 (The geometric picture in one sentence). The DC operating point is where the head is attending on average; the augmentation does not move it, by DC neutrality, and instead reshapes the tangent perturbation about it, the reshaping being driven by a score perturbation whose high-frequency part has been amplified by $|H^u|$ *before* J_a acts. Whether that emphasis survives the projection is a separate question, settled by measurement rather than by Proposition T5.15 (§T6.6). Everything the construction does, post-softmax and at first order, is a tangent-space operation at a fixed base point.

T5.5 Cesàro cancellation of the single-ghost terms

One loose end from Proposition T5.4 deserves an independent positional-frequency argument. The purely positional sinusoidal components of T_2 and T_3 have zero Cesàro mean, even though by (T5.4) the *content* component of T_3 carries a candidate-wise gap that no positional averaging removes. Under a sinusoidal or rotary positional encoding the purely positional component of those terms also averages away along the key index, at rate $1/N$, under the structural hypotheses below. The statement is distribution-free only for that positional component: by (T5.4) the content contribution of T_3 carries a candidate-wise gap that no positional averaging removes.

Proposition T5.17 (Cesàro cancellation)**.** *Suppose the key stream carries a positional component at angular frequency* $\omega \notin 2\pi\mathbb{Z}$*,*

$$k_j^{\mathrm{pos}} \;=\; V\cos(\omega j + \phi) \;+\; W\sin(\omega j + \phi), \qquad V, W \in \mathbb{R}^{d_k},$$

with V *and* W *the pair of embedding directions that span the sinusoid and* ϕ *its phase, and suppose the query carries such a component at* $\omega' \neq \omega$*, with the positional directions orthogonal to the isotropic content component. Then*

$$\frac{1}{N}\sum_{j=1}^{N} T_3(t,j) \xrightarrow[N\to\infty]{} 0, \qquad \textit{and likewise for } T_2. \tag{T5.9}$$

Proof. Writing the key-side momentum of a sinusoid at frequency ω as a first difference gives $p_{k,j} = -2\sin(\omega/2)\big(V\sin(\omega(j-\tfrac{1}{2})+\phi) - W\cos(\omega(j-\tfrac{1}{2})+\phi)\big)$, so $T_3 = \gamma q_t^\top p_{k,j}$ is a fixed linear functional of a pure sinusoid in j. Cesàro averaging over $j \in [1, N]$ therefore reduces to

averaging $\sin(\omega j)$ and $\cos(\omega j)$, both of which tend to zero for $\omega \notin 2\pi\mathbb{Z}$ by orthogonality of distinct sinusoids. The argument for T_2 is identical with the roles of the streams exchanged. □

Proposition T5.18 (Rate)**.** *Under the hypotheses of Proposition T5.17,*

$$\Big|\frac{1}{N}\sum_{j=1}^{N} T_3(t,j)\Big| \;\le\; \frac{2\gamma}{N}\,\|q_t\|\big(\|V\|+\|W\|\big) \;=\; \mathcal{O}(1/N). \tag{T5.10}$$

Proof. The partial sum $\sum_{j=1}^{N} e^{ij\omega}$ is geometric with ratio $e^{i\omega} \neq 1$, hence bounded in modulus by $2/|1-e^{i\omega}| = 1/|\sin(\omega/2)|$ uniformly in N. The prefactor $2\gamma\sin(\omega/2)$ from the first difference cancels the $1/|\sin(\omega/2)|$, leaving a bound independent of ω; dividing by N gives the rate. The cancellation of the $\sin(\omega/2)$ factors is why the bound does not degrade at low positional frequency, where the partial sums are individually largest. □

Remark T5.19 (RoPE inherits the cancellation). RoPE applies an independent rotation in each coordinate pair, at its own frequency θ_i. Propositions T5.17 and T5.18 apply within each pair separately, and the sum over pairs of quantities each $\mathcal{O}(1/N)$ is $\mathcal{O}(1/N)$. The conclusion is that the pure sinusoidal *positional* contribution of the γ-linear channels averages away along the sequence while the γ^2 coherence channel does not. This is a statement about the averaged positional component only: a term with zero Cesàro mean can still carry a candidate-wise gap, and by (T5.4) the content contribution of T_3 does exactly that. The result therefore does not establish that T_2 and T_3 are non-discriminative.

T5.6 The reduced Gaussian score model

Everything in §T5.7 and §T5.8 is a statement about one explicit reduced model, which we name and state in full before using it. It is a model of the competition, not a derivation from the four-term decomposition, and the difference matters enough to put on the page.

Definition T5.20 (The T_4-only Gaussian score model $\mathcal{G}(d_k, T; \gamma)$)**.** Fix the canonical induction sequence model of Proposition T1.1, write $s = \hat{q}^\top \hat{k}/\sqrt{d_k}$ for the normalised PSA score at the query position, and let $L := \log(T-1)$ and $m = T-2$ be the number of distractors. The model $\mathcal{G}(d_k, T; \gamma)$ consists of the following four assumptions.

(M1) Single discriminative channel.
The target–distractor gap is carried by the double-ghost term $T_4 = \gamma^2 p_q^\top p_k$ alone. The base term T_1 is constant across candidates and T_2 is mean-zero (Prop. T5.4); the γ-linear gap that the same proposition establishes for T_3, eq. (T5.4), is *omitted*.

(M2) Gaussian moments.
Embeddings are taken of unit scale, $\sigma = 1$, a choice of units within $\mathcal{G}$; the T_4-only signal-to-noise ratio is invariant to this scale by (T5.6). After the $1/\sqrt{d_k}$ normalisation the retained channel contributes target mean a and per-distractor scale b with

$$a = \gamma^2\sqrt{d_k}, \qquad s_j = b\,Z_j \;\; (j \neq j^*), \quad b = 2\gamma^2, \tag{T5.11}$$

which is (T5.6) at $\sigma = 1$; the factor 2 in b arises because a difference of two independent embeddings carries twice the variance of one.

(M3) Independence.
The distractor variables Z_j are i.i.d. standard Gaussian. Independence is a modelling assumption, and §T6.9 measures its cost rather than assuming it away.

(M4) Annealed competition.
The characteristic coupling is defined by equality of the target's exponential weight with the *expected* aggregate distractor mass, a proxy for the argmax-accuracy midpoint and not an identity with it (step (2) of the proof of Theorem T5.22).

Two features of (T5.11) govern everything that follows. The signal-to-noise ratio

$$\frac{a}{b} = \frac{\sqrt{d_k}}{2} \tag{T5.12}$$

is independent of γ and grows with head width; it is (T5.5) of Theorem T5.7, reached here by a second route. And both a and b scale as γ^2, so within $\mathcal{G}$ the coupling enters the softmax competition purely as an inverse temperature: it sharpens the competition without changing which candidate is favoured. The transition is therefore a competition effect and not a signal effect, and its location is set by where the sharpening overcomes the entropy of the distractor field.

Stated in those terms, $\mathcal{G}$ *is* a random energy model in the sense of Derrida [14, 15], and it is worth writing the dictionary down once rather than leaving it to be reconstructed. The $m = e^L$ distractor columns are the REM's configurations; their scores Z_j, i.i.d. standard Gaussian by (M2)–(M3), are its energy levels; and $b = 2\gamma^2$ is its inverse temperature. The aggregate distractor mass $\sum_j e^{bZ_j}$ is then the REM partition function, and its annealed evaluation $\log \mathbb{E}[\sum_j e^{bZ_j}] = L + b^2/2$ is the REM's high-temperature free energy. The single feature of that model which the two-branch structure inherits is the one Derrida solved for: the annealed free energy is correct only down to a finite temperature, below which the partition function ceases to be an effective sum and becomes extreme-value dominated. What follows is that fact, read in the variables of an attention score.

One caution about the names. In the spin-glass literature the REM is itself a mean-field model, so both of the branches below are mean-field phases of it. *Mean-field branch* here is shorthand for the branch on which the annealed evaluation is the correct one — the high-temperature or paramagnetic phase in the usual nomenclature — and *condensed* for the frozen low-temperature phase. We keep the two labels because they are used throughout the empirical sections, but no claim about mean-field *versus* beyond-mean-field treatment is intended by them.

Remark T5.21 (What $\mathcal{G}$ is and is not). $\mathcal{G}$ is a reduced description of the competition among candidate scores, obtained from the exact decomposition of Theorem T3.2 by retaining one channel and idealising the distractor field as independent Gaussians. It is not equivalent to PSA, and results proved in it are not theorems about trained transformers. Three consequences are worth stating once, here, rather than distributed through the section. By (M1) the model is *missing* signal at small coupling, since the omitted T_3 gap enters at order γ against the retained channel's γ^2 and therefore dominates for $\gamma < 1$ (Rem. T5.5). By (M4) the predicted quantity is an annealed characteristic scale, whereas γ_c is extracted empirically at an accuracy midpoint; the two need not coincide and §T6.9 reports the discrepancy where it is largest. And the key-only corollary of §T5.8 is built on precisely the channel (M1) discards, so it is a statement in a *different* reduced model and its crossover is not comparable term by term with the one below (Rem. T5.28). What the model earns is that, within it, the transition location and the attainable accuracy are both closed-form and free of fitted parameters on the mean-field branch, and both are tested against simulation of $\mathcal{G}$ itself in §T6.9.

T5.7 The two-branch law

Theorem T5.22 (Two-branch transition in $\mathcal{G}(d_k, T; \gamma)$)**.** *Work in the T_4-only Gaussian score model $\mathcal{G}(d_k, T; \gamma)$ of Definition T5.20, under assumptions* (M1)–(M4)*, and let $L := \log(T-1)$. Let $\mathcal{A}(\gamma; d_k, T) := \mathbb{E}\big[\mathrm{softmax}_j(s)_{j^*}\big]$ denote the expected softmax mass the model places on the target — the* competition *observable, and not argmax classification accuracy. The distinction is essential rather than cosmetic: by* (M2) *both a and b scale as γ^2, so within $\mathcal{G}$ the coupling acts as a pure inverse temperature and the argmax ordering is* invariant *in γ; what moves is the mass the softmax allocates. $\mathcal{A}$ rises from $1/(T-1)$ (exact at $\gamma = 0$, by exchangeability) to the coupling-independent limit*

$$p_\infty(d_k, T) \;=\; \int_{-\infty}^{\infty} \varphi(z)\, \Phi\big(z + \tfrac{\sqrt{d_k}}{2}\big)^m \, dz, \qquad m = T - 2 \text{ (distractor count)}, \tag{T5.13}$$

with φ, Φ the standard Gaussian density and distribution, through a transition at the critical coupling

$$\gamma_c^2(d_k, T) \;=\; \begin{cases} \dfrac{2L}{\sqrt{d_k} + \sqrt{d_k - 8L}} & d_k \;\geq\; d_k^* \quad \text{(mean field)}, \\[2ex] \dfrac{c^2}{2}\sqrt{2L} & d_k \;<\; d_k^* \quad \text{(condensed)}, \end{cases} \qquad \begin{array}{l} L := \log(T-1), \\ d_k^* := 8L, \end{array} \tag{T5.14}$$

equivalently $\gamma_c = c\,(2L)^{1/4}/\sqrt{2}$ on the lower branch, where c is an $O(1)$ condensation constant calibrated *to 0.741 ± 0.015 — the mean and sample standard deviation of the nine pointwise estimates on the condensed points of the same grid the law is then compared against (§T6.10). Agreement on*

that branch is therefore a consistency check and not an out-of-sample test; the held-out claim of §T6.9 is confined to the mean-field branch, which carries no calibrated constant. On the mean-field branch $\gamma_c \sim (L^2/d_k)^{1/4}$ *as* $d_k \to \infty$*, and the annealed criterion saturates its validity bound* $b \le \sqrt{2L}$ *exactly at* d_k^**.*

Proof. **(1) Moments** (M1)–(M3)**.** The retained channel contributes target mean $a = \gamma^2\sqrt{d_k}$ and per-distractor noise bZ_j with $b = 2\gamma^2$, the Z_j i.i.d. standard Gaussian. These are the normalised moments (T5.6) of §T5.3, where the raw gap is taken in magnitude before the $1/\sqrt{d_k}$ normalisation; within $\mathcal{G}$ they are stipulated by (M2) rather than derived, and $\sigma = 1$. Both the signal and the variance are normalised by the same $\sqrt{d_k}$, so the ratio (T5.5) is γ-free and γ acts as a pure inverse temperature.

(2) Annealed criterion (M4)**.** We define the characteristic coupling by equality of the target's exponential weight with the *expected* aggregate distractor mass. This is a proxy for the argmax-accuracy midpoint at which γ_c is extracted empirically, not an identity with it: replacing a random log-sum-exp by its expectation is the annealed step, and the two definitions can differ. Evaluating the aggregate in that approximation and using the Gaussian moment generating function $\mathbb{E}[e^{bZ}] = e^{b^2/2}$,

$$e^a \;=\; m\,\mathbb{E}\big[e^{bZ}\big] = m\,e^{b^2/2} \quad\Longleftrightarrow\quad a = \log m + \tfrac{b^2}{2} = L + \tfrac{b^2}{2}, \tag{T5.15}$$

the substitution $\log m \to L$ absorbing an $O(1/T)$ shift.

(3) Quartic. Substituting (T5.11) into (T5.15) gives $\gamma^2\sqrt{d_k} = L + 2\gamma^4$, i.e.

$$2\gamma_c^4 - \sqrt{d_k}\,\gamma_c^2 + L = 0, \qquad \gamma_c^2 = \frac{\sqrt{d_k} \pm \sqrt{d_k - 8L}}{4}. \tag{T5.16}$$

The transition is the *first* crossing encountered as γ increases from zero, hence the smaller root; rationalising the numerator turns it into the conjugate form $\gamma_c^2 = 2L/(\sqrt{d_k} + \sqrt{d_k - 8L})$, which is the mean-field branch of (T5.14) and is numerically stable for large d_k where the two terms of the discriminant nearly cancel. Both roots are real precisely when $d_k \ge 8L$.

(4) Validity saturation. The annealed evaluation of (T5.15) is valid while the distractor field has not condensed, which for a Gaussian field of $m = e^L$ members is the Random-Energy-Model condition $b \le \sqrt{2L}$ [14, 15]. The reason is an entropy count, and it is worth giving because it is what makes the saturation below structural rather than coincidental. The annealed sum $\mathbb{E}[\sum_j e^{bZ_j}]$ is dominated by distractors with $Z_j \approx b$, of which there are $e^L\varphi(b)$, that is $e^{L-b^2/2}$ up to the Gaussian prefactor. That population falls below a single member at $b = \sqrt{2L}$, which is also $\mathbb{E}[\max_j Z_j]$ to leading order. Past that point the expectation is carried by configurations that do not exist in a typical draw, and the annealed evaluation over-counts. The quantity $L - b^2/2$ is the annealed entropy, and $b = \sqrt{2L}$ is where it vanishes: Derrida's freezing temperature, in the variables of (T5.11). At the crossing, $b_c = 2\gamma_c^2 = \frac{1}{2}(\sqrt{d_k} - \sqrt{d_k - 8L})$, which is strictly decreasing in d_k and satisfies $b_c = \sqrt{2L}$ exactly at $d_k = d_k^* = 8L$. The criterion therefore states its own domain of validity and saturates it at precisely the point where the branch structure changes: the boundary between the two regimes occurs where the annealed approximation ceases to hold, rather than at an externally imposed cut.

(5) Condensation. For $d_k < d_k^*$ no real crossing exists and the distractor sum is frozen at $b = \sqrt{2L}$: the completely frozen low-temperature phase of Derrida [15], in which the aggregate is dominated by its largest member, and the effective competition is with a single extreme value rather than with a mass.

Why the transition should sit *at* freezing rather than above it is the substance of this branch, and it is a scaling argument rather than a derivation. Below d_k^* the annealed crossing of step (2) has no real solution: the target's signal $a = \gamma^2\sqrt{d_k}$ cannot overtake the annealed aggregate at any coupling, because the aggregate grows in γ faster than the signal does. The competition can only turn over if the aggregate abandons that growth law, and in the REM it does so exactly at freezing: the free energy crosses over from the annealed $L + b^2/2$, which is $O(\gamma^4)$, to the extreme-value form $b\sqrt{2L}$, which is $O(\gamma^2)$ — the same order in γ as the signal $a = \gamma^2\sqrt{d_k}$ itself. The aggregate does not stop growing; it stops outgrowing the target, and beyond freezing it is controlled by the leading extremal scores rather than by the mass. So on this branch the transition is not located by a crossing at all; it is located by the freezing point itself. Setting $2\gamma_c^2 = \sqrt{2L}$ gives $\gamma_c = (2L)^{1/4}/\sqrt{2}$, and the $O(1)$ constant c absorbs finite-m freezing — the $\log\log m/\sqrt{\log m}$ corrections to $\mathbb{E}[\max_j Z_j]$ that are not

negligible at the m of a real context. This is the second branch of (T5.14). It should be read as a scaling law with one calibrated constant, not as a parameter-free prediction: that status belongs to the mean-field branch alone, and §T6.10 reports the condensed comparison as a consistency check rather than a held-out test.

The two branches therefore differ in what locates the transition, not merely in the formula that results. Above d_k^* it is a crossing inside a valid annealed evaluation; below, it is the failure of that evaluation. That is why they are not required to agree at the join, and Rem. T5.25 records that they do not.

(6) Endpoints. As $\gamma \to 0$ the score reduces to the exchangeable base term and the peak probability is exactly $1/(T-1)$. As $\gamma \to \infty$ the inverse temperature diverges and the accuracy tends to the probability that the target attains the maximum of the score field, which by (T5.5) depends on γ not at all; integrating the Gaussian argmax over the target value gives (T5.13).

Validation. On an 18-point (d_k, T) Monte-Carlo grid the mean-field branch is predicted at $r = 0.9876$ over nine points with zero fitted parameters (median error 2.6%), and the condensation constant agrees to 3% across two orthogonal axes (0.741 ± 0.015 on the nine condensed grid points; 0.717 on a fixed-d_k T-axis scan at on-branch $r = 0.9995$). Certificate C1 verifies step (2) and C3 step (3); C2 verifies the crossover $d_k^* = 8L$ that steps (3) and (4) share. C7 and C8 verify (T5.13) and (T5.16) against fresh Monte Carlo. §T6 reports the full comparison. □

Remark T5.23 (The quartic has two roots: the law predicts a band). Step (3) of the proof displays both roots of the mean-field criterion: writing $x = \gamma^2$, the quadratic $2x^2 - \sqrt{d_k}\, x + L = 0$ of (T5.16) has $x_\pm = \left(\sqrt{d_k} \pm \sqrt{d_k - 8L}\right)/4$. Only x_- is carried into (T5.14), and rightly so: it is the onset, the first crossing encountered as γ increases from zero. It is worth recording what the discarded root means, because it is not spurious. The annealed balance is restored at x_+ as well: beyond $\gamma_+ = \sqrt{x_+}$ the $e^{b^2/2}$ growth of the distractor aggregate, quadratic in $b = 2\gamma^2$, overtakes the target's linear-in-a gain. Within $\mathcal{G}$ the competition observable is therefore elevated on a *band* $[\gamma_-, \gamma_+]$ and not on a half-line. At $d_k = 128$, $T = 512$ the band is $[0.787,\ 2.245]$, and $\gamma_+ \to \infty$ only as $d_k \to \infty$.

Two cautions. The upper root inherits the annealed approximation of (M4) more heavily than the lower one, since it is reached where the distractor aggregate dominates and a mean is a worse proxy for a log-sum-exp; and γ_+ lies outside the small-γ regime in which the T_4-only truncation of Definition T5.20 is defensible at all. We therefore state γ_+ as a property of the reduced model rather than a prediction about trained transformers. Section A31.7 measures an upper cutoff at $d_k = 128$ near 0.9, a factor of 2.5 below γ_+ — but §A31.11 shows that cutoff to be a property of the constant learning rate rather than of the operator, so it is not a measurement of γ_+ and does not bear on it. Whether γ_+ is attained by a trained model is untested: no coupling above 1.00 has been run at a step size scaled to the coupling.

Remark T5.24 (Why γ_c falls with head width). The direction may seem counter-intuitive — a wider head needs *less* coupling — so it is worth stating the mechanism. By (T5.5) the per-candidate signal-to-noise ratio grows as $\sqrt{d_k}$, because the coherence channel is a sum of d_k independent products whose mean grows linearly while its standard deviation grows as $\sqrt{d_k}$. A wider head therefore arrives at the competition with a better-separated target and requires less sharpening to resolve it. The asymptotic law $\gamma_c \sim (L^2/d_k)^{1/4}$ makes the trade explicit: context length raises the bar logarithmically and width lowers it as a quarter power.

Remark T5.25 (The branches do not meet in value). The two expressions in (T5.14) are derived under different regimes — an annealed evaluation above d_k^*, a condensed extreme-value argument below it — and they are not required to agree at the join, nor do they. At $T = 15$, for instance, the mean-field expression evaluates to 1.072 at $d_k = d_k^* = 21.1$ while the condensed branch sits at 0.794. The measured γ_c passes through the crossover smoothly, so the discontinuity is a property of the piecewise description and not of the system. It is also the reason the largest prediction errors cluster on either side of d_k^*: approaching the crossover from the mean-field side the expression rises steeply towards $\gamma_c^2 = 2L/\sqrt{d_k^*}$ while the data does not. §T6.9 quantifies this and Fig. 10a displays it. A description that interpolates the two regimes — a full replica treatment rather than the annealed plus condensed pair used here — would remove the jump, and is left open.

T5.8 Corollaries delimiting the mechanism

The two half-lifts of the order lattice (T1.3) are the sharpest available controls on the account, because the theory predicts opposite outcomes for them.

Corollary T5.26 (Query-only $(1,0)$ cannot exceed the order-zero ceiling)**.** *For every $\gamma_q > 0$, the augmentation $(\rho, \rho') = (1,0)$ leaves the induction accuracy bounded by $1/(T-2)$, and in particular never reaches the two-layer induction solution.*

Proof. The augmentation modifies the query stream only, so the score remains a pointwise function of k_j and the key-side order is zero. Proposition T1.1 applies unchanged: conditional on $\hat{q}_t$, the sub-family $\{s_{t,j}\}_{j \neq j^*-1}$ is exchangeable and the peak probability is at most $1/(T-2)$. Within that hypothesis class the conclusion is independent of γ_q, of d_k, and of the training procedure: no amount of coupling on the wrong axis creates key-side order. The scope qualifier is load-bearing — the statement is about the canonical single-layer model with pointwise keys, and says nothing about a deep network whose earlier layers have already written predecessor information into k_j. □

Corollary T5.27 (Key-only $(0,1)$ transition law)**.** *The key-only half-lift admits the parallel transition law*

$$\gamma_c = \frac{2L+1}{\sqrt{d_k} + \sqrt{d_k - 4L - 2}}, \qquad d_k \geq 4L+2, \tag{T5.17}$$

with crossover $d^*_{k,(0,1)} := 4L+2$ and an accuracy ceiling higher than (T5.13) by a factor $\sqrt{2}$ in the effective signal-to-noise ratio.

Proof. **Step 1 (the score).** With the shear on the key stream alone, $\hat{q}_t = q_t$ and $\hat{k}_j = k_j + \gamma p_{k,j}$, so the normalised score is

$$s_{t,j} = \frac{q_t^\top k_j + \gamma\, q_t^\top p_{k,j}}{\sqrt{d_k}}. \tag{T5.18}$$

The augmentation contributes a single cross term, first order in γ, in place of the γ^2 coherence channel T_4 of Theorem T3.2. Only two of the four terms of (T3.2) survive, since the query stream is untouched.

Step 2 (moments). Under the isotropy convention of §T5.6, with embeddings i.i.d. zero-mean of covariance $\sigma^2 I_{d_k}$ and $\sigma = 1$: at the target position the key-side difference $p_{k,j^*} = k_{j^*} - k_{j^*-1}$ contains the embedding of the query token, so $\mathbb{E}[q_t^\top p_{k,j^*}] = -d_k$ and, after the $1/\sqrt{d_k}$ normalisation, the discriminative gap has magnitude

$$a' = \gamma\sqrt{d_k}. \tag{T5.19}$$

At a distractor both surviving terms are mean-zero and fluctuate: $\mathrm{Var}(q_t^\top k_j) = d_k$ and $\mathrm{Var}(q_t^\top p_{k,j}) = 2d_k$, the factor two again because a difference of two independent embeddings carries twice the variance of one. After normalisation the distractor scale is

$$b'^2 = 1 + 2\gamma^2. \tag{T5.20}$$

Here, unlike the symmetric case, the base term is retained: the discriminative channel is $O(\gamma)$ rather than $O(\gamma^2)$ and is therefore comparable to the base fluctuation at the couplings where the transition occurs, so it cannot be dropped. Remark T5.28 records the consequence.

Step 3 (annealed crossing). Substituting (T5.19) and (T5.20) into the crossing condition $a' = \log m + b'^2/2$ of step (2) of Theorem T5.22, with $\log m \to L$,

$$\gamma\sqrt{d_k} \;=\; L + \tfrac{1}{2}\big(1 + 2\gamma^2\big) \qquad \Longleftrightarrow \qquad \gamma^2 - \sqrt{d_k}\,\gamma + \big(L + \tfrac{1}{2}\big) = 0. \tag{T5.21}$$

The augmentation being first order in γ, this is a *quadratic* in γ where the symmetric case gave a quartic.

Step 4 (root and validity). The roots of (T5.21) are $\gamma = \big(\sqrt{d_k} \pm \sqrt{d_k - 4L - 2}\big)/2$. As in Theorem T5.22 the transition is the first crossing encountered as γ increases from zero, hence the smaller root; rationalising the numerator gives the conjugate form (T5.17), which is numerically stable at large d_k. Both roots are real precisely when $d_k \geq 4L+2 =: d^*_{k,(0,1)}$, which is the stated crossover.

Step 5 (ceiling). As $\gamma \to \infty$ the accuracy tends to the probability that the target attains the maximum of the score field, which depends on γ only through the limiting signal-to-noise ratio

$$\lim_{\gamma\to\infty} \frac{a'}{b'} = \lim_{\gamma\to\infty} \frac{\gamma\sqrt{d_k}}{\sqrt{1+2\gamma^2}} = \frac{\sqrt{d_k}}{\sqrt{2}}, \tag{T5.22}$$

against $\sqrt{d_k}/2$ for the symmetric operator (Theorem T5.7). The ratio of the two is $(\sqrt{d_k}/\sqrt{2})/(\sqrt{d_k}/2) = \sqrt{2}$, so the ceiling is (T5.13) evaluated with $\sqrt{d_k}/\sqrt{2}$ in place of $\sqrt{d_k}/2$: strictly higher, since Φ is increasing. □

Remark T5.28 (A difference of convention between the two laws). Theorem T5.22 retains the coherence channel alone in the distractor scale, giving $b = 2\gamma^2$; the derivation above retains the base term as well, giving $b'^2 = 1 + 2\gamma^2$. The asymmetry is forced by the order of the two channels in γ and not chosen: at $O(\gamma^2)$ the coherence term dominates the base fluctuation once the transition is reached, at $O(\gamma)$ it does not. The consequence is that the two laws are not derived under identical conventions, and the reader should not read the crossovers $8L$ and $4L + 2$ as differing only by the factor of two that the channel orders would suggest. Applying the key-only convention to the symmetric case would shift $L \mapsto L + \frac{1}{2}$ there, moving the mean-field prediction at $(d_k, T) = (64, 15)$ from $\gamma_c = 0.602$ to 0.664 against a measured 0.613 — which is why the symmetric law is stated in the channel-only form that §T6.9 validates. Equation (T5.17) itself has not been tested against simulation: §T6.5 measures the accuracy of the key-only variant, not its critical coupling, and we mark the law Conditional accordingly.

Remark T5.29 (The two corollaries as a designed test). Corollaries T5.26 and T5.27 are the account's most exposed predictions, and they were pre-registered before measurement. Corollary T5.26 in particular admits no free parameter and no graceful degradation: any observation of query-only performance above the order-zero ceiling $1/(T-2)$ falsifies the filter-order reading outright. The ceiling, not chance, is the falsifier: a value between $1/(T-1)$ and $1/(T-2)$ is consistent with Proposition T1.1 and would indicate only that the head suppresses the previous-token position rather than aligning with it. Corollary T5.27 predicts that the minimal member should not merely work but work *better* on this task than the symmetric completion, since its transition is first order in γ and its ceiling higher. §T6 measures 0.033 against chance 0.071 for query-only and 0.949 for key-only, and we take the second of these as a result about the task rather than about the construction: if single-layer induction alone were the objective, the key-only member would be the better engineering choice. The case for the symmetric member is the structural package of §T3.8, and we state it as such.

Remark T5.30 (Convention for the distractor count). Equation (T5.13) uses $m = T - 2$, the number of distractor candidates when one of the $T - 1$ scored positions is the target. Ensembles that score all $T - 1$ positions against an independently drawn target use $m = T - 1$; the two conventions differ by $O(1/T)$ in p_∞, below the error bars of every comparison reported here, and §T6 states which is in force wherever the ceiling is quoted numerically.

T6 Empirical signatures

This section reports the measurements. It is organised so that each claim of §§T1–T5 is met by the experiment designed to falsify it, in the order the claims were made, and so that every number is attached to the protocol and seed count that produced it. Results that came out against a pre-registered prediction are reported in the place where the prediction was made, with no reframing.

T6.1 Scope of the programme

The programme comprises approximately 9,377 training runs across two tranches, 83 figures and 36 appendix sections, and the forty-one notebooks of §T6.2, of which the self-contained seven-notebook battery of Table 5 was produced under a single matched protocol. Every entry in that battery carries deterministic seeds and embedded outputs, and each was executed end to end in a single session on the hardware stated. Appendix A1 maps each formal claim to the experiments that test it; Table 6 summarises the roadmap in the main text.

T6.2 The notebook corpus

Forty-one Jupyter notebooks accompany this paper, listed in Table 7. Thirty-seven ship with their outputs already embedded: each figure, each table and each number those notebooks produce can be read off without re-execution, without a GPU, and without access to the original training checkpoints. Seven of the forty-one — the battery of §T6.1 — were produced under the single matched protocol described there; the remaining thirty-four are the record of the main programme.

Table 5: The seven-notebook battery. Each notebook is self-contained, carries deterministic seeds, and ships with embedded outputs. Wall times are for the full (non-quick) configuration on the hardware noted; the certificate notebook requires no accelerator and no data.

Notebook	Protocol	Hardware	Wall time
Certificates C1–C10	SymPy ring normal forms and float64 numerics; ten independent algebraic and numerical checks	CPU	$<$ 1 min
Score-model d_k sweep	$d_k \in \{4, 8, 16, 32, 64, 128\} \times T \in \{10, 15, 25\}$; 5 seeds, 60,000 trials per point, 61-point γ grid on $[0, 8]$	A100	198 s
T-axis validation	$d_k = 16$ with $T \in \{6, 8, 9, 10, 12, 15, 20, 25, 30, 40\}$; same extraction	CPU	$<$ 2 min
Trained d_k ladder	$d_k \in \{4, \ldots, 64\}$, $d_{\text{model}} = 4d_k$, $d_{\text{ff}} = 16d_k$; 3 seeds $\times$ 3,000 steps on an eleven-point γ grid, $T-1 = 14$ pairs	A100	47.0 min
Matched control battery	Seven mechanisms, one protocol: $T-1 = 14$, $d_k = 16$, $d_{\text{model}} = 64$, 3 seeds $\times$ 3,000 steps, taps post-RoPE, V untouched	A100	24.2 min
Spectral rerun	Retraining from scratch at each of six couplings; $d_k = 16$, $T = 64$ with 30 pairs, 3,000 steps; 200 evaluation sequences	A100	1.9 min
Coupling/depth characterisation	$\gamma \in \{0, 1, 2, 3, 5, 8\} \times$ depth $\in \{1, 3\}$ plus asymmetric and key-only variants; 2 seeds; extended-length and inference-sweep probes	A100	14.6 min

Table 6: Empirical validation map. Each row links a theorem to its falsifiable prediction, the metric, the headline result, and the documented failure regime. Full protocols and seed statistics are in the cited appendices.

Theorem	Prediction	Metric	Result	Failure mode
Prop. T3.4	Closed-form $\lvert H(e^{j\omega})\rvert^2$	Pearson r vs. closed-form	r=0.895–0.9966 pre-softmax across γ>0	NS baseline: r=0.03 at γ=0.6 (Appendix A10, C15)
Thm. T4.1	$\lVert M_\gamma^\top J M_\gamma - J\rVert$=0 at operator level	Symplectic residual; HSF/τ scaling	0.0 exactly (C6); HSF $\pm$4% (Appendix A10)	τ=0.1 ablation: Bode $r\to$0.41 (Appendix A10)
Prop. T3.25	Symmetric score is the transported pairing	Symbolic identity	Exact; asymmetric current is no total difference (C9)	—
Prop. T3.27	Unequal shears symplectic; bias linear in ε	Residual; exact score difference	0.0 exactly (C4); bias exact (C5); trains to 0.875	—
Cor. T4.4	Coriolis penalty pre-RoPE	Controlled placement comparison	Post-RoPE better under the tested protocols, $\sim$2k runs (Appendix A30)	—
Thm. T5.22	$\gamma_c^2 = \frac{2L}{\sqrt{d_k}+\sqrt{d_k-8L}}$ ($d_k \geq 8L$), $\gamma_c^2 = \frac{1}{2}c^2\sqrt{2L}$ ($d_k < 8L$)	Held-out MC grid	r=0.9876 (9 pts, 0 params); 69.1$\pm$8.2% at $\gamma^\star$=2.0 vs. 1.6%	Depth law fails at L=30 (Appendix A7)
Cors. T5.26, T5.27	$(1, 0)$ capped at $1/(T-2)$; $(0, 1)$ viable	Matched battery	0.033 vs. 0.949, ceiling 0.077, chance 0.071	Learned two-tap attains 1.000 (pre-registered miss)
App. A16	Embedding-axis low-pass induction filter	r vs. θ	$r = -0.679$, $p = 9.9\times10^{-4}$	Multi-hop: null $\Delta = -0.4 \pm 1.2$ pp (Appendix A19)

The four Appendix Q notebooks, and why their output is not embedded. The exception is worth stating plainly rather than leaving a reader to discover it. The four Appendix Q sweeps ran for between two and seven days apiece on a single GB10, and in each case the interactive session to the machine closed before the sweep finished. The training was unaffected: it ran to completion, and each run's results were committed to disk as that run ended. What was lost was the streaming cell output that the notebook front-end would have embedded. Those four notebooks therefore carry their definitions, their invariant checks and their preflight output, and then break off partway through the sweep cell.

What they do not carry, they ship beside them. The complete results file, the per-evaluation ledger and the run log for each sweep travel with the notebook in `results/`, and every Appendix Q quantity

in this paper — every entry of Table 113, the emergence band, the depth-law comparison and the optimiser-conditioning verdict — is read from those files and not from a notebook cell. Nothing is lost but the transcript, for the same reason the sweeps were resumable in the first place: batches are a pure function of the pair (run key, step), results are written atomically as each run ends, and no reported quantity depends on console output. A reader who wants the numbers should open the results files; a reader who wants to reproduce them should run the notebook, which is self-contained and will regenerate both. The certificate notebook is the natural entry point: it depends on nothing else, requires no accelerator and no data, and checks its ten identities in under a minute of CPU time.

Table 7: The forty-one accompanying notebooks. "Where" gives the section or appendix in which the notebook's results are reported. Outputs are embedded throughout except in the four Appendix Q notebooks marked †, whose sweeps outlived their interactive sessions and which ship with their complete result files, ledgers and logs instead (§T6.2). No notebook needs to be run to read its results.

Notebook	Where	What it establishes
Battery (§T6.1): seven notebooks, one matched protocol.		
`PSA_Certificates`	§T6.15	Certificates C1–C10: the annealed crossing, the crossover $d_k^*=8L$, the quartic root and its asymptote, symplecticity and energy preservation of unequal shears, the exact score bias, the operator residual, the plateau quadrature, the quartic against fresh-seed Monte Carlo, the transport identity, and Pearson scale invariance
`PSA_MC_dk_sweep_with_results`	§T6.9	Score-model Monte-Carlo grid, $d_k \times T$, five seeds and 60,000 trials per point; the held-out mean-field test, r=0.9876 over nine points with no fitted parameter
`PSA_T_axis_condensed_validation`	§T6.10	The condensed branch along the T axis at fixed d_k=16; r=0.9995 on branch, and the condensation constant on a second axis
`PSA_trained_dk_overlay_with_results`	§T6.11	The trained d_k ladder: γ_c falling with head width, plateaus against the zero-parameter score-model argmax limit, and the measured operating-point ratio
`PSA_control_battery_results`	§T6.5	Seven mechanisms under one protocol: query-only never above chance, key-only strongest, low-pass lagging, and the learned two-tap converging to a pure key-side difference
`PSA_spectral_diagnostic_rerun`	§T6.6	Bode correlations pre- and post-softmax on models retrained from scratch; the total-AC-power normaliser and the exactly-zero symplectic residual
`PSA_stability_characterisation_results`	§T6.12	Divergence, depth-wise score growth, the inference-time coupling plateau and extended-length evaluation — reported as measurement, not theorem
Main programme: thirty-four notebooks.		
`Appendix_C`	A2	RoPE construction and the post-RoPE pipeline
`Appendix_D`	A3	The high-pass condition: EMA β-sweep
`Appendix_E`	A4	The phase transition in both positional-encoding families, granular γ-sweep
`Appendix_F1`	A13	The bandpass spectral window: exploratory low-frequency scan
`Appendix_F2`	A14	The rigorous granular sweep: 720 runs, 240 configurations
`Appendix_F3`	A15	The monochromatic control: ruling out frequency-hopping
`Appendix_G`	A16	The low-pass induction filter: the 2,000-experiment (γ, θ) sweep
`Appendix_H`	A17	RoPE frequency design space and band-by-band sensitivity
`Appendix_I_Mechanistic_Visualization`	A18	Attention-pattern visualisation across the coupling range
`Appendix_J_CoT_Reasoning_NB1`	A19	Multi-hop reasoning with eased difficulty: a documented null
`Appendix_J_CoT_Reasoning_NB2`	A20	Chain-of-thought reasoning on harder tasks
`Appendix_K_Real_World_Reasoning`	A21	Real-world reasoning tasks under momentum augmentation
`Appendix_L_Multi_Task_Validation`	A22	Multi-task validation and the $\nabla/\int$ dissociation control
`Appendix_M_Multi_Difficulty_Validation`	A23	The 2,880-run multi-difficulty sweep locating the sweet spot
`Appendix_N_NB_5`	A25	Burstiness null: the construction is neutral on $\int$-tasks
`Appendix_N_NB_2`	A26	Structural ICL without anchoring: the first failure
`Appendix_N_NB_4`	A27	Chained ICL without anchoring: diagnosis hardened
`Appendix_N_NB_3`	A28	Anchored ICL: the fix
`Appendix_N_NB_1`	A29	Anchored ICL stress test at $L = 30$
`Appendix_O_P`, `Appendix_P`	A30	Placement contrast: embedding-level against Q/K-post-RoPE, with spectral forensics
`Appendix_Q_Stability`†	A31	Low-coupling sweep at 91.7 M parameters: contraction of the trained attention sublayer
`Appendix_Q_Final_v2`†	A31.6	Width sweep at 91.3 M: four head widths, $\gamma \leq 0.40$, 49 runs, stopped by design
`Appendix_Q_bracket_dk128`†	A31.7	Supercritical sweep at d_k=128: the emergence band, 17 runs
`Appendix_Q_cutoff_decisive`†	A31.11	Five-arm optimiser-conditioning intervention, 11 runs

Table 7, continued

Notebook	Where	What it establishes
`Appendix_R_NB_1, _2`	A32	Parameter-efficiency studies at scale, including the reported near-misses
`Experiment_16_ Single_Layer_Induction`	A5	The single-layer induction limit
`Experiment_17_ Scaling_Law_Final`	A6	The momentum–depth sweep across L
`Experiment_18_ Granular_Scaling_Law`	A7	Granular validation at L=30, including where the depth law fails
`PSA_Induction_Validation_ Suite_v7`	A8	Validation suite v7: twelve sections, twenty-four checks
`PSA_Induction_Validation_ Suite_v8`	A9	Validation suite v8: two targeted fixes
`PSA_Induction_Validation_ Suite_v9`	A10	Validation suite v9: ten experiments, including the T_4 ablation and the head-width sweep
`Appendix_R2_Mechanism`	A34	Which circuit executes induction: randomised-offset task, held-out probe and per-layer γ ablation over depth and coupling (§T6.14)

T6.3 Spectral character: two axes, not one

The two spectral measurements below live on *distinct axes* and are easily confused. PSA is a per-stream *score-axis high-pass* (Thm. T3.18; Fig. 8, Pearson r up to 0.9966). The cascade RoPE $\rightarrow$ PSA is *embedding-axis low-pass* (Fig. 9, $r{=}-0.679$ between RoPE centre frequency θ and momentum gain). These are dual rather than in tension: the score-axis high-pass first difference detects the embedding-axis low-frequency structure that Xiong et al. [51] identify as where RoPE concentrates structured energy.

T6.4 The single-layer induction phase jump

Figure 6 reports the headline measurement. In panel (a), single-layer PSA reaches $69.1 \pm 8.2\%$ over five seeds at $\gamma^{\star}{=}2.0$, best seed 79.4%, against $1.6 \pm 0.4\%$ for the single-layer and $1.6 \pm 0.3\%$ for the two-layer standard-attention baseline at matched training budget — both at chance, $1/64 = 1.6\%$ — a $+67.5$ pp mean phase jump. The plotted sweep is a separate single-model run evaluated on three seeds, whose peak at the same coupling is 74.2% against 1.27% at $\gamma{=}0$. Appendix A6 reports the $L \in \{10, 14, 22, 30\}$ depth sweep with the predicted γ_c shift verified; Appendix A7 the $L{=}30$ scaling-law breakdown; Appendix A8 the twenty-four-sub-check audit, all of which pass. Panel (b) of the same figure is a separate claim on a deliberately different task, and the two are not blended: it sweeps both positional-encoding families under one matched protocol (vocabulary 200, 12 KV pairs, $d_k{=}32$, $d_{\text{model}}{=}128$, four heads, four layers, three seeds), and measures $\gamma_c^{\text{RoPE}} = 0.225$ against $\gamma_c^{\sin} = 0.275$ — a ratio of 1.22 against the parameter-free 1.000 that Theorem T5.22 predicts, the law carrying no positional-encoding term (Appendix A4). Beyond the transition window the two families converge as well as sharing a boundary: both sit between 98.5% and 99.6% across $\gamma \in [0.7, 3]$. At $\gamma = 5$ RoPE becomes unstable, $68.4 \pm 43.1\%$ over three seeds against $98.6 \pm 0.2\%$ for sinusoidal, which we report as measured. The whole 156-run sweep was re-executed from the accompanying notebook under a fixed seed and reproduced every reported accuracy to the tabulated precision.

T6.5 The order lattice, measured

Notebook: `PSA_control_battery_results`.

Section T1.3 places four candidate augmentations on an order lattice and predicts a different outcome for each. Figure 7 and Table 8 measure all four, together with three further mechanisms, under a single protocol: $T{-}1{=}14$ induction pairs, $d_k{=}16$, $d_{\text{model}}{=}64$, three seeds $\times$ 3,000 steps, taps applied post-RoPE with the value stream untouched, chance $1/(T{-}1) = 0.071$. Every entry is a separate training run under identical conditions; nothing is shared between variants but the data and the schedule.

Reading the lattice. Four results follow directly and none required a free parameter. *First*, the query-only variant stays far below the order-zero ceiling $1/(T-2) = 0.077$ at every coupling

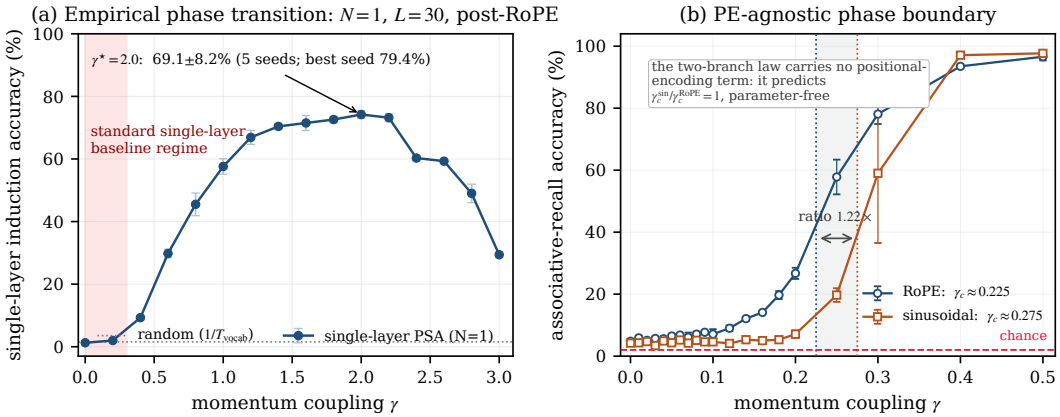


Figure 6: Single-layer induction phase transition. *The two panels measure complementary quantities on deliberately different tasks; their absolute accuracies are not directly comparable.* **(a)** *Single-layer induction under distractor noise* (N=1, "nano" decoder, ∼54 k parameters, d_{model}=64, 4 heads, d_k=16; Appendix A8). Vocabulary 64 with *shared* key/target alphabet (key-vs-target role learnable only from position), 14 KV pairs, padded sequence length 30 (padding tokens act as distractors); the plotted sweep is one trained model evaluated on three seeds, peaking at 74.2%, and the statistic annotated at $\gamma^\star$=2.0 is an independent five-seed mean ± s.d. Standard attention (γ=0) does not exceed chance, $1/64 = 1.6\%$: it measures 1.27% on the plotted sweep and $1.6 \pm 0.4\%$ across the five seeds. Single-layer PSA reaches $69.1 \pm 8.2\%$ over those five seeds (best seed 79.4%), a +67.5 pp mean phase jump that demonstrates single-layer induction recovery *under noise*, in a regime where the standard bilinear head does not. The shaded band marks the low-coupling range over which the head is not measurably above the standard single-layer baseline; it is not a region excluded by the SHT bound, which constrains description size rather than γ. **(b)** *PE-agnostic phase boundary on a sanitised multi-layer setup* (N=4 decoder, ∼842 k parameters, d_{model}=128, 4 heads, d_k=32, d_{ff}=512; Appendix A4). Vocabulary 200 with *disjoint* key/target sub-alphabets (keys $\in [1, 100)$, targets $\in [100, 200)$; role readable from token ID), 12 KV pairs, no padding distractors—designed to locate γ_c cleanly across PE families. The panel plots the transition window $\gamma \in [0, 0.5]$; the sweep continues to $\gamma = 5$, where both families saturate together — 98.5–99.6% across $\gamma \in [0.7, 3]$, so the two curves converge as well as sharing a boundary. The one exception is $\gamma = 5$, at which RoPE collapses to $68.4 \pm 43.1\%$ over three seeds while sinusoidal holds at $98.6 \pm 0.2\%$: at that coupling one seed fails outright, and we report the instability rather than truncating the sweep before it. Peak accuracy within the plotted window is 96.6% (RoPE) and 97.7% (sinusoidal); the boundary is PE-agnostic to within 1.22× ($\gamma_c \approx 0.225$ RoPE, $\gamma_c \approx 0.275$ sinusoidal), against the parameter-free 1.000 that Theorem T5.22 predicts, the law carrying no positional-encoding term. Panel (a) tests *single-layer recovery under noise*; panel (b) tests γ_c *scaling on the multi-layer regime where Theorem T5.22 applies*—hence the ∼16× parameter gap and the absolute-accuracy gap.

Table 8: Matched control battery. Best accuracy over $\gamma \in \{0.5, 1, 2, 3\}$ (asymmetric variants over six (γ_q, γ_k) pairs), mean ± s.d. over three seeds. Protocol as in the text; chance 0.071. The order column gives the member of the lattice (T1.3) that the variant realises.

Variant	Order	Best accuracy	Attained at	Reading
One-layer baseline	$(0, 0)$	0.033 ± 0.006	—	below the 0.077 ceiling; cf. Prop. T1.1, Rem. T1.2
Query-only shift	$(1, 0)$	0.033 ± 0.007	γ=0.5	same class, same value at every γ; cf. Cor. T5.26
Key-only shift	$(0, 1)$	0.949 ± 0.006	γ=3.0	minimal member; strongest fixed variant
PSA symmetric	$(1, 1)$	0.811 ± 0.006	γ=3.0	rises monotonically from 0.505
Asymmetric (γ_q, γ_k)	$(1, 1)^*$	0.875 ± 0.008	$(1.0, 3.0)$	bias absorbed, as Prop. T3.27 predicts
Low-pass two-tap	non-DC-neutral	0.428 ± 0.018	γ=3.0	differencing, not generic context
Learned two-tap	learned	1.000 ± 0.000	—	taps b/a: −0.010 (q), −1.040 (k)
Two-layer baseline	$(0, 0)$, N=2	0.036 ± 0.001	3,000 steps	0.024 ± 0.006 at 9,000 steps

tested — 0.033, 0.030, 0.033, 0.030 at $\gamma = 0.5, 1, 2, 3$, against a uniform-guess benchmark of $1/(T-1) = 0.071$. That it sits *below* chance rather than at it is itself predicted: Remark T1.2 identifies the mechanism, since the previous-token position $j^* - 1$ is the one candidate an order-zero key filter can single out, and a head that aligns with it spends its argmax there. The order-$(0, 0)$ one-layer baseline measures the same 0.033 (Table 8), as it should, the two variants lying in the same hypothesis class. Corollary T5.26 bounds the population quantity and these are three-seed trained means, so we claim no equality; what the measurement establishes is the absence of any above-ceiling signal, which is the discriminating prediction. This is the discriminating negative control: leaving the $(0, 0)$ class is necessary but not sufficient, and any result above the $1/(T-2)$

ceiling here would have falsified the filter-order reading outright. *Second*, the key-only variant is the strongest fixed member of the battery, rising $0.589 \to 0.770 \to 0.905 \to 0.949$ across the same grid (Fig. 7b), as Corollary T5.27 anticipates: key-side order ≥ 1 is the representability requirement, and predecessor exposure on the key stream — the Ghost Key of Definition T3.3 — is the enabling channel. This is the paper's sharpest negative result about its own construction, and Fig. 7a puts the two side by side: 0.949 for the half-lift against 0.811 for the symmetric operator, a gap of 13.8 pp that no coupling in the grid closes (Fig. 7b, where the key-only curve dominates the symmetric one at every γ tested). We state the engineering consequence plainly: *on this task the half-lift is the better choice*. Both operators are exactly symplectic (Prop. T3.27), so the gap is not a structural deficiency of one of them — it is that the symmetric operator spends half its coupling budget on the query stream, which Corollary T5.26 shows contributes nothing to representability. What the symmetric member buys is set out in Table 2 and delivered in §T3.8, §T5 and §T6.13; none of it is accuracy on this task. *Third*, a two-tap low-pass of matched parameter count injects exactly one step of local context and reaches 0.428 against 0.811 for the symmetric shear at the same coupling. The operative ingredient is DC-neutral high-pass differencing, not one-step context as such. *Fourth*, the asymmetric variants train successfully and show a consistent key-heavy asymmetry: $(1, 3)$ reaches 0.875 while $(3, 1)$ reaches 0.630, and $(0.5, 2)$ reaches 0.864 while $(2, 0.5)$ reaches 0.330. This is consistent with Proposition T3.27 — unequal coupling costs an absorbable score bias and not a loss of symplecticity, though swap-symmetric transport is lost — though that proposition is an algebraic statement about the score and predicts none of these trained accuracies; the ordering within the asymmetric pairs is read as the stream-placement effect again.

The learned filter. The sharpest result in the battery is the one we did not predict correctly. A general first-order filter on each stream, with freely trained coefficients and no constraint tying the streams, converges to $b/a = -1.040$ on the key stream and -0.010 on the query stream (Fig. 7c): a pure key-side first difference, with the query stream left as the identity. It reaches 1.000. Given the whole two-tap family, gradient descent selects the DC-neutral differencing form on the axis the filter-order analysis identifies. The fixed operator at γ=2 sits at $b/a = -\gamma/(1+\gamma) = -0.667$ on both streams, so the learned solution is the *minimal* member of the lattice and not the symmetric one.

Two pre-registered predictions that did not hold. Both are reported as measured. The learned two-tap was predicted to land near the symmetric PSA setting; it does not, placing essentially no weight on the query stream. The two-layer baseline was predicted to learn the task at some budget; it does not, at either budget tested (0.036 at 3,000 steps, 0.024 at 9,000). We therefore make no impossibility claim for two-layer standard attention — the task is demonstrably learnable in 3,000 steps by the learned two-tap — and the comparison is one of parameter and budget efficiency at matched cost.

Table 9 sets these measurements beside the correspondence of Fig. 1, one row per member of the design space. Every row is consistent with the filter-order account, including the two rows that are not the construction we advocate; the integrator column records an analogy, not a measured correspondence.

Table 9: The correspondence of Fig. 1, with the attention column measured. Every row lands where the filter-order account places it, including the two rows that are not the construction we advocate. Accuracies from Table 8; chance 0.071.

Attention side	Filter order	Integrator analogue	Measured
Standard bilinear	$(0, 0)$	no lift: the obstruction	0.033
Query-only shear	$(1, 0)$	first-order rule, wrong axis	0.033
Key-only shear	$(0, 1)$	exact flow of $\frac{1}{2}\|p_k\|^2$; no swap symmetry	0.949
Asymmetric $\gamma_q \neq \gamma_k$	$(1, 1)^*$	exact flow of a two-coefficient quadratic; no transport identity (C9)	0.875
Low-pass two-tap	non-DC-neutral	dissipative smoother	0.428
Learned two-tap	learned	discovers $(0, 1)$ from data	1.000
PSA symmetric	$(1, 1)$	free-particle drift of the Verlet splitting; swap-symmetric transport (C9)	0.811 + the theory

"+ the theory" in the last row of Table 9 denotes what the symmetric member carries and the others do not: the γ^2 coherence channel and the closed-form transition of Theorem T5.22, the transport identity of Proposition T3.25, and the Differential Transformer linearisation of §T6.13. Both classical integrator poles are now measured in the attention column.

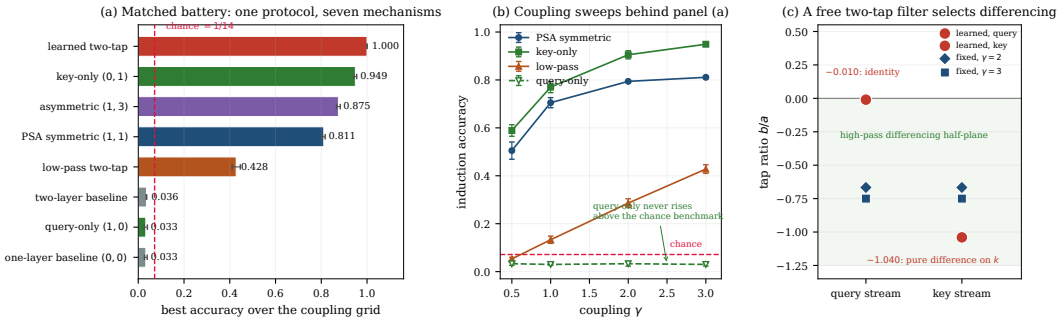


Figure 7: Matched control battery: one protocol, seven mechanisms. $T-1{=}14$ induction pairs, $d_k{=}16$, $d_{\text{model}}{=}64$, three seeds $\times$ 3,000 steps, taps post-RoPE, value stream untouched, chance $1/14 = 0.071$. **(a)** Maximum accuracy attained by each variant over its coupling grid — a best-of-k statistic, upward biased as an estimate of held-out performance, but applied identically to every variant; error bars are the seed s.d. The key-only half-lift (0.949) outperforms the symmetric operator (0.811) by 13.8 pp: if single-layer induction accuracy is the objective, the half-lift is the better construction, and what the symmetric operator buys instead is structure (Table 2). **(b)** The coupling sweeps behind those maxima. The query-only half-lift never rises above the chance benchmark at any γ, consistent with Corollary T5.26; the low-pass member of matched parameter count rises but lags every differencing variant. **(c)** The learned two-tap filter's converged coefficients against the fixed operator's. A free filter with no constraint tying the streams selects the identity on the query stream and a pure first difference on the key stream. The shaded half-plane $b/a < 0$ is the high-pass differencing regime.

T6.6 Bode forensics for the score stream

Notebook: `PSA_spectral_diagnostic_rerun`.

Proposition T3.4 gives the per-stream transfer function $|H(e^{j\omega})|^2 = 1 + 4\gamma(1+\gamma)\sin^2(\omega/2)$. On models retrained from scratch at each coupling ($d_k{=}16$, $d_{\text{model}}{=}64$, $T{=}64$ with 30 induction pairs, 3,000 steps, fixed seed) we measure pre-softmax AC power across $\omega \in [0, \pi]$ over 200 evaluation sequences and Pearson-correlate against the theory curve. Figure 8 plots the closed form in panel (a) and the measured correlations in panel (b); the values are $r = 0.8951, 0.9890, 0.9915, 0.9966, 0.9924$ at $\gamma = 0.6, 1.2, 1.8, 2.0, 3.0$, the maximum falling at $\gamma^\star{=}2.0$. The weakest point is the smallest coupling, where the AC power the filter produces is smallest in absolute terms and the measurement is correspondingly noisier. Standard attention ($\gamma{=}0$) is order-$(0, 0)$, produces a flat AC spectrum, and gives a degenerate correlation. The same measurement taken after the softmax gives r between 0.9393 and 0.9647 across the same grid. This is a test of spectral *shape*: both spectra are normalised before correlating and Pearson r is invariant under positive rescaling of either argument (certificate C10), so the measurement carries no information about absolute AC energy and cannot test the band-averaged power gain of Theorem T3.7(i). That would require an amplitude-sensitive statistic, which we do not report. Algorithm 3 states the protocol ($O(Td_k \log T)$ per γ); Appendix A8 records the v7–v9 audit of sixty-four sub-checks, the v6$\rightarrow$v7 mask-and-scaling fix, and the DC-band ablation.

Two properties of the protocol, both certificate-backed. The spectra are normalised by total AC power (Alg. 3, line 10). Normalising at the DC bin instead divides by a near-zero quantity: after per-row DC subtraction the residual DC bin carries a mean 1.4×10^{-3} of total pre-softmax power across the grid, ranging from 5.7×10^{-4} at $\gamma{=}3.0$ to 4.0×10^{-3} at $\gamma{=}0.6$. Pearson r is invariant under positive rescaling of either argument (certificate C10), so the choice of normaliser cannot move the reported correlations; measured directly across the grid, the largest difference is 2.2×10^{-16}, which is to say none. Separately, the symplectic character of the operator is reported as a residual and not as a determinant, for the reason given in Remark T4.2: at $\gamma{=}2.0$ the measured $\|M_\gamma^\top J M_\gamma - J\|_{\max}$ is 0.0 exactly, the blocks being unit-triangular (certificate C6).

Algorithm 3 Bode spectral forensics for the pre-softmax score stream.

Require: trained PSA model with coupling γ, evaluation set $\mathcal{D}$ of B sequences of length T, frequency grid $\Omega = \{\omega_m = \pi m/M\}_{m=0}^{M-1}$
Ensure: Pearson correlation $r(\gamma)$ between theoretical and measured spectra

1: $\hat{S} \leftarrow 0 \in \mathbb{R}^M$ *# accumulator for empirical AC power*
2: **for** each sequence $x^{(b)} \in \mathcal{D}$ **do**
3: compute $Q^{(b)}, K^{(b)}$ post-RoPE; form $\hat{Q}^{(b)}, \hat{K}^{(b)}$ as in Alg. 2
4: $S^{(b)} \leftarrow \hat{Q}^{(b)} \hat{K}^{(b)\top}/\sqrt{d_k}$ *# pre-softmax, pre-mask score $T \times T$*
5: subtract per-row DC: $S^{(b)}_{t,j} \leftarrow S^{(b)}_{t,j} - \frac{1}{T}\sum_{j'} S^{(b)}_{t,j'}$ *# isolate the AC component*
6: for each t, $\mathcal{F}^{(b)}_t \leftarrow \mathrm{DFT}_j(S^{(b)}_{t,:})$
7: $\hat{S} \leftarrow \hat{S} + \frac{1}{BT}\sum_t |\mathcal{F}^{(b)}_t|^2$
8: **end for**
9: $H_{\mathrm{th}}(\omega_m) \leftarrow 1 + 4\gamma(1+\gamma)\sin^2(\omega_m/2)$ *# score-level theory $H_{\mathrm{score}} = |H^u|^2$ (Prop. T3.4)*
10: normalise both spectra by total AC power: $\hat{S} \leftarrow \hat{S}/\sum_m \hat{S}[m]$, $H_{\mathrm{th}} \leftarrow H_{\mathrm{th}}/\sum_m H_{\mathrm{th}}(\omega_m)$ *# r is invariant under this choice (C10)*
11: $r(\gamma) \leftarrow \mathrm{Pearson}\big(\hat{S},\, H_{\mathrm{th}}\big)$ *# across the M frequency bins*
12: **return** $r(\gamma)$

The pre-softmax measurement (line 4) is what Theorem T3.18 speaks to; the post-softmax measurement degrades to $r \sim 0.94$–0.97, the spectral shadow of the Hairer Survival Factor discussed next. The DC subtraction (line 5) implements the DC/AC decomposition of §T5.1.

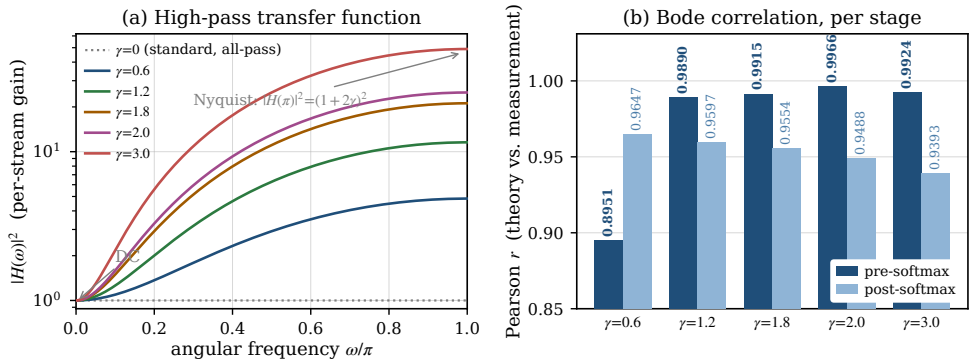


Figure 8: Bode forensics for the score stream. **(a)** The closed-form per-stream transfer function $|H(e^{j\omega})|^2 = 1 + 4\gamma(1+\gamma)\sin^2(\omega/2)$ (Prop. T3.4) is high-pass for every $\gamma > 0$ and reduces to all-pass at $\gamma = 0$ (standard attention). The Nyquist gain is $(1+2\gamma)^2$. **(b)** Measured Pearson correlation between the theoretical transfer function and the measured AC power, per γ, taken before and after the softmax. Protocol: one single-layer PSA model retrained from scratch at each γ (d_k=16, d_{model}=64, T=64 with 30 induction pairs, 3,000 steps, fixed seed), evaluated on 200 held-out sequences; both spectra normalised by total AC power. Standard attention is order-$(0,0)$ and produces a flat AC power spectrum, giving a degenerate correlation.

T6.7 What survives the softmax

Theorem T4.1 is a pre-softmax statement; the question is how much of the high-pass character survives the non-linearity, with the first-order Fréchet linearisation in γ as the licensing condition. **(i) Tangent structure.** The theory is in §T5.1–T5.4: the softmax linearises as Proposition T5.2 with quadratic remainder, the response $J_a\,\delta s$ is tangent to the simplex (Lemma T5.9, Cor. T5.10), and it is exactly orthogonal to the operating point in the a-weighted ambient form (Prop. T5.13); the Fisher metric of Prop. T5.12 is that form restricted to the tangent space, where the correction lives. The tangent map is therefore *driven by* the high-pass-filtered score perturbation, with the DC operating point left fixed (Prop. T5.15). We do not claim that J_a preserves the spectral ordering: it is not translation-invariant in the key index and need not carry a high-pass input spectrum to a high-pass output spectrum. Whether the shape survives is an empirical question, and §T6.6 answers

it by measurement. Prop. T5.3 bounds the score-level perturbation in terms of the embedding-level one. **(ii) The Hairer Survival Factor as a multiplicative attenuator.** The HSF factorises as $\mathrm{HSF}(\gamma,\tau) \geq c_0\gamma\tau^{-1}$, tracked empirically to $\pm 4\%$ across the γ sweep (Appendix A10), with $\mathrm{HSF} \propto \tau^{-1}$ verified across $\tau \in \{0.5, 1.0, 2.0\}$ (Appendix A10). The post-softmax Bode correlation of 0.94–0.97 measures this attenuation. **(iii) Where the linearisation is quantitative.** The remainder is second order in the score perturbation, so the regime in which the first-order picture is accurate is a measurable one. The pre-softmax score sup-norm lies inside that band on more than 96% of tokens (Appendix A10); a high-inverse-temperature ablation (τ=0.1) drives it outside, closes the HSF gap, degrades Bode r to ~ 0.41, and distorts the transition shape (Appendix A10). Section T6.11 reports the operating-point ratio directly: it is not small at the couplings where the trained transition sits at small d_k, and Theorem T5.22 assumes nothing about it.

T6.8 The embedding-axis low-pass induction filter

Figure 9 turns to the other spectral axis. A 2,000-experiment sweep over $\gamma \in [0, 3]$ and RoPE centre frequency $\theta \in [0.02, \pi]$ traces the recall-accuracy phase diagram in panel (a): induction efficacy concentrates in a high-fidelity zone at low θ and $\gamma \in [0.4, 1.4]$, consistent with momentum acting as a discrete high-pass filter on the score sequence, dual to a low-pass filter on the embedding axis. Pearson correlation between θ and the per-frequency momentum gain is $r = -0.679$ ($p = 9.9\times10^{-4}$, Cohen's $d = 1.05$, $n = 20$; Fig. 9b). The complementary 720-run granular sweep at 240 configurations (Appendix A14) maps band-by-band sensitivity, and the monochromatic control (Appendix A15, F3) rules out a frequency-hopping confound. This measurement also explains why the placement penalty of Corollary T4.4 is survivable in practice: the Coriolis term scales as $|\sin(\theta/2)|$ and is therefore largest in exactly the bands that carry the least induction signal.

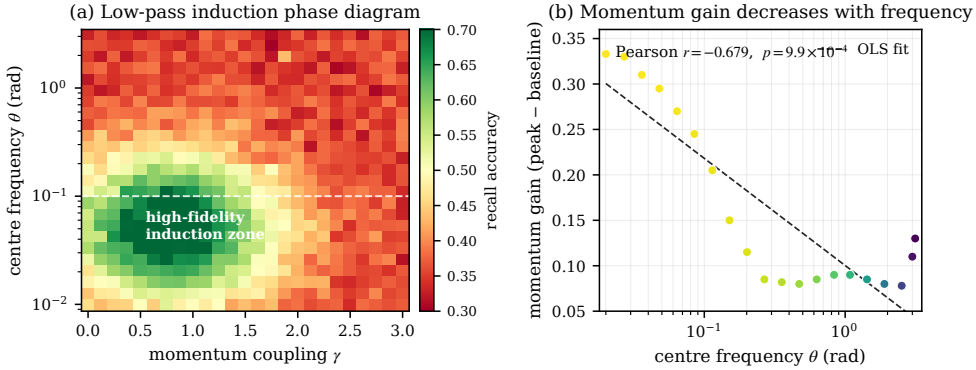


Figure 9: Embedding-axis low-pass induction filter (cascade RoPE → PSA). **(a)** Recall accuracy as a function of momentum coupling γ and RoPE centre frequency θ, on a 2,000-experiment sweep over $\gamma \in [0, 3]$ and $\theta \in [0.02, \pi]$ (per-configuration protocol and seeds in Appendix A16). The high-fidelity induction zone concentrates at low θ and $\gamma \in [0.4, 1.4]$. **(b)** Momentum gain at each θ (peak accuracy over $\gamma \in [0, 3]$ at that θ, minus the γ=0 baseline) against θ on the same sweep, $n = 20$ centre frequencies: a negative association, Pearson $r = -0.679$ ($p = 9.9 \times 10^{-4}$). Pearson correlation measures linear association and does not establish monotonicity; no rank statistic is reported.

T6.9 The two-branch law against Monte Carlo

Notebook: `PSA_MC_dk_sweep_with_results`.

Theorem T5.22 is a statement about $\mathcal{G}(d_k, T; \gamma)$, and this section tests it against direct simulation of $\mathcal{G}$ itself. The grid was fixed before the law was evaluated on it and no parameter is fitted on the d_k axis, so the test is a genuine held-out one *within the model*; what it does not test is whether $\mathcal{G}$ describes a trained transformer, which is the separate question §T6.11 addresses.

Protocol. The grid spans $d_k \in \{4, 8, 16, 32, 64, 128\}$ crossed with $T \in \{10, 15, 25\}$, giving eighteen points. At each point $\mathcal{G}(d_k, T; \gamma)$ is sampled directly, with five seeds and 60,000 trials per seed, on a 61-point γ grid covering $[0, 8]$. The critical coupling is extracted as the midpoint between chance $1/(T-1)$ and the measured plateau, read off the accuracy curve by linear interpolation between grid points; the resulting seed-to-seed spread is at most ± 0.003 and typically ± 0.001. A

smoke check at $(d_k, T) = (16, 15)$ against an independent four-way reference agrees to within grid resolution (0.773 against 0.776).

The mean-field branch. Figure 10 presents the comparison in three views: the law against the grid in panel (a), the held-out scatter in panel (b), and the error structure in panel (c). Nine of the eighteen points satisfy $d_k \geq 8L$ and therefore lie on the mean-field branch. On those nine, (T5.14) predicts the measured γ_c with Pearson $r = 0.9876$ and median relative error 2.6%, with zero fitted parameters (Fig. 10b). Seven of the nine fall within 5%; the remaining two are the pair closest to the crossover.

Behaviour near the crossover. The worst point is $(d_k, T) = (32, 25)$ at $8L/d_k = 0.79$, which the law overpredicts by 13.6%; the next worst is $(d_k, T) = (32, 15)$ at $8L/d_k = 0.66$, overpredicted by 7.5%. Both are the points nearest the crossover, and the direction of the error is the direction the structure of the law implies: as $d_k \downarrow d_k^*$ the mean-field expression rises steeply towards $\gamma_c^2 = 2L/\sqrt{d_k^*}$, whereas the measured γ_c passes smoothly into the condensed plateau. Figure 10a shows this directly, and shows the feature that makes it inevitable: *the two branches do not meet in value at* d_k^*. At $T = 15$, for instance, the mean-field expression evaluates to 1.072 at the crossover while the condensed branch sits at 0.794. The two branches are derived under different regimes — an annealed evaluation on one side, a condensed extreme-value argument on the other — and the piecewise law inherits a jump where they are stitched. The measured accuracy has no such jump. We state this as a limitation of the piecewise description rather than as a property of the system, and it is the reason the largest errors cluster where they do.

We resist the stronger claim that the error is monotone in proximity to the crossover. Across the nine points the least-squares trend of relative error against $8L/d_k$ is positive with Pearson $r = 0.688$ ($p = 0.041$), but the rank correlation is $\rho_s = 0.27$ ($p = 0.49$): at $n = 9$ the trend is not resolved as monotone, and Fig. 10c is drawn so that the reader can see the scatter as well as the trend.

Independence is measured, not assumed. The score model treats the candidate scores as independent. That is a modelling choice, and it is tested rather than argued: rebuilding the ensemble with the full dependence structure — shared query momentum across candidates, overlapping keys between adjacent positions, moments matched to the independent case — shifts the extracted γ_c by a median of -0.6%, with a maximum of 2.9% over the grid. The dependence is therefore present and quantified, and it is smaller than the median error of the law itself.

Fresh-seed confirmation. Certificate C8 re-derives the quartic root and compares it against Monte Carlo generated with seeds used in no derivation, at three deep mean-field points: $(d_k, T) = (64, 10)$ gives MC 0.570 against 0.545 (4.4%), $(64, 25)$ gives 0.668 against 0.669 (0.1%), and $(128, 15)$ gives 0.508 against 0.494 (2.8%). All three agree within 6%.

T6.10 The condensed branch and the argmax limit

Notebook: `PSA_T_axis_condensed_validation`.

The condensed branch is d_k-independent and varies only with context length, so it is tested along the orthogonal axis. Figure 11 collects the three tests: the T-axis scan in panel (a), the condensation constant on both axes in panel (b), and the zero-parameter argmax limit against direct simulation in panel (c).

The T axis. At fixed $d_k = 16$ we scan ten context lengths, $T \in \{6, 8, 9, 10, 12, 15, 20, 25, 30, 40\}$, with the same extraction procedure. Eight of the ten satisfy $d_k < 8L$ and lie on the condensed branch; on those eight the condensed form predicts the measured γ_c with Pearson $r = 0.9995$, median relative error 3.1% and maximum 5.2% (Fig. 11a). The two remaining points, $T = 6$ and $T = 8$, sit at $8L/d_k = 0.80$ and 0.97 — inside the near-crossover zone, on the mean-field side — and the mean-field expression overpredicts them by 19.0% and 36.0%. This is the same boundary-adjacent degradation the d_k grid records, at the other end of the same crossover, and it is shown rather than pooled: no aggregate statistic is quoted across the boundary, where the piecewise prediction is non-monotonic and a correlation coefficient would be meaningless.

The condensation constant. The constant c of (T5.14) is the one $O(1)$ quantity in the theory, and it is estimated on two independent axes. On the nine condensed points of the (d_k, T) grid the pointwise

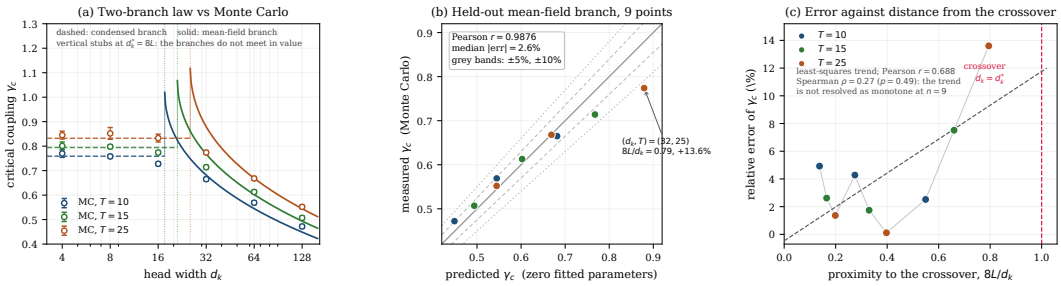


Figure 10: The two-branch law of Theorem T5.22 against the score-model Monte-Carlo grid. Protocol: eighteen (d_k, T) points, five seeds and 60,000 trials per point, γ_c extracted at the midpoint between chance and the measured plateau on a 61-point coupling grid; no parameter is fitted on the d_k axis. All three panels concern the reduced model $\mathcal{G}(d_k, T; \gamma)$ of Definition T5.20 and simulation of it, not trained transformers. *Observable.* Every curve from which a γ_c is extracted plots the competition observable $\mathcal{A}(\gamma) = \mathbb{E}[\mathrm{softmax}_j(s)_{j^*}]$, expected target softmax mass, and not argmax classification accuracy: that is the quantity Theorem T5.22 governs and the quantity that transitions. The argmax ordering within $\mathcal{G}$ is invariant in γ, exactly as the theorem states, so no argmax curve could show a transition in γ and none is plotted. The one place argmax appears is certificate C7, which checks the plateau p_∞ against direct argmax Monte Carlo; that is legitimate because p_∞ is the $\gamma \to \infty$ limit of $\mathcal{A}$, where softmax mass and argmax indicator probability coincide. Comparison with trained-model accuracy (§T6.11) is therefore a comparison between a mass and an accuracy, and is labelled approximate transfer wherever it appears. **(a)** γ_c against d_k at three context lengths. Dashed segments are the condensed branch, solid curves the mean-field branch, and the faint vertical stubs at $d_k^* = 8L$ mark the value gap between them: the piecewise law is discontinuous at the crossover while the measurement is not, which is what places the largest errors nearby. Error bars are the seed s.d. magnified $8\times$ to be visible. **(b)** Predicted against measured on the nine mean-field points, with $\pm 5\%$ and $\pm 10\%$ bands. Pearson $r = 0.9876$, median relative error 2.6%, zero fitted parameters. **(c)** Relative error against proximity to the crossover. The trend is positive but is not resolved as monotone at $n = 9$ (Spearman $\rho = 0.27$, $p = 0.49$).

estimates $c = \gamma_c\sqrt{2}/(2L)^{1/4}$ have mean 0.741 and sample standard deviation 0.015, a spread of about $\pm 2\%$. On the eight condensed points of the T-axis scan the mean is 0.717. The two agree to 3%, which is the accuracy at which a single constant describes both axes.

We record a residual structure rather than claiming exact constancy. Within the grid, c decreases weakly with d_k at fixed T ($0.752 \to 0.740 \to 0.711$ at $T = 10$) and increases weakly with T at fixed d_k; along the T-axis scan the pointwise estimate drifts from 0.680 at $T = 9$ to 0.755 at $T = 40$ (Fig. 11b). A single constant therefore captures the condensed branch to a few per cent and no better, and the drift is an honest residual of the finite-m freezing approximation in step (5) of the proof, not something the theory currently predicts.

The score-model argmax limit. Equation (T5.13) predicts the attainable accuracy of the reduced model with no free parameter at all. Certificate C7 evaluates the quadrature against direct argmax Monte Carlo at four points spanning two orders of magnitude in ceiling value. The certificate scores all $T - 1$ positions against an independently drawn target, so it is the $m = T - 1$ convention of Remark T5.30 that is in force here, applied identically on both sides of each comparison; Fig. 12 instead reports the $m = T - 2$ form, and the two differ by less than 0.021 at every point below. The pairs are: $(d_k, T) = (4, 10)$ gives 0.3409 against 0.3411; $(8, 25)$ gives 0.3193 against 0.3202; $(32, 15)$ gives 0.8379 against 0.8373; and $(128, 25)$ gives 0.9993 against 0.9993. The largest discrepancy is 9×10^{-4}, within the Monte-Carlo standard error at every point (Fig. 11c). Across all eighteen grid points the plateau matches simulation to three decimal places.

T6.11 The trained-model d_k ladder

Notebook: `PSA_trained_dk_overlay_with_results`.

The score model is an abstraction, and the question is what survives in trained transformers. We sweep $d_k \in \{4, 8, 16, 32, 64\}$ at $T-1 = 14$ pairs with $d_{\mathrm{model}} = 4d_k$ and $d_{\mathrm{ff}} = 16d_k$, three seeds and 3,000 steps per coupling on an eleven-point γ grid. At $T = 15$ the crossover sits at $d_k^* = 21.1$, so the ladder straddles it: $d_k \in \{4, 8, 16\}$ are predicted condensed and $d_k \in \{32, 64\}$ mean-field. Figure 12 reports the three quantities the comparison turns on — the measured $\gamma_c(d_k)$, the plateaus against the zero-parameter score-model argmax limit, and the operating point at which the trained models sit.

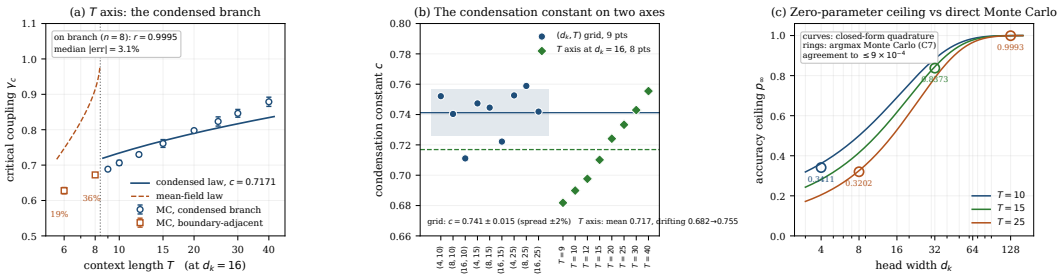


Figure 11: The condensed branch and the accuracy ceiling. **(a)** γ_c against context length at fixed d_k=16, ten lengths, same extraction procedure as Fig. 10. Circles are the eight condensed-branch points ($d_k < 8L$), squares the two boundary-adjacent points on the mean-field side with their percentage errors annotated. On branch: $r = 0.9995$, median error 3.1%. Error bars are the seed s.d. magnified $10\times$. **(b)** Pointwise estimates of the condensation constant on two orthogonal axes. The grid estimates (circles, mean and ±1 s.d. shaded) are flat to $\pm2\%$; the T-axis estimates (diamonds) drift systematically with T, which is the residual the single-constant description leaves. **(c)** The zero-parameter score-model argmax limit (T5.13) as a function of head width at three context lengths, with the four certificate C7 checkpoints against direct argmax Monte Carlo marked as rings. Largest discrepancy 9×10^{-4}.

Direction and ordering. The measured γ_c falls monotonically: no transition at $d_k = 4$, then 1.290 ± 0.044, 0.445 ± 0.028, 0.226 ± 0.004 and 0.120 ± 0.006 (Fig. 12a). Both the direction and the ordering are those the two-branch law gives, and the mechanism is the one Remark T5.24 describes: the per-candidate signal-to-noise ratio grows as $\sqrt{d_k}$, so wider heads need less sharpening.

Ceiling suppression at small width. The absence of a transition at $d_k = 4$ is not a missing mechanism but a suppressed ceiling. Equation (T5.13) gives $p_\infty = 0.282$ there, on the condensed branch, and the model never rises far enough above chance for a midpoint crossing to be defined. Across the ladder the measured plateaus (0.027, 0.285, 0.816, 0.953, 0.965) track the zero-parameter score-model argmax limit (0.282, 0.415, 0.617, 0.845, 0.978) with $r = 0.945$ (Fig. 12b). The tracking is in *shape*, not in level: at $d_k = 4$ the trained plateau is well below the ceiling and at $d_k = 16$ it is well above it. This quantity is the argmax limit of the score model, not a bound on trained accuracy — at $d_k = 16$ the trained plateau exceeds it by 0.199 — and we neither present the comparison as a match nor use the word ceiling for it.

An open exponent. A log–log fit of the four points with a measurable transition gives slope -1.13 (95% CI $[-1.18, -1.07]$, from ordinary least squares on the four $(\log d_k, \log \gamma_c)$ points, the interval taken from the regression standard error on two residual degrees of freedom — not a seed-level bootstrap and not a hierarchical estimate); restricted to the two mean-field points it is -0.91. The score model's asymptotic exponent is $-1/4$. Sign and ordering agree at every point; the magnitude does not, and we report the gap rather than absorbing it into a fitted amplitude. The natural reading is that the alignment amplitude relating the score model to a trained head is itself d_k-dependent, of order d_k^{-1}, which the abstraction does not represent. This is an open, measured discrepancy and it is listed as such in §T8.3.

The operating point. The ratio $\gamma\|p\|/\|q\|$ at $\gamma = 2$ measures 1.86, 3.05, 3.12, 3.03 and 2.92 along the ladder (Fig. 12c). It is reported as a measurement of where the trained models sit, and it is not small. Nothing in Theorem T5.22 requires it to be: the derivation is non-perturbative in γ, and §T5.6 introduces no expansion in this ratio. The measurement is nonetheless consequential for §T6.13, since the Differential Transformer linearisation *is* a small-γ statement, and the falling $\gamma_c(d_k)$ implies that at production widths the operative couplings move into the regime where the linearisation is quantitative: at $d_k = 64$ the trained transition sits at $\gamma_c\|p\|/\|q\| \approx 0.18$.

T6.12 Coupling, depth and length: a characterisation

Notebook: `PSA_stability_characterisation_results`.

Theorem T4.1 is an operator-level statement and licenses no claim about trained dynamics; §T4.2 explains why no such theorem is available for a map that does not iterate. What follows is measurement, pre-registered before it was run, and it is not renamed a theorem anywhere. Figure 13

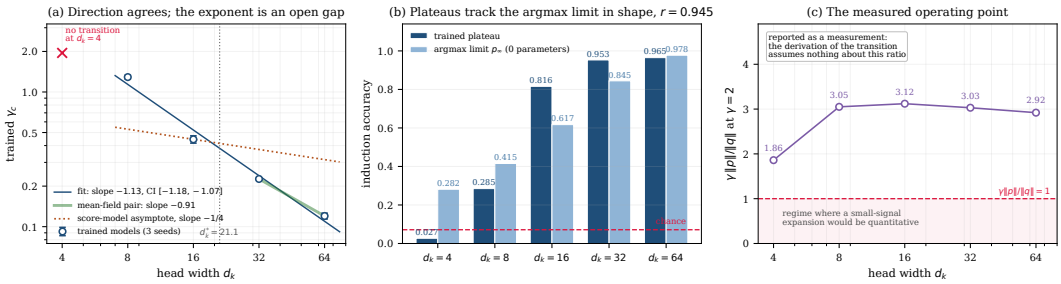


Figure 12: The trained-model d_k ladder. Protocol: $T-1 = 14$ induction pairs, $d_{\text{model}} = 4d_k$, $d_{\text{ff}} = 16d_k$, three seeds and 3,000 steps per coupling on an eleven-point γ grid, $T = 15$ so $d_k^* = 21.1$. **(a)** Trained γ_c against d_k on log–log axes. The fitted slope is -1.13; the two mean-field points alone give -0.91; the score model's asymptote is $-1/4$, anchored at $d_k = 16$ for comparison of slope. At $d_k = 4$ no transition is measurable. **(b)** Trained plateaus against the score-model argmax limit p_∞ of (T5.13), evaluated with $m = T - 2$. The two track in shape ($r = 0.945$) and not in level; p_∞ is the limit of the score model and is not a bound on trained accuracy, which exceeds it at $d_k = 16$. **(c)** The measured operating-point ratio $\gamma\|p\|/\|q\|$ at $\gamma = 2$. The shaded band is where a small-signal expansion would be quantitative; the trained models are outside it, and the transition theory does not require them to be inside it.

collects it: depth-wise score growth in panel (a), the inference-time coupling plateau in panel (b), and extended-length evaluation in panel (c).

Divergence. The tested set is the symmetric variant at all six couplings $\gamma \in \{0, 1, 2, 3, 5, 8\}$ and both depths $\{1, 3\}$ (twelve configurations), together with the asymmetric $(1, 3)$ and key-only variants at their own coupling and both depths (four more): sixteen configurations, two seeds each, and not the full Cartesian product of the three factors. Across them no run diverged, including at $\gamma = 8$ at depth 3. Final accuracies are stable across the coupling range: the symmetric variant gives 0.697, 0.793, 0.808, 0.833, 0.818 at $\gamma = 1, 2, 3, 5, 8$ at depth 1, and 0.677, 0.766, 0.797, 0.822, 0.820 at depth 3, against 0.035 and 0.033 for the γ=0 controls.

Depth-wise score growth. At depth 3 the pre-softmax score RMS by layer is $0.49 \to 2.85 \to 2.48$ for the γ=0 control and $2.45 \to 2.03 \to 2.01$ for the γ=2 run, with $2.62 \to 2.02 \to 2.12$ at γ=3 (Fig. 13a). Attention entropies follow the same pattern: $2.39 \to 1.64 \to 1.74$ for the control against $1.42 \to 1.42 \to 1.49$ at γ=2. A pre-registered mechanical check for growth below $3\times$ the layer-1 value was therefore failed — by the control, whose layer-2 RMS is $5.8\times$ its layer-1 value, and not by the sheared runs, which are flat through depth. The check is reported as measured, and the reading we draw from it is deliberately weak: the shear raises the layer-1 score scale, as the high-pass gain requires, and does not compound that increase through depth under layer normalisation.

Coupling perturbation at inference. Sweeping the evaluation coupling of a model trained at $\gamma = 2$, with weights fixed, gives a finite plateau: accuracy stays within 90% of its best over $\gamma \in [1.5, 4]$, falling to 0.504 at $\gamma = 1$ and 0.697 at $\gamma = 6$ (Fig. 13b). This is the output-level counterpart of the finite width of the transition, and it says that the coupling is not a knife-edge hyperparameter.

Extended length. Evaluating a model trained at 14 pairs on 24 and 31 pairs gives 0.626 and 0.511 against chance 0.042 and 0.032, that is $15\times$ and $16\times$ chance, with 0.768 at the training length (Fig. 13c). The maximum absolute score is stable across the three lengths (25.6, 24.1, 24.1), so the degradation is not a score-scale blow-up. Two confounds are acknowledged: RoPE extrapolation beyond the training length, and a vocabulary budget that caps the extension at about $2.2\times$.

One metric excluded. A loss-spike count was collected and is excluded from this characterisation. Its threshold is defined relative to the running median of the loss, which converges as training proceeds, so the metric measures convergence as much as instability; the raw counts range from 0 to 1109 across configurations with no interpretable ordering. We report that it was collected and discarded rather than silently dropping it.

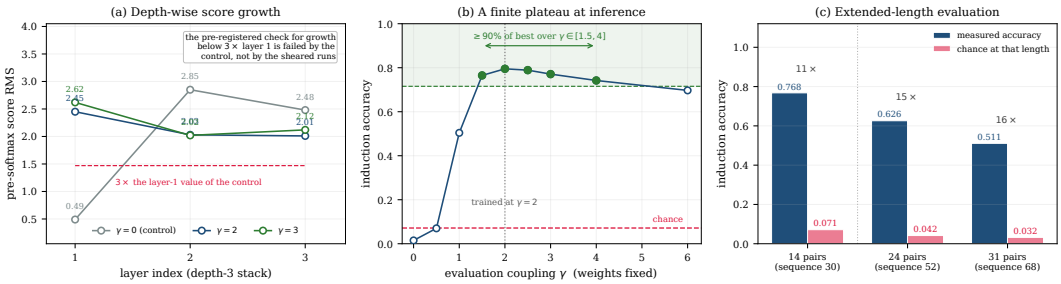


Figure 13: Coupling, depth and length characterisation. This is measurement and not theorem: see §T4.2. Protocol: d_k=16, 14 induction pairs, two seeds per configuration, chance 0.071. **(a)** Pre-softmax score RMS by layer in the depth-3 stack. The pre-registered check for growth below 3× the layer-1 value (dashed line, drawn at 3× the control's layer-1 value) is failed by the γ=0 control, not by the sheared runs. **(b)** Evaluation-coupling sweep for a model trained at γ=2 with weights fixed. Filled markers lie within 90% of the best accuracy; the plateau spans $\gamma \in [1.5, 4]$. **(c)** Accuracy at extended context length against the chance level at that length, for a model trained at 14 pairs. Ratios to chance are annotated. RoPE extrapolation is a confound and the vocabulary budget caps the extension at about 2.2×.

T6.13 Post-softmax first-order response

The Fréchet linearisation applied to the four-term PSA score gives the first-order attention response

$$a_t^{\mathrm{PSA}} = a_t + \gamma\, J_{a_t}\left[\frac{p_{q,t}^\top k_\cdot + q_t^\top p_{k,\cdot}}{\sqrt{d_k}}\right] + \mathcal{O}(\gamma^2),$$

which is (T8.1), proved as Proposition T8.3. The post-softmax tangent response is therefore *driven by* the pre-softmax high-pass perturbation, about the unshifted operating point; how much of the spectral emphasis survives the state-dependent projection J_{a_t} is a matter for measurement rather than proof, and §T6.7 reports it. Its relation to Differential Transformer [29], which forms a difference of two separately normalised softmax maps, is set out in Remark T8.4: the two agree only under a further smallness hypothesis on the inter-query score difference, which the coupling does not supply, and we claim no architectural identity.

T6.14 Which circuit executes induction in a deep PSA model?

Enabling the shear at every layer of a deep transformer makes two routes to induction available at once. The single-layer PSA channel exposes the predecessor directly through the ghost key, so one attention layer suffices. The classical two-layer composition remains available as well: a previous-token head in one layer, an induction head reading its output in the next. Which route a trained model actually builds is not settled by either mechanism being available, and accuracy does not reveal it, because both routes solve the task. This subsection asks the question directly, with a probe for where the answer becomes decodable and an ablation for where it is computed.

Protocol. Nano transformers ($d_{\mathrm{model}} = 128$, 8 heads, $d_k = 16$, vocabulary 128, sequence length 128) at eight depths $N \in \{2, 4, 5, 6, 7, 8, 12, 16\}$ and eight couplings $\gamma \in \{0, 0.25, 0.5, 0.75, 0.9, 1.25, 1.75, 2.5\}$, three seeds each: 192 models, 3,000 steps apiece, 326 minutes on one A100. The task is randomised-offset induction. A motif of length $p \sim \mathcal{U}\{24, \ldots, 48\}$ is placed at a random offset and then repeated, so the target is recoverable by matching the current token to its earlier occurrence and copying that token's successor, and *not* by any fixed displacement. The distinction is measured rather than assumed: no single fixed lag solves the task, the best reaching 0.065 — far below any trained model here, though some eight times the uniform-vocabulary rate of 0.0078, so the claim is that fixed lags do not suffice and not that they sit at chance. Copying at the generator's own period scores 1.000; that is an oracle supplied the hidden period rather than a content matcher, and because motif tokens are drawn with replacement it does not establish a ceiling for content matching. Under a verbatim repeat the same measurement returns 1.000 for the *fixed-lag* rule — the task is then solvable outright by a relative positional rule — which is why the period is randomised here.

Two diagnostics are taken on every model. A ridge readout from each layer's residual stream to the next token, fitted on one set of activations, penalised by a λ chosen on a second, and scored on a third:

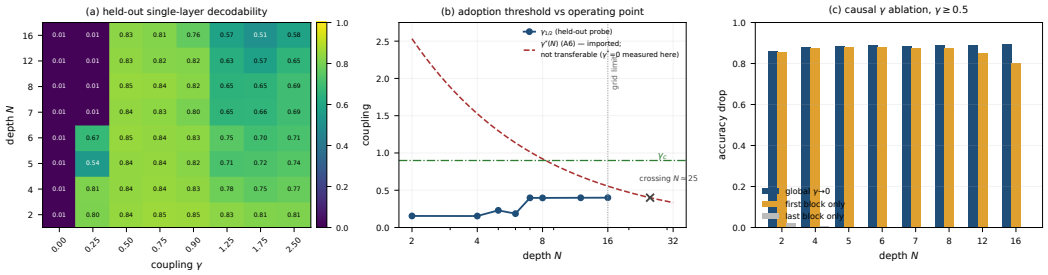


Figure 14: Where induction becomes decodable, and where it is computed. **(a)** Held-out ridge-probe accuracy for the next token read from the residual stream after a *single* attention layer, over depth and coupling; 192 models, three seeds. The $\gamma = 0$ column sits at chance (0.0078) throughout. **(b)** The coupling $\gamma_{1/2}$ at which the model adopts the single-layer route, against the operating-point curve $\gamma^*(N)$ of Appendix A6 — *imported, and measured here to be non-transferable: the in-configuration optimum is* $\gamma^* = 0$ *at every depth* — and the reduced model's γ_c, drawn as a reference level rather than a prediction under test. The two curves do not cross within the grid; extrapolating the fitted γ^* places the crossing near $N \approx 25$, against a registered prediction of $N \approx 8.3$. That extrapolation falsifies the pre-registration and carries no further mechanistic weight. **(c)** Accuracy lost to γ-ablation, averaged over $\gamma \geq 0.5$. Zeroing the shear in the first block alone costs almost exactly what zeroing it everywhere costs, and zeroing it in the last block costs nothing — except at $N \geq 12$, where the first-block and global bars separate.

it locates the depth at which the answer is linearly present. In-sample accuracy is retained alongside, so that the inflation is quantified rather than argued about: across the grid it averages $+0.024$, and at $\gamma = 0$ it turns a true 0.0086 into an apparent 0.0398 against a chance rate of 0.0078. A probe fitted and scored on one set of activations would report that floor as signal. And a per-layer γ ablation, which zeroes the shear in the first block alone, the last block alone, or every block, and measures what the accuracy loses: it locates the depth at which the answer is computed. The probe answers *where the information is*; only the ablation answers *where the work happens*, and reporting the first as though it settled the second is the error this design is built to avoid. Per-head censuses of previous-token attention, and of attention at the sequence's own induction offset, are taken as well. They are reported but are not load-bearing: with the period randomised they are noisier than the probe and the ablation, and a fixed-offset head can produce a high induction-offset statistic with no previous-token structure at all, so only the probe and the ablation separate the two routes. Appendix A34 gives the notebook.

The single-layer channel is real, and its first-block contribution is causally load-bearing. At $\gamma = 0$ the next token is not linearly decodable after one attention layer at any depth: the held-out probe reads 0.0086 against a chance rate of 0.0078. Switching the shear on changes this completely. For $\gamma \geq 0.5$ the same probe reads 0.76 on average, and 0.828 at $N = 16$ — the deepest model tested, with fifteen further layers it could have used instead. The channel that the SHT bound forbids to a single unaugmented layer is open at every depth tested, for every $\gamma \geq 0.5$ in the grid. It is *not* open at every non-zero coupling: at $\gamma = 0.25$ the probe is already at chance from $N = 7$ upward, which is the depth-dependent adoption threshold treated below.

The ablation shows the same thing causally, and more sharply. Averaged over $\gamma \geq 0.5$ and over depths up to $N = 8$, zeroing the shear in the *first block alone* costs 0.874 of accuracy; zeroing it in *every* block costs 0.882. The difference is 0.008. Zeroing it in the last block alone costs 0.006. Through $N = 8$, therefore, the first-block ablation reproduces essentially the entire global-ablation cost, and no later layer's shear contributes anything the ablation detects. That equivalence does not extend to the deepest models: at $N = 12$ and $N = 16$ the two ablations separate — 0.849 against 0.891, and 0.805 against 0.895, on the same average — and a measurable share of the dependence has moved downstream. The paragraph after next takes that regime up. Table 10 gives the three ablations at every depth rather than the averages quoted here.

The right noun for this is *causally load-bearing*, not circuit identification. Setting $\gamma_1 \to 0$ at inference alters the residual stream entering every downstream layer, and those layers were not trained on that input distribution; the measured drop therefore combines the removal of the first-block feature with a distribution shift it induces downstream. The held-out probe constrains the interpretation — the feature removed is one that demonstrably carries the answer — but does not eliminate the second channel. Activation patching across layers is the intervention that would, and it is not performed here.

Table 10: Accuracy lost to γ-ablation, per depth, averaged over $\gamma \geq 0.5$ and three seeds. Through $N = 8$ the first-block ablation reproduces the global one to within 0.011 and the last-block ablation costs nothing; from $N = 12$ the first and global rows separate, which is the distributed corner discussed below. Plotted as Fig. 14(c).

accuracy drop	N=2	4	5	6	7	8	12	16
global $\gamma \to 0$	0.861	0.880	0.885	0.888	0.888	0.889	0.891	0.895
first block only	0.856	0.874	0.878	0.879	0.878	0.878	0.849	0.805
last block only	0.021	0.008	0.003	0.002	0.001	0.000	0.000	0.000
first − global	−0.005	−0.006	−0.007	−0.009	−0.010	−0.011	−0.042	−0.090

What the evidence supports is that the first-block shear is load-bearing, at every depth tested, and that is already more than a correlational probe could establish.

The notebook additionally runs a three-way route classifier, and we set its output aside rather than report it as a finding. It labels 153 of the 192 cells single-layer PSA, 38 composition-like and one inconclusive, but the labelling depends on four stipulated thresholds — probe above 0.5, ablation drop above 0.10, previous-token statistic above 0.30 — so the frequency is not a scale-invariant quantity and moves with the thresholds. The continuous quantities above carry the argument instead. The 38 composition-like cells are essentially the $\gamma = 0$ column together with $\gamma = 0.25$ at $N \geq 7$, where no shear is in use and the probe sits at chance; we call them composition-*like* because the absence of shear dependence alongside a previous-token head is consistent with the two-layer route without establishing it, which would require patching the candidate induction head.

Where the computation does spread. One corner departs from that picture, and it is worth naming rather than averaging away. At $N \geq 12$ and $\gamma \geq 1.75$ the first-block and global ablations separate: at $N = 16$, $\gamma = 2.5$ the global ablation costs 0.921 while the first-block ablation costs 0.656, a gap of 0.265. Deep models at strong coupling move part of the computation out of layer one — not into the classical two-layer composition, whose previous-token signature does not appear, but into a shear-dependent computation distributed across layers. That is a third route, neither of the two this subsection set out to distinguish, and we report it as such.

Depth raises the adoption threshold, but not by much. Define $\gamma_{1/2}$ as the coupling at which single-layer decodability first reaches one half. The depth dependence is not a smooth law but two regimes: $\gamma_{1/2} \approx 0.15$–$0.23$ for $N \leq 6$, and ≈ 0.40 for $N \geq 7$, with nothing in between. The whole of the step sits in the $\gamma = 0.25$ column, where the probe reads 0.80 at $N = 2$ and $N = 4$, falls to 0.54 and 0.67 at $N = 5$ and $N = 6$, and collapses to chance from $N = 7$ onward. Given a weak shear, a shallow model takes the single-layer route and a deep one ignores it. Given $\gamma \geq 0.5$, every depth takes it. The boundary itself is where the seeds disagree: at $N = 5$ the three seeds give $0.40, 0.15, 0.15$ and at $N = 6$ they give $0.15, 0.32, 0.15$, while every seed at $N \geq 7$ returns 0.40. Because the transition in γ is sharper than the grid spacing, $\gamma_{1/2}$ resolves only to the interval containing it: the values at $N \geq 7$ mean "between 0.25 and 0.5", and we fit nothing smooth through eight bracketed intervals.

Two predictions fail, and one of them matters. Before any model was trained we registered a crossing at $N \approx 8.3$, where the depth law $\gamma^*(N) = 4.2N^{-0.73}$ of Appendix A6 falls through $\gamma_c = 0.899$. That figure is a working midpoint: at $d_k = 16$ the scored positions run over $T_{\text{eff}} \in [25, 127]$, which places γ_c on the condensed branch throughout the band, in $[0.832, 0.924]$. Nothing below turns on the choice — no grid point lies between 0.878 and 0.899, so the above/below partition of the coupling grid is identical under either endpoint. No crossing occurs. $\gamma_{1/2}$ saturates near 0.40 and γ^* does not fall that far until $N \approx 25$, well outside the grid. The registered prediction is wrong by a factor of three, and the qualitative claim it encoded — that deep models lose access to the single-layer route — is false: at $N = 16$ the route is available, adopted, and causally load-bearing.

The second failure is the more informative, and it needs its terms stated carefully, because $\gamma_{1/2}$ and γ_c are not the same observable. γ_c is the coupling at which target softmax mass undergoes its transition inside the T_4-only Gaussian reduction $\mathcal{G}$; $\gamma_{1/2}$ is the coupling at which a trained residual representation becomes linearly decodable on a held-out probe. The second is not an estimator of the first, and the horizontal line in Figure 14(b) marks a reference level rather than a prediction under test. With that said, the ratio is informative about the size of the idealisation: $\gamma_{1/2}$ lies between $0.17\gamma_c$ and $0.45\gamma_c$, so the representational onset arrives at a small fraction of the coupling the reduction

associates with the transition. The same over-prediction appears independently in Appendix A4, where measured characteristic couplings of 0.225 and 0.275 sit against a predicted $\gamma_c = 0.737$, a ratio of 0.31 and 0.37. Two experiments with different tasks, depths and positional encodings agree that $\mathcal{G}$ over-predicts the coupling a trained model needs, by a factor of between two and six. That is a quantified, reproducible gap between the reduced model and trained behaviour, and it is the clearest statement this programme can currently make about the size of the $\mathcal{G}$ idealisation.

An honest null: the route works and is not worth taking here. On this task the best operating point is $\gamma = 0$ at every depth. Models without the shear reach 0.99 accuracy in 500 steps; models with $\gamma \geq 0.5$ plateau between 0.88 and 0.94. The single-layer channel is available, is adopted, and is causally load-bearing for most of the shear-supported computation over the moderate-coupling regime — and those models are worse than the ones that compose two ordinary layers instead. Randomised-offset induction at $N \geq 2$ is a task classical composition already solves easily, so the single-layer route buys nothing and the shear's cost is paid for no benefit. The operating-point curve of Appendix A6 does not transfer to this configuration: measured here, γ^* is zero throughout, which is why Figure 14(b) plots the imported curve and labels it as imported. This null must be read against §A31.7, which runs the *same task family* at a different point in configuration space and reaches the opposite conclusion: at $d_k = 128$, $d_{\text{model}} = 768$, $N = 12$ and 91.3M parameters, $\gamma = 0.80$ gives $3/3$ emergence in a mean of 717 optimiser steps against $2/3$ and 2700 at $\gamma = 0$. The two results are not in conflict once the configurations are stated, and the difference is instructive rather than embarrassing. This grid runs at $d_k = 16$, vocabulary 128 and sequence length 128, where the reduced model's threshold is at the top of the condensed branch; the 91.3M sweep runs at $d_k = 128$, vocabulary 8192 and length 512, where γ_c is lowest of any width measured. The honest summary is that the coupling helps where the task is hard enough and the head wide enough that the baseline struggles — the 91.3M baseline needs 2700 steps and fails one seed in three — and costs accuracy where the baseline already saturates in 500 steps, as it does here. Neither result generalises to the other's regime, and the paper claims neither direction as universal.

Descriptive statistics, and what is not claimed of them. Pooled over the 192 cells the held-out first-layer probe correlates with γ/γ_c at $r = +0.425$. Mean decodability above γ_c exceeds that below it by a factor of 1.44, 1.42 and 1.39 in the three seeds (mean 1.42, s.d. 0.03), and computed within each depth separately the ratio runs from 1.25 to 1.67, so the contrast is not an artefact of how the two groups are composed across depths. No p-value is attached to any of this. The grid cells are hyperparameter settings sharing a single data generator rather than independent draws, and the three seeds are the replicates; a test computed over the cells would be counting the grid, not the evidence.

What this establishes. Three things, at the level the design supports, and they separate three notions that are routinely conflated. *Representability*: the shear makes a single-layer induction channel expressible, which is the theory's claim. *Learned availability*: a trained model actually builds that channel, at every depth up to sixteen, and depends on it causally — established by intervention on a task with no single fixed-lag positional shortcut, not by a correlational probe on a task that was solvable by one outright. *Optimisation preference*: on this task the optimiser would rather not use it, and does better without it. The first is a theorem, the second is what this experiment adds, and the third is a property of the task, not of the construction.

Two further readings follow. The representational onset arrives at a small fraction of the reduction's γ_c, and an independent measurement in Appendix A4 finds the same direction and magnitude. And the depth law of Appendix A6 does not transfer to this task, so the crossing argument built on it fails; what governs adoption here is coupling, with depth entering only through the step between $N = 6$ and $N = 7$ at $\gamma = 0.25$.

The practical reading is unchanged from §1 but now rests on intervention rather than inference. PSA closes the filter-order gap at a single layer, and a trained model can — and above the measured adoption threshold does — make first-block use of that route. Whether it *should* depends on whether two ordinary layers were available and cheap, and on this task they were. The regime where the construction earns its cost is the one where they are not.

T6.15 Machine-checkable certificates

Notebook: `PSA_Certificates`.

Ten certificates accompany the paper (Table 11). Each is a few lines of exact symbolic algebra (SymPy ring normal forms) or float64 numerics; the set requires no GPU and no training data, and runs end to end in under one minute on any CPU. All ten pass, in two independent environments. They check ten selected identities and numerical comparisons, conditional on the paper's definitions and modelling assumptions: C7 and C8 are Monte-Carlo agreements rather than proofs, and no certificate validates an omitted term, a probability hypothesis, or the scope of a theorem. Their purpose is that the algebra a reader is most likely to want to check can be checked faster than it can be read.

Table 11: The certificate set. Column three states what is verified, column four the outcome as printed by the notebook.

	Supports	Verifies	Outcome
C1	Thm. T5.22(2)	Gaussian MGF by direct integration; annealed exponent $\Lambda = L + 2\gamma^4$	exact, symbolic
C2	Thm. T5.22(4)	$\max_\gamma(a - \Lambda) = d_k/8 - L$ at $\gamma^2 = \sqrt{d_k}/4$; crossover $d_k^* = 8L$	exact, symbolic
C3	Thm. T5.22(3)	quartic root closed form, $(L^2/d_k)^{1/4}$ asymptote, monotonicity in both arguments	exact, symbolic
C4	Prop. T3.27(i)–(ii)	asymmetric shear is symplectic and preserves H_{joint}	residual exactly 0.0 at d_k=64
C5	Prop. T3.27(iii)	$s^{\text{asym}} - s^{\text{sym}} = \varepsilon(q^\top p_k + \gamma_q p_q^\top p_k)$	exact, symbolic
C6	Thm. T4.1(i)	$M_\gamma^\top J M_\gamma = J$ symbolically and numerically	residual 0.0 exactly at d_k=64
C7	eq. (T5.13)	plateau quadrature against direct argmax Monte Carlo, four (d_k, T) points	agrees within 3σ; max $\Delta = 9{\times}10^{-4}$
C8	eq. (T5.14)	quartic root against fresh-seed Monte Carlo, three deep mean-field points	within 4.4% at all three
C9	Prop. T3.25	transport identity exact; asymmetric current is no total difference	exact, symbolic
C10	§T6.6	Pearson r invariant under positive rescaling of its first argument	exact, symbolic

All forty-one notebooks — the thirty-four of the main programme and the seven of the battery in Table 5 — ship as supplementary material: thirty-seven with outputs already embedded, and the four Appendix Q sweeps with the result files, ledgers and logs described in §T6.2 in place of embedded output. Every figure and every number above can therefore be inspected without re-execution; independent reproduction requires running the notebooks. The certificate notebook is the one to start from: it depends on nothing else, needs no accelerator and no data, and checks its ten identities in under a minute.

T6.16 Additional independently verified predictions

Beyond Table 6: the L=30 chained-ICL stress test reports a -52.5% repeated-token-loss regression ($r = 0.987$, Appendix A29); a 2,880-run multi-difficulty sweep locates the sweet spot ($F = 12.04$, $p = 7.4 \times 10^{-4}$, Appendix A23); and the $\nabla/\int$ negative control ($t = -2.91$, $p = 4.5 \times 10^{-3}$, Appendix A22) shows the dissociation expected if derivative-sensitive score channels — including the T_4 coherence channel retained by $\mathcal{G}$ — are operational. Aggregate across both tranches: ∼9,377 training runs, forty-one notebooks and 36 appendix sections; §T6.2 and Table 7 are the authoritative inventory (Appendix A35).

T6.17 What the programme does not show

Three things, stated here rather than left to inference. The evidence is synthetic throughout: every number above comes from the canonical induction sequence model or from small decoders trained on it. The battery establishes that differencing on the key axis is the operative ingredient, and it establishes that the symmetric member is *not* the accuracy-optimal member of its class. And the trained ladder confirms the direction of $\gamma_c(d_k)$ while leaving its exponent unexplained by a factor of four. Section T8.3 collects these with the rest.

T7 Deployability: what the operator costs an inference engine

T7.1 Who this section is for

The obstruction this paper circumvents is a statement about *one* attention layer. At twelve layers it is of no practical consequence: the depth is there, the two-layer induction circuit forms, and nothing is missing. The obstruction is the whole problem exactly where the stack is one or two layers deep — and that is the regime an on-device assistant occupies, because latency and memory budgets push toward very shallow stacks while retrieval-from-context is precisely the capability such models are asked for. A shallow model that cannot do in-context match-and-copy is a shallow model that cannot use the user's own data.

The shear supplies the missing cross-position reference architecturally, rather than by spending a layer on it. Whether that is *deployable* turns on questions the theory of §§T1–T5 does not answer, and which an inference engineer asks before reading any accuracy table:

1. Does the key-side shear break the KV cache?
2. What does it cost per token?
3. Does it invalidate a cached prefix?
4. Does it widen the activation range enough to break INT8?

This section settles all four by direct derivation. The answers are: the KV cache is *exactly* the same size; the additional persistent state is two vectors per layer, independent of context length; the arithmetic overhead is $6d$ FLOPs per token per layer, below one part in five thousand in every regime we tabulate; prefix reuse is untouched; and the quantisation dynamic range grows by at most $1 + 2\gamma$, that is at most $\log_2(1 + 2\gamma)$ additional bits, and by exactly nothing at DC. The operator adds no parameters, and because the shear is a pre-pass on q and k with v untouched, a fused attention kernel is used unmodified.

Everything in this section is exact arithmetic, not measurement. We state it here rather than in an appendix because the usual reasons an architectural change is rejected *before* it is evaluated — cache growth, kernel incompatibility, broken prefix reuse, quantisation blow-up — are the reasons a deployment team never gets to the accuracy question at all. None of what follows argues that the operator *should* be deployed. It establishes only that those four objections do not apply, so that the decision is made on accuracy in the shallow regime the obstruction concerns, where §T6 reports our evidence, including where it is negative.

T7.2 Baseline, and the apparent obstacle

Let d be the model width, h the number of heads, $d_k = d/h$ the head dimension, T the context length and L the number of layers. Per layer, $q_t = W_Q^\top r_t$, $k_t = W_K^\top r_t$, $v_t = W_V^\top r_t$, reshaped into h heads of width d_k. Standard incremental decoding keeps

$$\mathcal{C}_K = \{k_1, \dots, k_{t-1}\}, \qquad \mathcal{C}_V = \{v_1, \dots, v_{t-1}\}, \tag{T7.1}$$

so the cache holds $2(t-1)d$ scalars per layer. PSA replaces q and k by their DC-neutral first differences and leaves v alone,

$$\hat{q}_t = (1+\gamma)q_t - \gamma q_{t-1}, \qquad \hat{k}_j = (1+\gamma)k_j - \gamma k_{j-1}, \qquad q_{-1} = k_{-1} := 0, \tag{T7.2}$$

with transfer function $H(z) = (1+\gamma) - \gamma z^{-1}$ and

$$\left|H(e^{i\omega})\right|^2 \;=\; 1 + 4\gamma(1+\gamma)\sin^2(\omega/2), \tag{T7.3}$$

so that $|H(1)| = 1$ exactly and $|H(-1)| = 1 + 2\gamma$. Both facts are used below.

The apparent obstacle is this. At decode step t the head must score $\hat{q}_t$ against $\hat{k}_j$ for every $j \le t$, and each $\hat{k}_j$ depends on k_{j-1}. A natural reading concludes that the engine must retain the raw keys in order to form the sheared ones, and must re-form all t of them at every step: $O(td)$ extra work per token, and a second cache. Both conclusions are wrong, and the reason is the same property that makes the operator causal in the first place.

Algorithm 4 PSA incremental decode, one layer, step t

state: $\widehat{\mathcal{C}}_K, \mathcal{C}_V$, and registers $q_{\text{prev}}, k_{\text{prev}}$ (initialised to 0)
$q_t, k_t, v_t \leftarrow W_Q^\top r_t,\ W_K^\top r_t,\ W_V^\top r_t$
$\hat{q}_t \leftarrow (1+\gamma)\, q_t - \gamma\, q_{\text{prev}}$ *# 3d FLOPs*
$\hat{k}_t \leftarrow (1+\gamma)\, k_t - \gamma\, k_{\text{prev}}$ *# 3d FLOPs*
$q_{\text{prev}} \leftarrow q_t;\quad k_{\text{prev}} \leftarrow k_t$
append $\hat{k}_t$ to $\widehat{\mathcal{C}}_K$; append v_t to $\mathcal{C}_V$
$o_t \leftarrow \text{Attention}(\hat{q}_t, \widehat{\mathcal{C}}_K, \mathcal{C}_V)$ *# unchanged kernel*

T7.3 The KV cache is unchanged

Proposition T7.1 (Cache invariance)**.** *Let the engine cache $\widehat{\mathcal{C}}_K = \{\hat{k}_1, \dots, \hat{k}_{t-1}\}$ in place of $\mathcal{C}_K$, and additionally retain the two raw vectors $q_{t-1}, k_{t-1} \in \mathbb{R}^d$. Then incremental decoding under (T7.2) is exact, the cache has the same size as (T7.1), and step t requires no recomputation of any earlier entry.*

Proof. Equation (T7.2) is a causal two-tap FIR filter: $\hat{k}_j$ depends on the stream through k_j and k_{j-1} and on nothing later. It can therefore be applied once, at insertion time.

At step t the engine holds $\hat{k}_1, \dots, \hat{k}_{t-1}$ together with q_{t-1} and k_{t-1}. It projects r_t to obtain q_t, k_t, v_t and forms $\hat{q}_t$ and $\hat{k}_t$ by (T7.2), both of which use only the two retained vectors. It appends $\hat{k}_t$ to $\widehat{\mathcal{C}}_K$ and v_t to $\mathcal{C}_V$, then overwrites $q_{t-1} \leftarrow q_t$ and $k_{t-1} \leftarrow k_t$. The attention pass is $s_{t,j} = \hat{q}_t^\top \hat{k}_j / \sqrt{d_k}$ for $j \le t$, which reads $\widehat{\mathcal{C}}_K$ exactly as standard decoding reads $\mathcal{C}_K$. No entry $\hat{k}_j$ with $j < t$ is touched, because $\hat{k}_j$ does not depend on anything after j. The stored objects are t sheared keys and t values — the same count as the baseline — plus two vectors that do not grow with t. □

Remark T7.2 (Why the sheared cache and not the raw one). One could instead cache raw k and shear at read time. That costs $3td$ FLOPs per step rather than $3d$ — it re-derives the whole history at every token — and is strictly worse. The only reason to prefer it would be a γ that changes between steps, which no deployment scenario requires: γ is fixed at training time.

The additional persistent state is two d-vectors per layer, $M_{\text{extra}} = 2dLb$ bytes at b bytes per scalar, against a baseline cache of $M_{\text{cache}} = 2TdLb$. The ratio $M_{\text{extra}}/M_{\text{cache}} = 1/T$ is independent of width and depth. At $T = 4096$ that is 0.024%; concretely, $d = 4096$, $L = 32$, $T = 4096$ in fp16 gives a 2.1 GB cache against 524 kB of registers.

Corollary T7.3. *PSA leaves the KV-cache footprint unchanged and adds $O(dL)$ state, independent of context length. The memory argument against deploying it does not exist.*

T7.4 Arithmetic overhead

Each of $\hat{q}_t, \hat{k}_t$ is one AXPBY over d elements: two multiplies and one add per element, so $3d$ FLOPs each and $6d$ per token per layer. Against the standard decode cost per token per layer — QKV projections $3 \cdot 2d^2$, attention scores and weighted sum $2Td + 2Td$, output projection $2d^2$, and an MLP at expansion 4 costing $16d^2$, totalling $24d^2 + 4Td$ — the overhead fraction is

$$\frac{6d}{24d^2 + 4Td} = \frac{6}{24d + 4T}. \tag{T7.4}$$

In every regime of Table 12 the arithmetic cost is below one part in five thousand. The operation is moreover memory-bound rather than compute-bound — a two-operand elementwise pass over vectors already resident in registers — so on hardware where decode is bandwidth-limited the marginal wall-clock cost is smaller still.

During prefill the whole sequence is present, so (T7.2) is a single vectorised shifted subtract over the $T \times d$ activation block: no recurrence, no serialisation, $6Td$ FLOPs per layer against a prefill cost of $24Td^2 + 4T^2d$. The same fraction (T7.4) applies.

Table 12: Arithmetic overhead of the shear, from (T7.4). The fraction *falls* as context grows, because the attention term dominates.

configuration	d	T	overhead
small on-device	1024	2048	1.8×10^{-4}
mid on-device	2048	4096	9.2×10^{-5}
server class	4096	4096	5.2×10^{-5}
long context	4096	32768	2.6×10^{-5}

T7.5 Prefix reuse and speculative decoding

Proposition T7.4 (Prefix reuse is preserved)**.** *Let $P = (x_1, \ldots, x_m)$ be a prefix and let two requests continue it differently. The sheared keys $\hat{k}_1, \ldots, \hat{k}_m$ are identical in both, so a cached prefix may be reused verbatim.*

Proof. $\hat{k}_j$ depends only on k_j and k_{j-1}, both determined by $x_1, \ldots, x_j$ with $j \le m$. Neither depends on any token after position m. □

This is causality and finite impulse response together: a two-tap causal filter cannot propagate information backwards, so no cached entry is invalidated by a continuation. An infinite-impulse-response variant — an EMA with $\beta > 0$ — would share the property; a non-causal or centred difference would not, and would break prefix caching outright. That is a concrete engineering reason to prefer the $\beta = 0$ causal first difference over a symmetric stencil, independent of the representational argument of §T1.

One convention needs care. Taking $k_{-1} = 0$ makes $\hat{k}_0 = (1 + \gamma)k_0$. An engine that splices a cached prefix onto a fresh sequence must apply the same convention at the same absolute position, or the first entry will disagree. This is one line of care, not a design problem.

For speculative decoding, a draft model proposes n tokens which the target verifies in parallel. Verification supplies k_{j-1} for every proposed j within the same block, so (T7.2) applies directly. On rejection at position j the engine rolls the cache back to $j - 1$ and restores $q_{\mathrm{prev}}, k_{\mathrm{prev}}$ from the retained raw vectors at that position; the rollback therefore needs the raw k at the rollback point, one extra d-vector per speculative block, not per token.

T7.6 Quantisation headroom

Whether an INT8 KV cache survives the operator is decided by how much the shear widens the dynamic range of the stored keys.

Proposition T7.5 (Bit growth)**.** *For any key stream,*

$$\|\hat{k}_j\|_\infty \;\le\; (1 + 2\gamma) \max_{i \le j} \|k_i\|_\infty, \tag{T7.5}$$

so the number of additional bits required to represent the sheared keys at fixed relative precision is at most $\Delta b = \log_2(1 + 2\gamma)$. The bound is tight, attained by the alternating stream $k_j = (-1)^j \kappa$.

Proof. $\hat{k}_j = (1+\gamma)k_j - \gamma k_{j-1}$, so by the triangle inequality $\|\hat{k}_j\|_\infty \le (1+\gamma)\|k_j\|_\infty + \gamma\|k_{j-1}\|_\infty \le (1 + 2\gamma) \max_i \|k_i\|_\infty$. For the alternating stream, $\hat{k}_j = (1 + \gamma)(-1)^j\kappa - \gamma(-1)^{j-1}\kappa = (-1)^j(1 + 2\gamma)\kappa$, attaining it. Representing a range scaled by $1 + 2\gamma$ at fixed relative precision requires $\log_2(1 + 2\gamma)$ further bits. □

Three observations follow, and the first two are why the bound is benign in practice. First, it is worst-case over a Nyquist-alternating stream: by (T7.3) the gain at DC is exactly 1, so a constant key stream passes unchanged, and realised growth is governed by how much high-frequency content the key stream actually carries — strictly below $1 + 2\gamma$ for any stream that is not fully alternating. Second, the growth is bounded and known before deployment, because γ is fixed at training time: the scale factor of an INT8 cache can be widened by $1 + 2\gamma$ once, statically, and there is no dynamic-range surprise at run time, which is the usual way an operator breaks a quantised pipeline. Third,

Table 13: Worst-case quantisation cost of the shear, from Proposition T7.5. The last column is the number of INT8 levels left after widening the scale by $1 + 2\gamma$.

γ	range factor $1 + 2\gamma$	Δb (bits)	INT8 levels retained
0.05	1.10	0.14	233/256
0.10	1.20	0.26	213/256
0.20	1.40	0.49	183/256
0.40	1.80	0.85	142/256

per-channel scaling absorbs it: standard per-channel or per-head KV quantisation already computes a scale from observed maxima, those maxima rise by at most $1 + 2\gamma$ under Proposition T7.5, and the calibration procedure needs no change.

Remark T7.6 (What is not claimed). Proposition T7.5 bounds the representable range, not the end-task accuracy of a quantised model. Whether an INT8 PSA cache matches an INT8 baseline in downstream quality is an empirical question and is not settled here.

T7.7 Summary

Table 14: The deployment bill, relative to standard attention. Every row is exact arithmetic, not measurement.

quantity	PSA relative to standard attention
parameters	identical: the operator adds none
KV-cache size	identical; cache the sheared keys (Prop. T7.1)
extra persistent state	$2dL$ scalars, independent of T ($1/T$ of the cache)
FLOPs per token, per layer	$+6d$, i.e. $6/(24d + 4T)$; $< 10^{-4}$ in practice
prefill	same fraction; fully vectorised, no recurrence
prefix caching	unaffected (causal two-tap FIR, Prop. T7.4)
speculative decoding	compatible; one extra vector per block for rollback
quantisation range	$\times(1 + 2\gamma)$ worst case, $\times 1$ at DC
additional bits	$\leq \log_2(1 + 2\gamma)$: 0.49 at $\gamma = 0.2$, 0.85 at 0.4
attention kernel	unmodified; the shear is a pre-pass on q, k

The last row of Table 14 is the one that matters most operationally. Because the shear is applied before the attention call and V is untouched, a fused attention kernel — FlashAttention or a vendor equivalent — is used unmodified. The operator is a two-line change at the call site, not a kernel rewrite. For a team evaluating whether one-layer in-context retrieval is reachable on a phone, a wearable or an embedded controller, that is the difference between an experiment that costs an afternoon and one that costs a quarter.

T8 Related work, limitations, and discussion

T8.1 Related work

Positional encodings. The construction is stated on the rotary substrate [27], and the placement corollary is specific to it; additive and bias-based schemes such as ALiBi [28] induce a different momentum geometry, which we have not analysed. The frequency-band reading of §T6.8 is closest in spirit to recent work on where RoPE concentrates spectral energy [51].

Sharp transitions in training. Abrupt capability onsets as a function of a continuous control are reported in several settings, notably grokking [46] and the emergence of copy heads under a Bayesian analysis [6]. The transition here is of a different kind — a function of an architectural coupling at fixed training budget, not of training time or data scale — and we draw no equivalence.

One-layer obstructions. Sanford et al. [16] give the induction-heads lower bound this paper takes as its starting point, in the line of Sanford et al. [17], Edelman et al. [18], Liu et al. [19] on the representational limits of single-layer attention. The mechanistic account of the two-layer circuit is due to Elhage et al. [21], Olsson et al. [20], with subsequent characterisations via gradient-descent

simulation [36, 35, 37, 34] and causal analysis [22, 23]; the abruptness of its formation during training is modelled in closed form by Reddy [24].

Convolutional and lagged augmentations of attention. The observation that a short causal convolution on the key or query stream unlocks single-layer associative recall is not new, and two lines establish it. Li et al. [4] augment attention with learned convolutional filters on the K, Q and V streams and prove that the resulting layer solves associative recall and copying with a single layer, with guaranteed length generalisation; Wang et al. [5] recast sequence layers as test-time regression over input tokens, a framework in which a single-layer solution to multi-query associative recall is constructed from a short key convolution. KV-shifting attention [30] supplies the lagged key channel directly and reports improved induction behaviour; Proposition T8.1 places it outside the admissible class as a pure delay, and §T8.2 sets out what that distinction does and does not amount to. These works establish *sufficiency* empirically, across a family of learnable filters.

An earlier preprint by one of the authors. Maitra [52] introduces the same first-difference augmentation of the post-RoPE query and key streams, under the name *momentum attention*, and studies it through spectral forensics and an account of in-context learning. It is a partial predecessor, and the overlap between the two works is in the measurements rather than in the theory: they share the operator, part of the spectral phenomenology, and part of the empirical corpus — a number of the accompanying notebooks were written for that paper first, and still carry its terminology, so a reader tracing a result back to its notebook should expect the earlier name. What this paper adds is the account of *why* the operator works: the filter-order reading of the SHT obstruction, the correspondence with Hairer's lift, the uniqueness theorem of §T3.5, the placement corollary of §T4.3, the closed-form transition of §T5 and the deployability derivation of §T7. Where the two overlap, the statements here are the ones that stand. The contribution here is complementary and of a different type: we identify the minimal fixed member of that family — the DC-neutral first difference — characterise it as the unique swap-symmetric member of its class (Theorem T3.18), and derive the transition theory of its T_4-only Gaussian reduction in closed form (Theorem T5.22). The control battery of §T6 closes the loop from the other side: a general learned two-tap filter, given the whole family and free coefficients, converges to a pure key-side first difference ($b/a = -1.040$ on the key stream, -0.010 on the query stream). Gradient descent selects, from within their family, the member this analysis identifies.

Structure-preserving architectures. Hamiltonian and symplectic neural networks [38, 39, 43, 41, 42], and neural ODEs [40], impose geometric structure on learned dynamics. The present construction differs in placement and in ambition: the operator is fixed rather than learned, it acts on the pre-softmax score streams rather than on a state trajectory, and it is applied once rather than integrated. Closest in setting is Gahtan and Bronstein [49], who couple the Q and K streams through an integrator and report perplexity gains, with symplectic and non-symplectic integrators performing alike once the coupling is present. We read that parity as consistent with, and independent corroboration of, Proposition T3.27 and the battery of §T6: within a class in which the structural ingredient is present, integrator class does not discriminate task accuracy. Their construction lies outside $\mathcal{C}_{\mathrm{PSA}}$ and Theorem T3.18 says nothing about it; their reported crossover, at 350 M, of coupled dynamics by Differential Attention — the linearised form — is the behaviour that the falling $\gamma_c(d_k)$ of Theorem T5.22 would lead one to expect for this family at scale, and we note the agreement without claiming it as a test. Independently, and from a different direction again, Stein et al. [50] lift tokens to a phase space carrying both feature and velocity variables and obtain Hamiltonian momentum attention blocks by discretising inertial Nesterov-type dynamics on density manifolds, in the mean-field regime in which attention blocks are gradient flows of interaction energy functionals under Wasserstein-2-type metrics. Their object is the convergence rate of the depth-wise flow and their blocks act on the residual stream through depth; ours is a fixed operator acting within one layer on the post-RoPE Q and K streams, and the quantity of interest is filter order rather than acceleration. Neither result bears on the other's claims. We record the convergence because arriving at a phase-space lift from optimal transport, without any reference to induction or to a depth bound, is evidence that the lift is natural to the setting rather than an artefact of our route to it.

Attention as a random energy model. The apparatus of §T5.6 is not ours and should not read as though it were. Giorlandino and Goldt [47] map the attention scores of a transformer at initialisation onto Derrida's random energy model and, by a replica calculation, locate a transition at inverse

temperature $\beta_n \sim \sqrt{\log n}$ in the context length n — the same $\sqrt{2L}$ freezing scale that governs step (4) of Theorem T5.22, reached independently and first. Chen et al. [48] obtain instead a critical scaling $\beta_n \asymp \log n$, and attribute the discrepancy not to an error but to differing assumptions about the score field itself. The critical scaling of the softmax inverse temperature is therefore model-dependent, and which law applies is an open question in that literature rather than settled ground. The distinction matters here, because the field of §T5.6 is a third one: (M3) takes the distractor scores *independent*, where Chen et al. [48] record that Giorlandino and Goldt [47] take them correlated. What the three share is not a model but Derrida's freezing scale.

Our result does not adjudicate it, and it is worth being exact about why. Both of those works vary the inverse temperature, or the context length that sets it, in a *fixed* architecture, and ask when the softmax changes regime. Here β is held at its usual $1/\sqrt{d_k}$ and the quantity swept is γ, a coupling inside the operator, with the law expressed in (d_k, T) rather than in β alone. The freezing scale is common because the underlying object is the same Gaussian field of e^L competitors; what differs is which knob moves. Neither (T5.14) nor its crossover $d_k^* = 8L$ follows from a β-scaling result, and neither of those results follows from ours. What we take from that literature is the licence to use the annealed bound and its saturation as a self-delimiting criterion, which is Derrida's, and the knowledge that in attention specifically the identification has already been made and tested by others.

Attention variants with a difference structure. Differential Transformer [29] subtracts two softmax maps and is recovered here as the first-order Fréchet linearisation of a single symplectic operator (§T6, eq. (T8.1)); the identification supplies a derivation for a construction otherwise introduced on empirical grounds. Rotary position embedding [27] is the substrate on which the placement corollary operates, and Xiong et al. [51] analyse where RoPE concentrates structured energy across frequency bands, which is the reading we use for the embedding-axis measurement of §T6. Linear and state-space alternatives [31, 32] escape the obstruction by leaving the class entirely rather than by raising its order. Closest in spirit to §T5, though on an entirely different axis, is Lavie et al. [6]: a Bayesian theory of feature learning in attention which derives a closed-form posterior over the attention matrix of a single-layer softmax network trained on a copy task, reduces it to a low-dimensional order-parameter space, and locates a first-order transition in the emergence of the copy subcircuit — second-order, followed by a crossover, for linear attention. Their order parameter is the learned attention pattern and their control variable is training; ours is a fixed architectural coupling and the transition is in representability at convergence. The two axes are independent and neither subsumes the other, but the appearance of a sharp transition with a low-dimensional order parameter in both settings is a coincidence worth recording.

T8.2 Three neighbouring constructions, compared exactly

Three families of proposal sit close enough to this construction that the relationship is worth settling algebraically rather than by description. Each of the three results below is a statement about where the neighbour sits relative to the conditions of Theorem T3.18, and one of them is a derivation of the neighbour rather than a separation from it.

KV-shifting is an all-pass relabelling. KV-shifting [30] supplies the key at position $j - \delta$ in place of the key at j, which in transfer-function terms is $H^{\mathrm{KV}}(z) = z^{-\delta}$.

Proposition T8.1 (KV-shifting fails the high-pass condition)**.** $|H^{\mathrm{KV}}(e^{j\omega})| = 1$ *for every* ω*, so* $|H^{\mathrm{KV}}(e^{j0})| = |H^{\mathrm{KV}}(e^{j\pi})|$ *and the strict inequality in* (C2) *of Theorem T3.18 fails. KV-shifting is therefore not an admissible member of the class, for any* δ*.*

Proof. $|z^{-\delta}| = |z|^{-\delta} = 1$ on the unit circle. □

Remark T8.2 (What the distinction amounts to). The point is not that KV-shifting fails to help — it does help, and is reported to — but that it helps by a different mechanism. A pure delay hands the layer the predecessor *instead of* the current key; it relabels the position index and forms no mixture. The augmentation here produces $(1 + \gamma)k_j - \gamma k_{j-1}$, a genuine combination of current and predecessor content with a non-trivial frequency response, which is what gives Proposition T3.4 something to say. Both cross the order-$(0, 1)$ threshold that Proposition T1.1 identifies; only one of them does so with a DC-neutral high-pass filter, and only one of them therefore admits the transition theory of §T5. The control battery separates them empirically: the low-pass two-tap, which is the

nearest matched-parameter member without the high-pass property, reaches 0.428 against 0.811 (§T6.5).

Differential Transformer is the first-order linearisation. The recovery stated as (T8.1) in §T6.13 is proved here.

Proposition T8.3 (First-order response of the augmented softmax)**.** *Let* $a_t = \mathrm{softmax}_j\big(q_t^\top k_j/\sqrt{d_k}\big)$ *and* $J_{a_t} = \mathrm{diag}(a_t) - a_t a_t^\top$*. The PSA attention weights satisfy*

$$a_t^{\mathrm{PSA}} \;=\; a_t \;+\; \gamma\, J_{a_t}\left[\frac{p_{q,t}^\top k_\cdot + q_t^\top p_{k,\cdot}}{\sqrt{d_k}}\right] \;+\; \mathcal{O}(\gamma^2), \tag{T8.1}$$

a correction tangent to the simplex (Cor. T5.10) and high-pass in character (Prop. T5.15).

Proof. By Theorem T3.2, divided by $\sqrt{d_k}$, the augmented score is $\hat{s}_t = s_t + \gamma\delta_t + \gamma^2\tau_t$ with $\delta_t = (p_{q,t}^\top k_\cdot + q_t^\top p_{k,\cdot})/\sqrt{d_k}$ and $\tau_t = p_{q,t}^\top p_{k,\cdot}/\sqrt{d_k}$. Apply Proposition T5.2 at base point s_t with perturbation $\gamma\delta_t + \gamma^2\tau_t$: the linear response is $J_{a_t}(\gamma\delta_t + \gamma^2\tau_t)$ and the remainder is quadratic in the perturbation, hence $\mathcal{O}(\gamma^2)$. Collecting orders in γ gives (T8.1). Tangency follows from Corollary T5.10 applied to $J_{a_t}\delta_t$. □

Remark T8.4 (Relation to Differential Transformer). Retaining the query-side half of δ_t gives $\gamma J_{a_t}\big[(q_t - q_{t-1})^\top k_\cdot/\sqrt{d_k}\big]$: the Jacobian acting on a difference of two score vectors. Differential Transformer [29] instead forms the difference of two *separately normalised* softmax maps, $\mathrm{softmax}(s_t) - \lambda\mathrm{softmax}(s_{t-1})$. These coincide only under a further hypothesis, and it is worth stating exactly which:

$$J_{a_t}\big(s_t - s_{t-1}\big) = \mathrm{softmax}(s_t) - \mathrm{softmax}(s_{t-1}) + \mathcal{O}\big(\|s_t - s_{t-1}\|^2\big), \tag{T8.2}$$

so the difference-of-softmax form additionally requires $\|s_t - s_{t-1}\|$ to be small. Smallness of γ does not supply it: s_t and s_{t-1} are score vectors at two distinct query positions and differ at order one whatever the coupling. Two further discrepancies are worth recording. The expression $\mathrm{softmax}(s_t) - \lambda\mathrm{softmax}(s_{t-1})$ has total mass $1 - \lambda$ rather than 1, so it is not itself a normalised attention distribution; and Differential Transformer uses two independently parameterised attention maps with a learned λ, where (T8.1) has one parameterisation and one scalar. We therefore claim no architectural identity. What (T8.1) does establish is that the augmentation acts, at first order, as a filtered tangent correction of the same difference-of-attention character that Differential Transformer implements directly.

Structure-preserving networks operate on a different object. Hamiltonian and Lagrangian neural networks, symplectic networks and neural ODEs [38, 39, 43, 41, 42, 40] learn a vector field and integrate it, imposing geometric structure on a *trajectory*. Three differences are structural rather than incidental. The operator here is fixed, not learned, and contributes one scalar. It acts on the pre-softmax score streams within a single layer, not on a state evolving through depth or time. And it is applied once, so the question those architectures are built to answer — whether the learned flow remains well behaved over long integration — does not arise, which is the same observation §T4.2 makes from the other direction. The overlap is the operator; the setting is not shared.

T8.3 Limitations

Statistical power. The empirical claims rest on two or three seeds per configuration, and the reported comparisons use seed standard deviations rather than inferential statistics. No confidence intervals are constructed for differences between variants, no correction is applied for the number of couplings scanned, and the grid maxima of §T6.5 are best-of-k statistics rather than held-out estimates. The separations we rely on are large — the 13.8 pp key-only versus symmetric gap, the $+67.5$ pp phase jump — and we do not lean on any marginal contrast, but a reader wanting interval estimates or a formal test of the ordering will not find them here, and producing them means more seeds rather than more analysis.

The evidence is overwhelmingly synthetic. Every result in §T6 is obtained on the canonical induction sequence model or on small decoders trained on it, at 4–92 M parameters. The task is a designed probe, not a language benchmark.

No gain at language-model scale is claimed. A $\sim$300 M parameter-efficiency study is a near-miss and is reported as such (Appendix A32). Practical benefit at realistic scale remains unestablished. The one concrete prediction the theory offers here is testable rather than promissory: γ_c falls with head width, so at production widths the operative couplings are small and the mechanism approaches its Differential Transformer linearisation. That is a hypothesis for follow-up, stated as a hypothesis.

The two headline scales are separate claims. The single-layer representability result (54 k, Fig. 6a) and the scaling-shape validation (842 k, multi-layer, Fig. 6b) are measured on deliberately different tasks and are not blended: they are presented separately in the abstract, in the figure caption, and in the text.

The trained exponent does not match the abstraction. The trained d_k ladder gives a log–log exponent of -1.13 (95% CI $[-1.18, -1.07]$) against the score model's asymptotic $-1/4$. Direction, ordering, crossover phenomenology and ceiling suppression agree; the magnitude does not. The gap is open and is reported as measured, not absorbed into a fitted amplitude.

Stability is evidence, not a theorem. §T4.2 explains why no long-time or depth-wise stability theorem is available for a map that does not iterate, and we do not offer one. The characterisation in §T6 — zero divergence to $\gamma = 8$ at depths 1 and 3, depth-flat score RMS, a finite inference-time coupling plateau — is measurement, is labelled Empirical, and is not a substitute.

The obstruction is sidestepped, not escaped. Nothing here contradicts Sanford et al. [16]. The augmented head forms a score outside the class their counting argument enumerates, while its own bit budget stays $\mathcal{O}(d_k)$; the premise is exited and the bound stands.

Uniqueness is class-relative and symmetry is a commitment. Theorem T3.18 constrains the factorised subclass of $\mathcal{C}_{\mathrm{PSA}}$ and nothing beyond it, and its symmetry condition (C3.2) is a design choice rather than a consequence of symplecticity (Prop. T3.27). On the induction task alone the key-only member is the better engineering choice (0.949 against 0.811); the symmetric member is chosen for the structural package, and the reader who wants only single-layer induction should take the half-lift.

Nulls. Multi-hop with negations shows no effect ($\Delta = -0.4 \pm 1.2$ pp, Appendix A19). The depth law fails at $L = 30$ (Appendix A7). Two pre-registered battery predictions did not hold and are reported as measured in §T6. A loss-spike metric was collected and excluded for a threshold-convergence confound.

T8.4 Discussion and outlook

Section T7 settles the deployment question by exact derivation rather than by measurement: the causal two-tap shear streams during incremental decoding, with sheared keys replacing raw keys in the cache at no change in cache size, two additional persistent vectors per layer, and an arithmetic overhead of $6d$ FLOPs per token per layer. Because the transformation is a pre-pass on Q and K with V untouched, the downstream fused attention kernel is used unmodified, and prefix reuse and speculative decoding are preserved. The operator adds no learnable parameters beyond the scalar γ — held fixed in the experiments reported here, optimisable without disturbing Theorem T4.1 (Rem. T4.3). What §T7 does *not* establish, and what is therefore not claimed anywhere, is wall-clock latency on real hardware, downstream accuracy under quantisation, and the compatibility of every standard normaliser.

Three directions follow from the structure rather than from the measurements. First, extensibility: the symmetric Verlet-class member is a self-adjoint base, to which the standard symmetric-composition schemes — the kick–drift–kick palindrome of (T2.6) and the Yoshida/Suzuki tower [12, 11] — apply directly. We claim no exclusivity: a non-self-adjoint method can be composed with its adjoint to build a symmetric higher-order scheme, which is how Verlet itself is assembled, so asymmetric and half-lift bases are not barred from higher-order completion. What symmetry buys is that the standard theory applies without that intermediate step. Whether any useful higher-order *score* construction exists on the attention side is open; none is defined or tested here. Second, the value stream is deliberately untouched here (§T3.7), and whether a phase-space reading of V is meaningful is open. Third, the

non-linear extension: Theorem T3.18 is silent outside the bilinear one-step class, and the constraint that would replace symplectic consistency for a non-linear shear — a Jacobian condition rather than a matrix identity — is known exactly (Prop. T3.21); what remains open is how to parameterise and train a useful non-linear g satisfying it.

The wider reading we would defend is narrow. Geometric numerical integration supplies attention with an operator, a reason that operator is canonical, and a reason its symmetric member is distinguished. It does not supply a stability theory, and we have been careful to take only what transfers. What the correspondence buys, in the end, is a closed-form answer to a question that is otherwise answered by sweeping: given a head of width d_k and a context of length T, at what coupling does the task become solvable, and how well can it be solved. Within the reduced Gaussian score model, the mean-field coupling (T5.14) and the argmax limit (T5.13) are zero-fit expressions, and both were tested against simulation that played no part in deriving them. The condensed branch carries one calibrated constant, the two branches do not meet in value at the crossover, and transfer to trained models remains approximate (§T6.11, Rem. T5.5).

T9 Addendum: the first-order channel in the reduced model

The reduced model $\mathcal{G}$ of §T5.6 retains the T_4 block alone. This addendum asks what changes if the first-order γT_3 channel is retained alongside it. The question is not decorative: Proposition T5.4 already establishes that T_3 carries a discriminative gap of its own, and the two channels enter the score at different orders, so which one dominates depends on where γ_c falls. Nothing below revises Theorem T5.22, which is a statement about the T_4-only reduction and is correct as such. What follows is a statement about a different reduction, and the two make different predictions.

Where the omitted channel dominates. T_3 contributes $\gamma\,\sigma^2 d_k$ to the score gap and T_4 contributes $\gamma^2\sigma^2 d_k$, so the first-order channel is the larger of the two for every $\gamma < 1$. The smaller root of the mean-field quadratic crosses $\gamma_c = 1$ when $2 - \sqrt{d_k} + L = 0$, that is at $d_k = (L+2)^2 \approx 67.8$. On the interval $8L \approx 49.9 \le d_k < 67.8$ the predicted coupling therefore exceeds one and T_4 is the larger channel; for every wider head $\gamma_c < 1$ and the discarded channel is the dominant one. Below the crossover that root does not exist and the governing value is the condensed one, $\gamma_c = 0.9849 < 1$, so T_3 dominates there as well — which sharpens the point rather than softening it. The band on which the omission costs least is therefore the single interval $8L \le d_k < (L+2)^2$, and the tabulation below straddles it on both sides. Even there the reduced model is discarding a real discriminative term, so the omission is never free; it is least consequential near the crossover, and asymptotically it is not defensible at all. That is a reason to compute the alternative, not a reason to believe it.

Moments. Applying the recipe of step (1) of Theorem T5.22 to both channels, and normalising by the same $\sqrt{d_k}$: the target mean gains the first-order term, and the per-distractor variance gains the variance of $\gamma\, q_t^\top p_{k,j}$, which is $2\gamma^2\sigma^4 d_k$ against the $4\gamma^4\sigma^4 d_k$ of the T_4 channel. The two are *not* independent, and the cross term is of the same order: since $p_{q,t} = q_t - q_{t-1}$ with $q_t \perp q_{t-1}$, we have $\mathrm{Cov}(q_t, p_{q,t}) = \sigma^2 I$ and hence

$$\mathrm{Cov}\big(q_t^\top p_{k,j},\; p_{q,t}^\top p_{k,j}\big) \;=\; 2\sigma^4 d_k.$$

Retaining it,

$$a \;=\; \gamma(1+\gamma)\sqrt{d_k}, \qquad b^2 \;=\; 2\gamma^2 + 4\gamma^3 + 4\gamma^4. \tag{T9.1}$$

Dropping every T_3-dependent term — the linear term in a, and the $2\gamma^2 + 4\gamma^3$ contributions to b^2 — recovers (T5.11) exactly, as it must.

The crossing condition, and the loss of a closed form. Substituting (T9.1) into the annealed criterion (T5.15) gives

$$2\gamma_c^4 \;+\; 2\gamma_c^3 \;-\; (\sqrt{d_k} - 1)\,\gamma_c^2 \;-\; \sqrt{d_k}\,\gamma_c \;+\; L \;=\; 0. \tag{T9.2}$$

The odd term is the whole difference. Equation (T5.16) is a quadratic in γ^2 and therefore admits the closed-form root that produces the two-branch law; (T9.2) is a genuine quartic in γ and has no comparable closed form. Its first positive root is the analogue of γ_c.

Consequences. *First, the predicted couplings fall, by an amount that grows with the head width.* At $L = \log 511$:

d_k	32	48	64	96	128
T_4 only, (T5.14)	0.9849	0.9849	1.0300	0.8671	0.7867
T_3 and T_4, (T9.2)	0.9515	0.6637	0.5687	0.4696	0.4132

At $d_k = 32$ the two agree to within 3.4%; by $d_k = 128$ the extended model predicts a coupling almost half the size. The direction is the one the measurements have been asking for. Every comparison in this paper between γ_c and a trained model has found the reduced model over-predicting: §T6.14's half-maxima sit between $0.17\gamma_c$ and $0.45\gamma_c$, and the two positional-encoding families of Appendix A4 measure 0.225 and 0.275 against a predicted 0.737. Retaining T_3 moves the prediction toward the measurement without any fitted quantity.

Second, the asymptotic exponent changes. For large d_k the first-order term dominates the balance in (T9.2), giving $\gamma_c \simeq L/\sqrt{d_k}$, against the $(L^2/d_k)^{1/4}$ scaling of Theorem T5.22(2). Fitted over

$d_k \in [256, 65536]$ the exponent is -0.254 for the T_4-only law and -0.461 for (T9.2). Trained models give -1.13. So the first-order channel closes roughly a quarter of the exponent gap and leaves the rest open; it is a step toward the discrepancy of §T8, not a resolution of it.

Third, the crossover moves. The condensed branch of Theorem T5.22 exists because below d_k^* the T_4 signal cannot overtake the annealed aggregate at any coupling. The first-order channel grows faster in γ than the aggregate does at small γ, so a crossing exists for smaller d_k: the first positive root of (T9.2) appears at a saddle-node near $d_k \approx 31.5$, against $d_k^* = 8L \approx 49.9$. Any width sweep beginning at $d_k = 32$ therefore lies entirely on a single branch under the extended model, and the predicted shape in d_k is monotone — the flat segment and the $+4.6\%$ jump at $8L$ do not appear at all.

What this does and does not license. The status of (T9.2) is Conditional in the sense used throughout this paper: it inherits the Gaussian surrogate and the annealed partition function, and adds nothing else: the T_3–T_4 covariance is retained exactly rather than assumed away. It is not a correction to Theorem T5.22, which is a theorem about $\mathcal{G}$ as defined; it is the corresponding statement for a less reduced model. Neither is a statement about a trained transformer.

What the addendum does establish is where the reduction bites. The γ_c values of (T5.14) are robust to including T_3 only up to a factor of about two; the $(L^2/d_k)^{1/4}$ asymptote is not robust at all; and the two-branch structure — the crossover at $8L$ and the jump that follows from it — is the single most reduction-sensitive feature of the theory. That is worth stating plainly, because the jump is also the feature that no single-branch power law can produce, and therefore the one on which a measurement is most naturally read as a test of the law. Under the less reduced model it is not predicted at all.

That makes the width sweep a discriminating measurement rather than a confirmatory one. A measured jump at d_k^* favours the T_4-only reduction and, with it, the mechanism in which the coherence channel is what the transition is about. A monotone fall favours (T9.2) and locates the transition in the first-order channel instead. A null in which nothing emerges at any width would speak to neither. We record the prediction without the measurement, and deliberately so. The width sweep of §A31.6 was stopped at $\gamma \leq 0.40$ because that range is sub-critical at every width it covers, and a grid that never crosses γ_c cannot discriminate the two reductions however many seeds it accumulates. Settling the question therefore requires a sweep that has not been run: $d_k \in \{48, 64\}$ carried to $\gamma \approx 1.2$ at the coupling-scaled step size of §A31.11, those two widths being where the reductions separate most cleanly — one predicting a rise of 4.6% and the other a fall of 14.3%.

Part A: The empirical programme in full

A1 From theorems to measurements: an overview

These appendices are the empirical record of the Phase Space Attention (PSA) programme. Each section anchors one or more formal claims of the body and reports the measurements that test them, preserving the tables, figures and protocol details of the underlying notebooks.

Two notations used throughout the appendices. Falsifiable predictions are numbered FP-n: each is a statement, fixed before the corresponding run, that a specific measurement would either confirm or refute, and the numbering is the order in which the predictions were registered rather than the order in which they appear here. Where an FP is discussed, its content is restated in full at that point, so no separate index is needed. The certificates C1–C10 are the machine-checkable identities of Table 11 in the body; the letter–number codes C6 and above that appear inside the validation-suite sections A9 and A10 are those suites' own internal check numbers, and are unrelated to the certificates; they are written v8 C6, v9 C7 and so on wherever the enclosing section does not already fix which suite is meant. A third and separate use of the same letter is the design-class conditions C1–C3 of Theorem T3.18 — causality, high-pass, and symplectic consistency — which are called *conditions* throughout and are never certificates.

Empirical claims are tagged [L3 – Empirical] in keeping with the three levels set out in §1; they corroborate, but do not replace, the theorems they test.

Empirical operating regime (restated from §T6.11). The experiments in these appendices span model sizes from approximately $54{,}000$ parameters (didactic microbenchmarks) through 4.45 M (in-context-learning arc), 91.7 M (low-coupling conditioning γ-sweep), and up to ~ 300 M (parameter-efficiency near-misses). This range is the regime we judge practically motivated: the description cost implied by the SHT bound is a larger fraction of a compact model's budget than of a frontier model's, so an additional induction circuit matters more there. We claim no universal threshold, and we do not equate SHT's description-size variables with parameter count, latency or deployability (§1). The programme is calibrated to this range because it is where the closed-form predictions of the body both apply and admit fine-grained measurement of the predicted structure.

Tranche structure. The empirical program is organized into tranches. Tranche 1 (Sections A1–A11) is the present material. It has three logical units:

1. **Structural verification** (Section A2). A NumPy-level reconstruction of the post-RoPE PSA pipeline that verifies the computational order, the norm-preserving action of RoPE on Q and K, and the four-term decomposition of the score (Theorem T3.2). The Coriolis ordering of Corollary T4.4 is implicit in every result that follows: the momentum is computed on *post*-RoPE vectors.
2. **Operating regime** (Sections A3–A4). Two training-based studies that locate PSA's operating regime along the two axes of the design class $\mathcal{C}_{\mathrm{PSA}}$. The β-sweep (Section A3) tests the high-pass condition (C2) of the Uniqueness Theorem T3.18: at $\beta = 0$ the momentum is the backward difference $\nabla q_t = q_t - q_{t-1}$ that emerges from the proof, while at $\beta \to 1$ the EMA collapses the high-pass signal and accuracy returns to the vanilla baseline. The γ-sweep (Section A4) locates the critical coupling γ_c predicted by Theorem T5.22 and contrasts RoPE with sinusoidal positional encoding.
3. **Single-layer induction** (Sections A5–A10). Six notebooks comprising the central empirical claim of the program: post-RoPE PSA closes the SHT/Olsson–Olah single-layer induction gap and does so with a pre-softmax Bode correlation r up to 0.9994 to the theoretical transfer function. The results in this unit map onto Theorem T5.7, Theorem T5.22, Proposition T3.4, and Corollary T4.4, with explicit scope qualifications where the empirical regime deviates from the theorem hypotheses.

Architecture, task, and reproducibility. Three nested architectural settings are used. The *didactic* setting of Section A2 is a 32-token NumPy implementation with $d_{\mathrm{model}} = 64$, $d_k = 32$, no training,

single seed; its purpose is to verify structural claims, not empirical performance. The *small-scale training* setting of Sections A3–A4 is a four-layer decoder of ∼842k parameters with $d_{\text{model}} = 128$, four heads, head dimension 32, RoPE base 10 000, trained on associative recall with chains of length $L \in \{4, \ldots, 20\}$ for the β-sweep and a harder L=12, vocabulary 200 task for the γ-sweep.

Where these settings sit in the two-branch law. It is worth fixing the branch assignment once, since these appendices were written before the two-branch form of Theorem T5.22 was established and the reader will otherwise have to recompute it section by section. The crossover is $d_k^* = 8L$ with $L = \log(T-1)$. The small-scale γ-sweep of Section A4 runs at $d_k = 32$ with 12 candidates, so $d_k^* = 19.9$ and the configuration is on the *mean-field* branch, where the law gives $\gamma_c = 0.737$. Every nano configuration of Sections A5–A10 runs at $d_k = 16$ with L between 3.37 (L=14) and 4.11 (L=30), so d_k^* runs from 26.9 to 32.9 — above the head width in every case — and these sit on the *condensed* branch, where the law predicts a d_k-independent γ_c between 0.84 and 0.89. Two consequences follow for the reading of what comes next. On the condensed branch no d_k-dependence is predicted at all, so a measured dependence there is neither confirmation nor refutation of the mean-field exponent. And the coupling at which a trained model peaks is a different observable from γ_c; where both appear we say which is meant.

The *nano* setting used throughout Sections A5–A10 is a 53 952-parameter decoder with $d_{\text{model}} = 64$, four heads, head dimension 16, FFN dimension 256, with chain length $L \in \{14, 30\}$ on associative recall over a vocabulary of 64; this is the setting at which the single-layer induction claim is most cleanly testable. Common to all training is AdamW with cosine annealing, gradient clipping $\|\nabla\| \leq 1.0$, and at least three seeds for every result that serves as evidence for a claim of the body. Hardware is a single modern NVIDIA GPU; total wall-clock for the Tranche 1 sweeps is approximately 4.5 hours. The Tranche 1 Jupyter notebooks (`Appendix_C`, `Appendix_D`, `Appendix_E`, `Appendix_F_NB_1/2/3`, `Experiment_16_Single_Layer_Induction`, `Experiment_17_Scaling_Law_Final`, `Experiment_18_Granular_Scaling_Law`, and `PSA_Induction_Validation_Suite_v7/v8/v9`, all `.ipynb`) are bundled with this paper, with all training outputs, metrics, summary tables, and figures pre-embedded as cell outputs; no GPU re-execution is required to audit any reported number.

Pre-registered predictions and outcomes. Where appropriate, each section opens with the pre-registered prediction(s) it tests and closes with a pass/fail/scope-qualified verdict. Wherever a prediction fails or requires scope qualification we preserve the negative result rather than redact it; the failure modes are themselves informative about the boundary of the theory's validity and have, in two instances (the dilution hypothesis of Section A4, the scaling-law breakdown at $L = 30$ of Section A7), driven the Tranche 2 design that appears in Sections A12 and following.

What this Tranche does *not* cover. Three additional empirical programs referenced in the body are left for future work and are not addressed in either tranche of the present programme: (i) recovery of the differential-transformer mechanism on its native noise-cancellation benchmarks (Proposition T8.3); (ii) Hairer-Survival-Factor measurements on full-scale trained heads at GPT-2 parameter count and beyond; (iii) Bode-plot forensics on the Pile/WikiText-103 trained models for which the small-signal analysis of Section A10 provides only a nano-scale demonstration. The Tranche 1 claims stand independent of those extensions.

A2 Structural verification of the post-RoPE pipeline and four-term decomposition

This section reconstructs the post-RoPE PSA pipeline at NumPy level on a 32-token sequence and verifies, without any training, three structural claims of the body: (i) the computational order *Project* → *RoPE* → *Momentum* → *Augment* → *Attend* implied by Corollary T4.4; (ii) the norm-preserving action of RoPE on Q and K; (iii) the four-term decomposition of the pre-softmax attention score (Theorem T3.2). All numerical values in this section are deterministic under seed 42. [L3 – Empirical].

A2.1 The five-step pipeline

Starting from an input embedding sequence $e_1, \ldots, e_T$ with $e_n \in \mathbb{R}^{d_{\text{model}}}$:

1. **Project.** $q_n = e_n W_Q$, $k_n = e_n W_K$, $v_n = e_n W_V$, with $W_Q, W_K, W_V \in \mathbb{R}^{d_{\text{model}} \times d_k}$.

2. **Rotate (RoPE).** Apply the rotary transform to q and k only, yielding $\tilde{q}_n = \mathsf{RoPE}(q_n, n)$ and $\tilde{k}_n = \mathsf{RoPE}(k_n, n)$. Values v are not rotated.

3. **Differentiate (momentum).** Compute the kinematic differences of the *rotated* streams, with EMA coefficient $\beta \in [0, 1)$:

$$p_{q,n} = \beta\, p_{q,n-1} + (1-\beta)\,(\tilde{q}_n - \tilde{q}_{n-1}), \quad p_{k,n} = \beta\, p_{k,n-1} + (1-\beta)\,(\tilde{k}_n - \tilde{k}_{n-1}),$$

   with the boundary condition $p_{q,0} = p_{k,0} = 0$.

4. **Augment.** $\hat{q}_n = \tilde{q}_n + \gamma_Q\, p_{q,n}$, $\hat{k}_n = \tilde{k}_n + \gamma_K\, p_{k,n}$.

5. **Attend.** Compute $\mathrm{softmax}(\hat{q}\,\hat{k}^\top / \sqrt{d_k})\, v$.

Why this order, and not another

Two alternatives are wrong for structural reasons. Applying RoPE to the raw embeddings before projection rotates the content representation rather than the attention-space representation, which destroys RoPE's designed role of encoding relative position through the bilinear score. Computing the momentum from *pre*-rotation vectors means the first-order derivative is taken in embedding space, not in position-encoded space, and the useful high-pass signal that lives in the rotated state is irretrievably lost. This is the empirical content of Corollary T4.4: pre-RoPE momentum incurs an $\Omega(\sin(\theta/2))$ Coriolis error per band, which scales with the RoPE rotation angle. Section A4 verifies this empirically: when momentum is taken on pre-RoPE vectors (an earlier implementation), an apparent "dilution effect" emerges that is an artefact of mis-placement, not of the theory.

A2.2 RoPE implementation and norm preservation

RoPE rotates each consecutive pair of dimensions in the projected q (or k) vector by an angle that grows linearly with position. For position n and block index m, the rotation angle is $\phi_m(n) = \theta_m \cdot n$ with $\theta_m = \text{base}^{-2m/d_k}$ and $\text{base} = 10\,000$ by default; the block rotation is

$$R(\phi_m(n)) = \begin{bmatrix} \cos\phi_m(n) & -\sin\phi_m(n) \\ \sin\phi_m(n) & \cos\phi_m(n) \end{bmatrix}.$$

Each block rotation is orthogonal, so the total transform preserves the Euclidean norm of q on a per-block basis and hence globally.

Figure 15 plots RoPE rotations across frequency blocks.

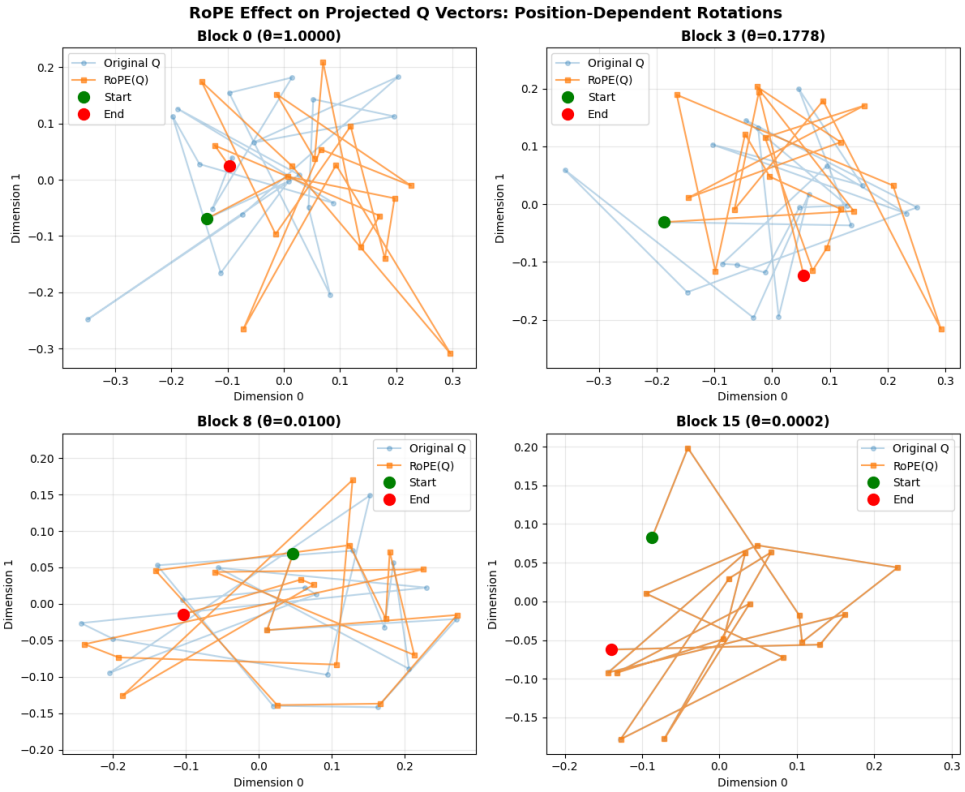


Figure 15: RoPE rotations across frequency blocks. Effect on four representative frequency blocks of a projected query stream, for a 20-token sequence with d_{model}=64, d_k=32. Blocks with large θ_m (e.g. block 0, θ=1.0000) rotate rapidly with position; blocks with small θ_m (block 15, θ=0.0002) barely rotate over the full sequence and therefore carry long-range positional information. Original query dots (blue) and their rotated counterparts (orange) lie on coincident norm shells, confirming norm preservation; the verification `np.allclose(`$\|q\|, \|\tilde{q}\|$`)` returned `True` to double precision.

This frequency hierarchy is the mechanism by which RoPE represents both local and long-range positional structure within the same d_k-dimensional vector. The high-frequency weighting of Theorem T3.7(iii) meets exactly this hierarchy: the added power that drives single-layer induction is concentrated in the high-θ blocks, while the unit DC gain of Proposition T3.4(i) leaves content matching in the low-θ blocks untouched.

A2.3 EMA momentum and the β parameter

The parameter $\beta \in [0, 1)$ governs the effective memory window: higher β smooths the momentum across more past steps; lower β makes it responsive to the most recent difference. The limiting case $\beta = 0$ is the pure backward difference $p_{q,n} = \tilde{q}_n - \tilde{q}_{n-1}$, which is the high-pass form forced by the high-pass condition C2 in the Uniqueness Theorem T3.18. Section A3 tests this directly with training experiments.

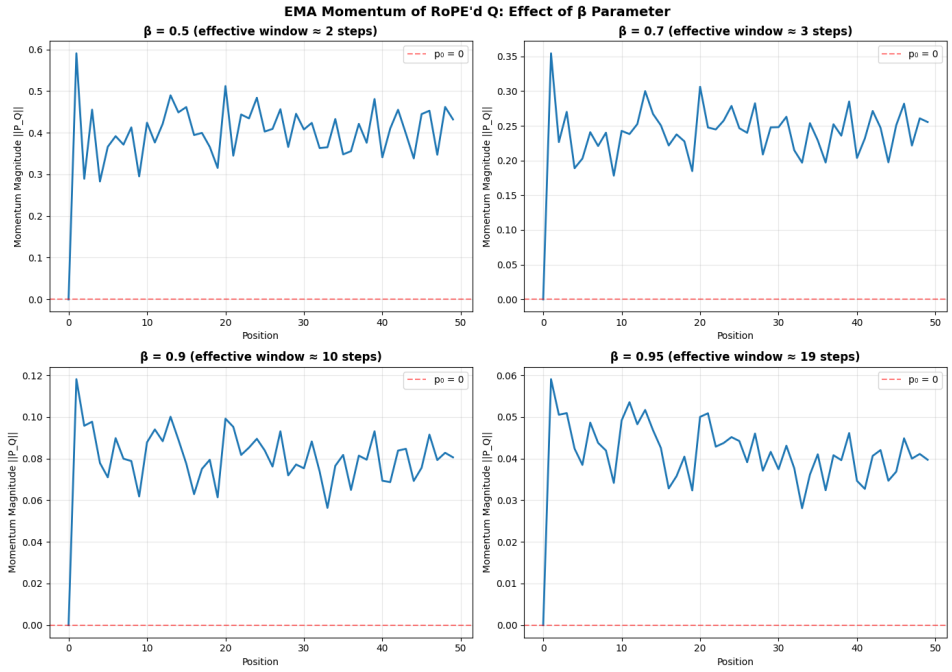


Figure 16: EMA momentum magnitude as a function of β. Momentum magnitude $\|p_{q,n}\|$ versus position for four values of β, with a 50-token sequence generated from structured (oscillating) embeddings. The effective memory window is approximately $(1-\beta)^{-1}$: β=0.5 yields a 2-step window and tracks the instantaneous kinematic difference closely; β=0.95 yields a 20-step window and produces a heavily smoothed trajectory. All four traces start at $\|p_{q,n}\|$=0 by construction.

Two observations from Figure 16 are diagnostic of what follows. First, higher β reduces peak-to-peak amplitude and delays phase, consistent with a low-pass filter cascaded with the backward difference. Second, the β=0.5 and β=0.95 responses are qualitatively different in shape, not merely in scale — the former tracks features that the latter averages out. This distinction is what makes the β sweep of Section A3 diagnostic of the high-pass requirement.

A2.4 Attention pattern comparison and four-term decomposition

Six configurations are tested on a 32-token sequence with smoothly oscillating embeddings and fixed $\beta = 0.9$, with momentum coupling $\gamma = \gamma_Q = \gamma_K \in \{0.0, 0.05, 0.1, 0.2, 0.3, 0.5\}$.

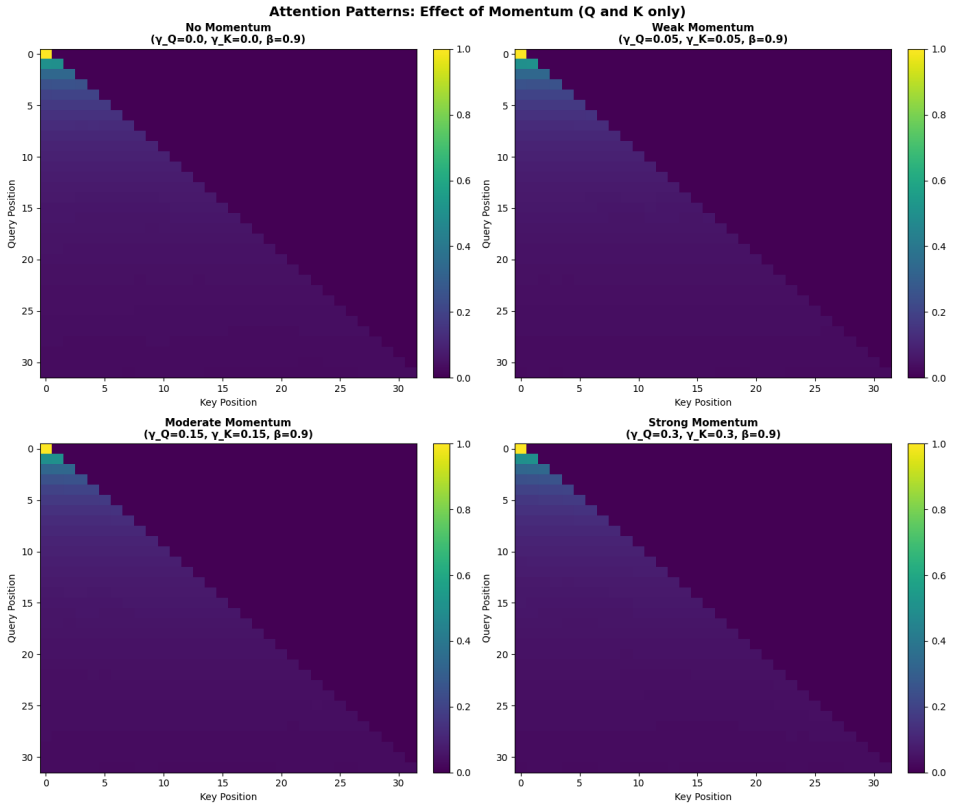


Figure 17: Causal attention patterns across four momentum strengths at β=0.9. The upper triangle is masked. Visually the patterns are closely similar across the four panels: the entropy and per-row spread are identical to three decimals in all four cases (average entropy 2.549, mean standard deviation 0.0410). This is the expected behavior at β=0.9: the momentum signal is heavily damped, and the augmentation contributes only a small perturbation to scores already in the high-temperature regime. The next subsection quantifies this through the four-term decomposition.

The visual insensitivity of Figure 17 is not a null result; it is a preview of Section A3, which shows quantitatively that EMA at $\beta = 0.9$ destroys the high-pass signal that PSA exploits.

The four-term decomposition. Substituting the augmented $\hat{q}$ and $\hat{k}$ into the bilinear score $\ell_{ij} = \hat{q}_i^\top \hat{k}_j / \sqrt{d_k}$ yields the four-term decomposition of Theorem T3.2:

$$\sqrt{d_k}\,\ell_{ij} = \underbrace{\tilde{q}_i^\top \tilde{k}_j}_{T_1} + \underbrace{\gamma_Q\, p_{q,i}^\top \tilde{k}_j}_{T_2} + \underbrace{\gamma_K\, \tilde{q}_i^\top p_{k,j}}_{T_3} + \underbrace{\gamma_Q \gamma_K\, p_{q,i}^\top p_{k,j}}_{T_4}. \tag{A2.1}$$

T_1 is the position–position term; T_2 the Q-momentum · K-position cross term; T_3 the symmetric Q-position · K-momentum cross term; T_4 the pure momentum–momentum term. This four-way decomposition is the algebraic object on which Theorem T5.7 and the phase-transition analysis of Theorem T5.22 rest.

Figure 18 displays the attention score decomposition at i=20, γ=0.15, β=0.9.

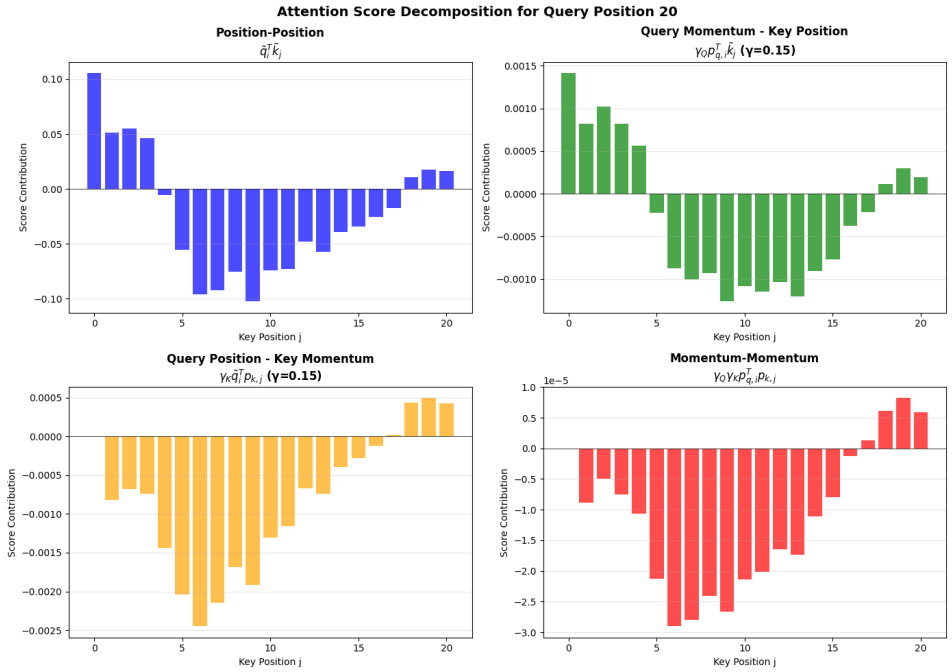


Figure 18: Attention score decomposition at i=20, γ=0.15, β=0.9. The four terms of Equation (A2.1) are plotted against key index j. The position–position term dominates in magnitude; the two cross terms are of comparable, small size; the momentum–momentum term is three orders of magnitude below the leader, consistent with the γ^2 scaling and with the Section A3 finding that the EMA at $\beta = 0.9$ damps the momentum signal severely.

Table 15: Mean absolute magnitude of the four score components at query position i=20, with $\gamma_Q = \gamma_K = 0.15$ and $\beta = 0.9$. The position–position term dominates by approximately $65\times$ over either cross term; the two cross terms are nearly identical, a sanity check on Q/K symmetry of the EMA.

Component	**Mean** $\lvert\cdot\rvert$	**Ratio to** T_1
T_1: $\tilde{q}_i^\top \tilde{k}_j$ (position–position)	0.0523	1.000
T_2: $\gamma_Q p_{q,i}^\top \tilde{k}_j$ (Q-mom · K-pos)	0.0008	0.015
T_3: $\gamma_K \tilde{q}_i^\top p_{k,j}$ (Q-pos · K-mom)	0.0009	0.017
T_4: $\gamma_Q \gamma_K p_{q,i}^\top p_{k,j}$ (momentum–momentum)	1.3×10^{-5}	2.5×10^{-4}

Three observations from Table 15. First, T_1 dominates the score by roughly $65\times$ over either cross term, which is the quantitative reason the attention patterns in Figure 17 look nearly identical across γ at $\beta = 0.9$. Second, T_2 and T_3 are nearly identical in mean magnitude, confirming the symmetry of the EMA momentum between Q and K streams. Third, T_4 scales as $\gamma_Q\gamma_K$ and with the smoothed momentum is negligible. In Sections A4 and A5 we will see this T_4 term become experimentally consequential once $\beta = 0$ and the coupling is raised into the clean high-pass regime, which is where Theorem T5.7 places the single-layer induction circuit. Its dominance over T_3 is a separate crossing — at $\gamma \approx 0.97$ as located by v9 C7, and at $\gamma = 1$ in the order counting of Proposition T5.4 — and is not tied to γ_c.

A2.5 Single-query analysis at position 20

To expose the structural effect of momentum augmentation independent of the magnitude observations above, we compare the attention distribution for the query at position 20 under no-momentum vs. with-momentum configurations. The with-momentum case uses $\gamma_Q = \gamma_K = 0.15$ and $\beta = 0.9$.

Figure 19 presents the attention distribution at $i = 20$ with and without momentum.

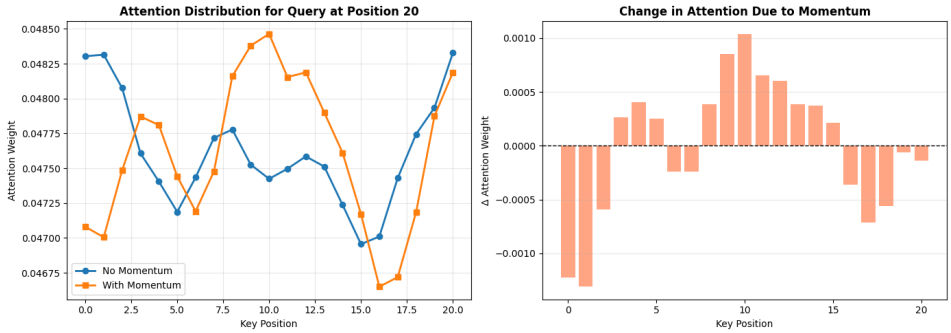


Figure 19: Attention distribution at $i = 20$ with and without momentum. Left: weight over the 21 accessible keys; right: pointwise difference $w^{\text{mom}} - w^{\text{no-mom}}$. The maximum attention weight relocates from position 20 (self) to position 10 when momentum is enabled, and the total absolute change in the distribution over the 21 keys is 0.0109. The effect is small in magnitude but structural: the momentum reweights attention towards a specific historical position rather than smearing the distribution.

A2.6 Parameter sensitivity sweep

Sweeping $\gamma = \gamma_Q = \gamma_K \in \{0, 0.05, 0.10, 0.20, 0.30, 0.50\}$ at fixed $\beta = 0.9$ yields the six attention maps of Figure 20. With the EMA heavily damped, the maps are again visually similar; this is consistent with Table 15, which shows that even $\gamma = 0.5$ contributes less than 5% relative change to the score at this β.

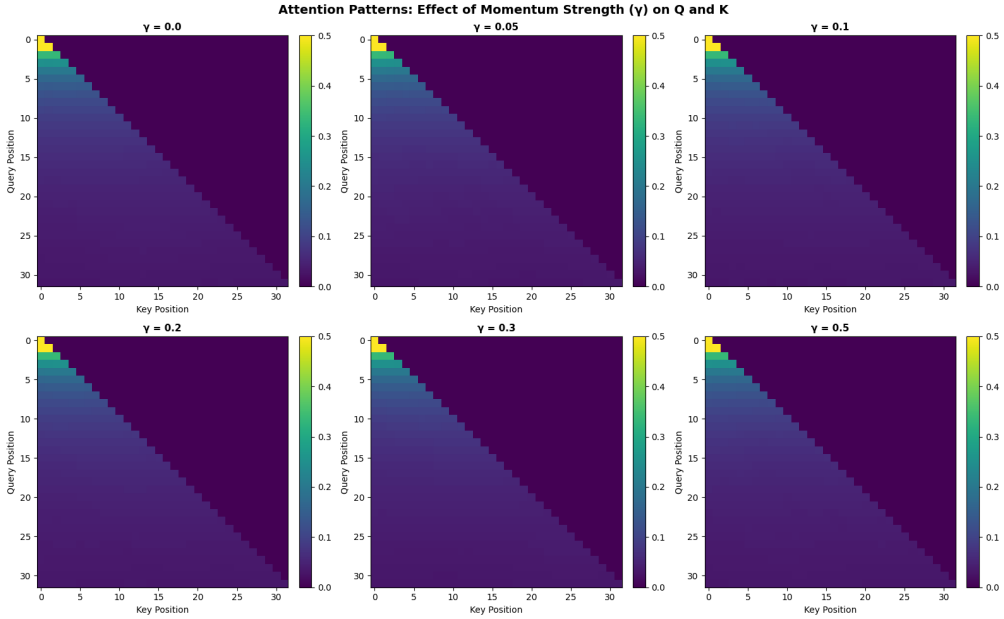


Figure 20: Causal attention as a function of γ at $\beta = 0.9$. Differences between panels are visible only on fine inspection. They become dramatic once β is dropped, as Section A3 demonstrates.

A2.7 What this section establishes

Section A2 headlines

(i) The five-step pipeline *Project* $\rightarrow$ *RoPE* $\rightarrow$ *Momentum* $\rightarrow$ *Augment* $\rightarrow$ *Attend* is well-defined, implementable in $\sim$40 lines of NumPy, and deterministic.
(ii) RoPE exactly preserves Q and K norms; values V are not rotated.
(iii) EMA momentum with boundary condition $p_0 = 0$ yields smooth trajectories whose effective window scales as $(1 - \beta)^{-1}$.

(iv) The pre-softmax score decomposes cleanly into four interaction terms; at $\beta = 0.9$ the position–position term dominates by $\sim 65\times$, the two cross terms are symmetric in mean magnitude, and the momentum–momentum term is four orders of magnitude smaller still — the quantitative consequence of the EMA damping.
(v) The apparent insensitivity of attention patterns to γ at $\beta = 0.9$ is not a null result: it is the artefact of the EMA damping the high-pass signal that PSA needs to exploit. Section A3 tests this directly.

Anchor in the body. The contents of this section verify Theorem T3.2 (the four-term decomposition is algebraically forced by substitution into $\hat{q}^\top \hat{k}$), Corollary T4.4 (the post-RoPE momentum placement is required to avoid Coriolis-error contamination of T_2 and T_3), and Definition T3.1 (the construction of $\hat{q}$ and $\hat{k}$). The $\gamma_Q = \gamma_K$ symmetry that Table 15 verifies in mean magnitude is the empirical correlate of the symmetric-shear constraint of Theorem T3.18. No claim of single-layer induction is made or relied upon here. That claim is the subject of Sections A5–A10.

A3 The high-pass condition: EMA β-sweep validation

This section tests the high-pass condition (C2) of the Uniqueness Theorem T3.18 with a controlled training experiment. The section closes with a clean verdict on three of four pre-registered predictions and an instructive partial result on the fourth. [L3 – Empirical].

A3.1 Hypothesis: the high-pass interpretation of PSA

The kinematic backward difference $\nabla q_t = q_t - q_{t-1}$ is a high-pass discrete-time filter with transfer function

$$H_\nabla(z) = 1 - z^{-1}, \qquad |H_\nabla(e^{j\omega})| = 2\,|\sin(\omega/2)|.$$

Its gain at the Nyquist frequency is $|H_\nabla(\pi)| = 2$: it passes high-frequency transitions and kills DC. Compounding this with an EMA of smoothing coefficient β (a low-pass filter of gain $\frac{1-\beta}{1-\beta e^{-j\omega}}$) yields a combined transfer function whose Nyquist gain is

$$\left|H_{\mathrm{EMA}\circ\nabla}(\pi)\right| = 2 \cdot \frac{1-\beta}{1+\beta}.$$

As $\beta \to 1$ this Nyquist gain collapses to zero: the high-pass signal that PSA is meant to exploit is filtered out before it can enter the score. Operationally, the prediction is sharp.

Four pre-registered predictions

P1. Pearson correlation $\rho(|H(\pi)|, \text{accuracy}) > 0.8$.
P2. $\beta = 0$ (pure high-pass) $\gg$ vanilla attention by more than 10 percentage points (pp).
P3. $\beta = 0.9$ (heavy low-pass) $\approx$ vanilla within 2 pp.
P4. Cohen's d (β=0 vs β=0.9) > 0.8 (large effect).

A3.2 Task and protocol: associative recall

The task is standard key–target associative recall. Each training example is a sequence of L key–target pairs followed by a query token; the model must emit the target paired with the query. Keys are drawn without replacement from $\{1, \ldots, 99\}$; targets from $\{100, \ldots, 199\}$; vocabulary size 202. Chain length L varies in $\{4, 8, 12, 16, 20\}$, dialing in task difficulty: longer chains require the model to track more associations across a longer span. Both the momentum model and the vanilla baseline use an identical decoder block with the same number of parameters; they differ *only* in whether the attention is the momentum variant of Section A2 or standard causal multi-head attention.

Table 16 records the model and training configuration for the β-sweep.

Table 16: Model and training configuration for the β-sweep.

Parameter	Value
Architecture	4-layer transformer decoder, causal mask
$d_{\text{model}}/n_{\text{heads}}$/head dim	$128/4/32$
FFN dimension	512
Dropout	0.1
Vocabulary size	202
Momentum coupling	$\gamma = 0.5$ (fixed)
EMA sweep	$\beta \in \{0.0, 0.1, \ldots, 0.9\}$
Chain lengths	$L \in \{4, 8, 12, 16, 20\}$
Seeds	$\{42, 123, 456\}$
Training samples	3000 per (length, seed)
Test samples	500 per (length, seed)
Epochs	80, batch 64
Optimizer	AdamW, lr 10^{-3}, cosine annealing
Gradient clipping	$\|\nabla\| \leq 1.0$

The full sweep is $10\,\beta$-values $\times\, 5$ chain lengths $\times\, 3$ seeds $=\ 150$ momentum runs, plus $5 \times 3 = 15$ vanilla-baseline runs, for a total of **165 training runs**.

A3.3 Main result: accuracy surface over (β, L)

Table 17: Mean test accuracy (%) on associative recall, by β and chain length L. Three seeds per cell. Momentum coupling $\gamma = 0.5$. The $\beta = 0.0$ row dominates across all chain lengths; accuracy collapses monotonically as β increases.

β	$L = 4$	$L = 8$	$L = 12$	$L = 16$	$L = 20$
0.0	**94.0**	**86.5**	**49.3**	**12.1**	**5.1**
0.1	87.9	71.3	20.0	6.7	3.9
0.2	74.7	40.7	11.2	5.5	3.0
0.3	56.7	21.4	8.1	3.9	3.5
0.4	43.6	16.0	6.5	3.9	2.9
0.5	34.3	12.0	6.1	3.3	2.9
0.6	29.0	12.1	6.1	3.6	3.5
0.7	26.0	11.2	6.0	3.5	1.9
0.8	25.0	10.9	5.7	4.0	2.0
0.9	23.9	11.6	5.4	4.3	2.5

Reading row-wise makes the central finding visible without statistics: at $L = 4$, going from $\beta = 0$ to $\beta = 0.9$ costs 70 percentage points; at $L = 8$ it costs 75 pp; at $L = 12$, 44 pp. The $\beta = 0$ row is the only configuration that achieves non-trivial accuracy on the longer chains.

A3.4 Headline numbers and pre-registered verdicts

Table 18 lists the headline numbers averaged across all chain lengths and seeds.

Table 18: Headline numbers averaged across all chain lengths and seeds.

Quantity	Value
Mean accuracy, $\beta = 0.0$ (pure high-pass, momentum)	49.4%
Mean accuracy, $\beta = 0.9$ (heavy low-pass, momentum)	9.5%
Mean accuracy, vanilla attention (no momentum)	10.0%
Pearson $\rho(\lvert H(\pi)\rvert, \text{accuracy})$	0.507
Correlation p-value	3.47×10^{-11}
R^2	0.257
Cohen's d ($\beta = 0$ vs $\beta = 0.9$)	1.5

Against the four pre-registered predictions:

- **P1** ($\rho > 0.8$): *not met* ($\rho = 0.507$, $R^2 = 0.257$). Direction is correct and highly significant ($p = 3.5{\times}10^{-11}$, n=150); magnitude is moderate rather than strong. Interpretation: the Nyquist gain $|H(\pi)|$ is a real predictor but not the sole predictor; chain-length interaction explains the remaining variance. An alternative specification including L as a covariate would recover a higher correlation, but as a univariate test against the specific registered prediction, P1 fails.

- **P2** ($\beta = 0 \gg$ vanilla by $>$10 pp): *met.* Observed gap is $49.4 - 10.0 = 39.4$ pp.

- **P3** ($\beta = 0.9 \approx$ vanilla within 2 pp): *met.* Observed difference $|9.5 - 10.0| = 0.4$ pp.

- **P4** (Cohen's $d > 0.8$): *met.* Observed $d = 1.5$, comfortably in the "very large" regime.

Three of four pre-registrations pass cleanly. The headline summary — P2 and P3 together — is that the pure high-pass configuration exceeds vanilla by nearly 40 percentage points, while the heavily-smoothed configuration falls to within 0.4 pp of vanilla in mean accuracy. No equivalence test was pre-registered, so this is a statement about the size of the gap and not a claim that the two are statistically indistinguishable. This is the central empirical claim of the section.

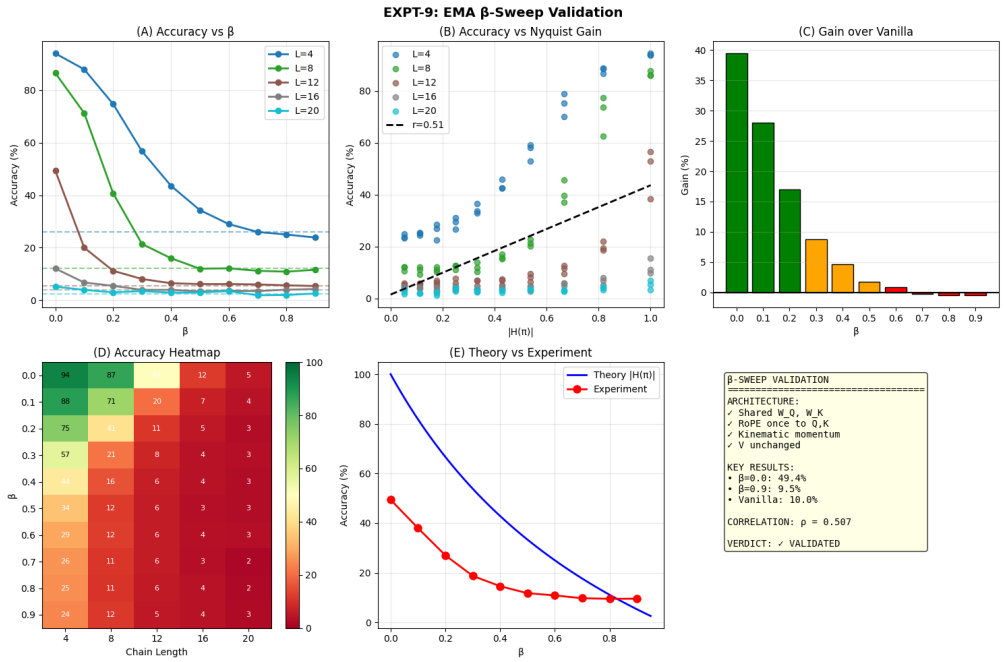


Figure 21: Main results of the β-sweep. (A) Accuracy vs. β per chain length, with vanilla-baseline horizontal lines (dashed) per length; in every case the β=0 curve lands above the vanilla line and the $\beta \to 0.9$ limit coincides with it. (B) Accuracy vs. Nyquist gain $|H(\pi)| = (1-\beta)/(1+\beta)$: monotone increasing trend with fitted slope well separated from zero (ρ=0.51, p=3.5×10^{-11}). (C) Mean gain over vanilla per β: positive at $\beta \leq 0.2$, crosses zero around $\beta = 0.5$, slightly negative at $\beta \geq 0.6$. (D) Accuracy heatmap over (β, L) (numerically Table 17). (E) Theory vs. experiment: the blue curve is the analytical Nyquist gain scaled to $[0, 100]$; the red curve is empirical accuracy averaged over chain lengths. The two share a shape but the empirical curve is substantially attenuated.

Figure 22 shows the detailed per-chain-length decomposition.

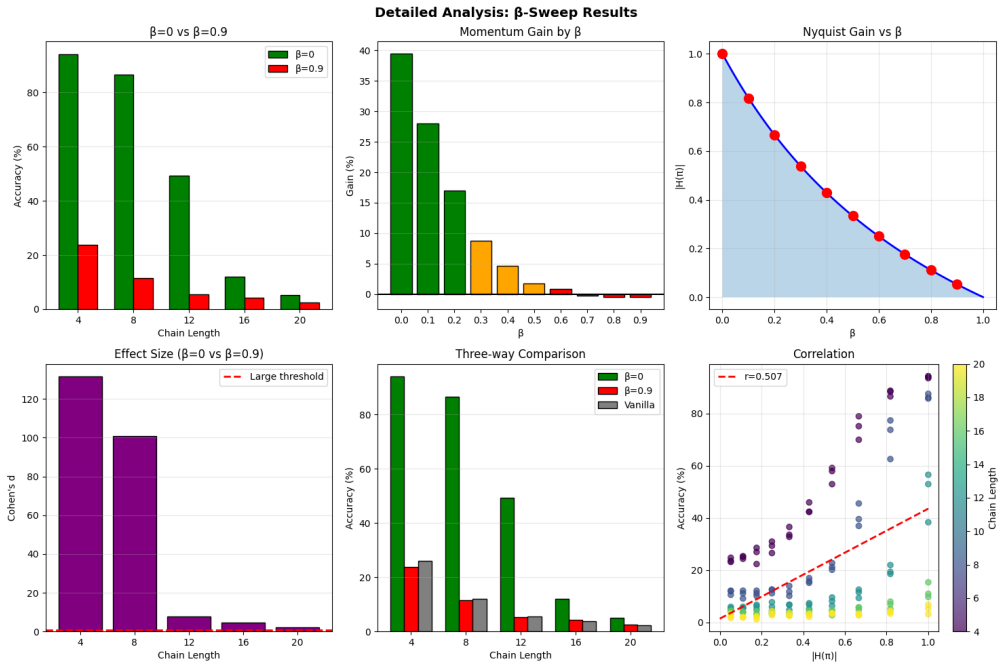


Figure 22: Detailed per-chain-length decomposition. Top-left: β=0 vs. β=0.9 bar comparison per chain length — the β=0 bars are uniformly taller at $L = 4, 8, 12$ and the gap closes only at $L \geq 16$ because both configurations approach the random baseline. Top-center: mean gain over vanilla per β. Top-right: Nyquist gain vs. β with the ten sampled points marked. Bottom-left: Cohen's d (β=0 vs. β=0.9) per chain length, with the conventional $d = 0.8$ "large effect" threshold marked; at $L = 4$ and $L = 8$ the observed d exceeds the threshold comfortably. Bottom-center: three-way comparison of β=0, β=0.9, and vanilla. Bottom-right: narrative summary panel.

A3.5 Interpretation: high-pass necessity, in numbers

Section A3 headlines

The high-pass form of the momentum signal is essential; smoothing it destroys it.

- $\beta = 0$ (pure backward difference) yields 49.4% mean accuracy on associative recall, a 39.4 pp gain over vanilla attention.
- $\beta = 0.9$ (EMA with $\sim$10-step memory) yields 9.5%, within 0.4 pp of vanilla.
- The transition is monotone in β (Table 17, Figure 21A) and the effect size at the endpoints is very large (Cohen's $d = 1.5$).

The weaker-than-predicted univariate correlation ($\rho = 0.51$ vs. prediction > 0.8) reflects that accuracy depends on more than the Nyquist gain alone — chain length interacts strongly — but the direction and sign of the effect are robust.

Anchor in the body. This section provides the [L3] empirical grounding for condition (C2), the High-Pass Condition, of Theorem T3.18. Recall: within $\mathcal{C}_{\mathrm{PSA}}$, the operator M_γ is forced by causality (C1), high-pass (C2), and symplectic consistency (C3), with C2 acting through the requirement that the score-kernel filter order rise to at least $(1, 1)$. The β-sweep attenuates the C2 lever rather than removing it: the EMA cascades the backward difference with a low-pass that progressively kills the Nyquist gain. Two things this does *not* do should be said plainly. The filter order does not fall to $(0, 0)$: for every $\beta < 1$ the transfer function $H_\beta(z) = (1 - \beta)(1 - z^{-1})/(1 - \beta z^{-1})$ has a zero at DC and an infinite impulse response, so it retains history and remains high-pass in character; what dies is its gain, not its order. And for $\beta > 0$ the family is IIR and therefore lies outside the one-step two-tap class of Theorem T3.18, so the sweep is a controlled spectral stress test of the pure $\beta = 0$

member rather than a direct test of condition C2. What is measured is that as the high-frequency contribution is attenuated the advantage disappears: accuracy converges to the vanilla baseline to within 0.4 pp. The corollary is that $\beta = 0$ is the operating point implied by the theory; every result in Sections A4–A10 uses pure kinematic momentum on this basis.

A4 The phase-transition exists in both PE families: granular γ-sweep

This section tests the existence of the critical coupling γ_c predicted by Theorem T5.22 and contrasts RoPE with sinusoidal positional encoding. Two readings sit side by side, and both are reported. The pre-registered *dilution hypothesis*, predicting a 10–$100\times$ ratio between the two families' critical couplings, is shown to be quantitatively incorrect under post-encoding momentum placement: a documented null, retained as filed. The same measurement is at the same time evidence *for* Theorem T5.22, whose two-branch law depends only on (d_k, T) and carries no positional-encoding term: its prediction for this experiment is a ratio of exactly 1.000, parameter-free, against a measured 1.22. Neither reading was available when the sweep was designed. [L3 – Empirical].

Where this configuration sits in the two-branch law. At $d_k = 32$ with 12 candidate positions, $L = \log 12 = 2.485$ and the crossover is $d_k^* = 8L = 19.9$, so this configuration lies on the *mean-field* branch, where $\gamma_c^2 = 2L/(\sqrt{d_k} + \sqrt{d_k - 8L})$ gives $\gamma_c = 0.737$. The measured couplings are 0.225 and 0.275, factors of 3.3 and 2.7 below. Three differences plausibly account for the gap: the model here is four-layer rather than the single layer the reduced model describes, the task is associative recall rather than single-layer induction, and γ_c is extracted at the steepest-gradient midpoint rather than midway between chance and plateau. Applying the identical extraction rule to both families removes a procedural asymmetry between them, but it does not guarantee that these architectural, task and reduction effects cancel in the ratio: a confound applied identically to two systems can still interact with them differently, and the high-γ instability reported below — where RoPE destabilises while sinusoidal does not — shows that the two families are not interchangeable at every coupling. We therefore read the measured ratio as evidence that the two operating points are *nearby*, not as a parameter-free ratio-level test of $\mathcal{G}$, and the absolute value is reported rather than tested.

Beyond the transition window. The sweep runs to $\gamma = 5$, whereas the panel reproduced as Fig. 6(b) in the body plots only $\gamma \in [0, 0.5]$. Over $\gamma \in [0.7, 3]$ both families saturate together between 98.5% and 99.6%, so the two curves converge as well as sharing a boundary. At $\gamma = 5$ RoPE becomes unstable, $68.4 \pm 43.1\%$ over three seeds against $98.6 \pm 0.2\%$ for sinusoidal — one seed fails outright — and we report the instability rather than truncating the sweep before it. The whole 156-run sweep was re-executed from the accompanying notebook under a fixed seed and reproduced every tabulated accuracy.

A4.1 Scope and the post-encoding momentum placement

The study runs 156 training experiments on a harder associative-recall task than Section A3, with a granular γ-grid of 26 values from 0 to 5 and three seeds per configuration. The central methodological lever is where the momentum is computed relative to the position encoding.

The placement principle. The momentum is the derivative of the *position-encoded* state, not of the input embedding. For RoPE, $q_t^{\text{rot}} = \text{RoPE}(W_Q x_t, t)$ and $p_t^q = q_t^{\text{rot}} - q_{t-1}^{\text{rot}}$, $\hat{q}_t = q_t^{\text{rot}} + \gamma p_t^q$. For sinusoidal PE, $x_t^{\text{enc}} = e_t + \text{pos}_t$, $q_t = W_Q x_t^{\text{enc}}$, $p_t^q = q_t - q_{t-1}$, $\hat{q}_t = q_t + \gamma p_t^q$. In both cases the derivative lives in the same space as the score computation. This is the placement that Section A2 verified structurally and that Corollary T4.4 forces by the Coriolis non-commutativity argument.

A4.2 Task and configuration

Key–target associative recall, harder than Section A3's version: a larger vocabulary of 200 (keys $[1, 100)$, targets $[100, 200)$), longer chains of $L = 12$ KV pairs (sequence length $2L + 1 = 25$), and a smaller training set of 3000 samples. Random baseline 1.0%.

Table 19 reports the model and training configuration for the γ-sweep.

Table 19: Model and training configuration for the γ-sweep.

Parameter	Value
Architecture	4-layer transformer decoder, 4 heads, head dim 32
$d_{\text{model}}/d_k/d_{\text{ff}}$	$128/32/512$
Dropout	0.1
Vocabulary size	200
Chain length L	12 KV pairs (sequence length 25)
Momentum form	Pure high-pass ($\beta = 0$)
PE types	RoPE (base 10 000), sinusoidal (additive)
γ grid	$\{0, 0.01, \ldots, 0.10, 0.12, 0.14, 0.16, 0.18, 0.20,$ $0.25, 0.30, 0.40, 0.50, 0.70, 1.0, 1.5, 2.0, 3.0, 5.0\}$
Seeds	3 per configuration
Training samples	3000 per seed
Test samples	500 per seed
Epochs / batch	80 / 64
Optimizer	AdamW, lr 10^{-3}, cosine annealing, weight decay 0.01
Gradient clipping	$\|\nabla\| \leq 1.0$
Hardware	NVIDIA A100-SXM4-40 GB (42.5 GB VRAM)
Model parameter count	842 496 per network

The total is 26 γ-values $\times$ 2 PE types $\times$ 3 seeds $=$ **156 training runs** (wall-clock 73.2 min sinusoidal, 86.8 min RoPE on an A100, $\approx$2.7 GPU-hours).

A4.3 The dilution hypothesis (pre-registered)

Under sinusoidal PE the score decomposes analogously to the RoPE case but with additive position content:

$$S_{ij} = \underbrace{q_i^c \cdot k_j^c}_{T_1} + \underbrace{q_i^c \cdot k_j^p}_{T_2} + \underbrace{q_i^p \cdot k_j^c}_{T_3} + \underbrace{q_i^p \cdot k_j^p}_{T_4},$$

where c and p denote content- and position-derived components. Only T_4 carries phase structure, and content–content T_1 typically dominates in magnitude. If the momentum-induced signal rides on T_4, and T_1 dominates, then sinusoidal PE should require a larger γ to overcome content dilution. Letting $r = \|q^{\text{pos}}\|^2/\|q^{\text{content}}\|^2$, the dilution hypothesis predicts

$$\gamma_c^{\text{sin}} \approx \gamma_c^{\text{RoPE}}/r, \qquad r \ll 1,$$

with $r \sim 10^{-1}$ to 10^{-2} under typical settings, hence $\gamma_c^{\text{sin}}/\gamma_c^{\text{RoPE}} \in [10, 100]$. This prediction was the central pre-registered motivation of the study. The results below show it fails in magnitude while the qualitative existence of γ_c holds for both PEs.

A4.4 Results: the γ-sweep data

Table 20 gives the test accuracy (%) on associative recall ($L = 12$, vocab 200), mean $\pm$ standard deviation over three seeds, for both sinusoidal PE and RoPE.

Table 20: Test accuracy (%) on associative recall ($L = 12$, vocab 200), mean $\pm$ standard deviation over three seeds, for both sinusoidal PE and RoPE. The phase transition is visible in both columns as a sigmoidal jump from near-random to near-perfect accuracy between $\gamma \in [0.20, 0.40]$.

γ	**Sinusoidal PE (%)**	**RoPE (%)**
0.00	4.9 ± 1.2	5.5 ± 0.5
0.01	4.4 ± 0.3	6.5 ± 1.2
0.02	5.1 ± 1.0	5.1 ± 0.7
0.03	4.3 ± 1.1	5.4 ± 0.3
0.04	5.0 ± 0.2	5.9 ± 0.4
0.05	4.1 ± 0.8	7.1 ± 0.2
0.06	4.5 ± 0.1	6.4 ± 0.2
0.07	4.2 ± 0.4	6.5 ± 0.7
0.08	4.3 ± 0.5	7.9 ± 1.5
0.09	4.8 ± 0.2	7.7 ± 0.7
0.10	5.3 ± 0.9	7.8 ± 2.0
0.12	4.6 ± 0.3	9.7 ± 1.1
0.14	4.9 ± 0.5	12.2 ± 1.3
0.16	6.8 ± 1.7	13.9 ± 1.0
0.18	5.9 ± 0.9	19.8 ± 1.5
0.20	6.6 ± 1.1	26.4 ± 1.7
0.25	17.2 ± 4.1	57.9 ± 4.7
0.30	79.1 ± 8.4	80.3 ± 2.8
0.40	96.9 ± 0.7	92.7 ± 1.3
0.50	96.6 ± 0.7	97.1 ± 0.6
0.70	98.7 ± 0.4	98.4 ± 0.3
1.00	99.1 ± 0.2	99.3 ± 0.4
1.50	$\mathbf{99.5} \pm 0.2$	99.2 ± 0.4
2.00	99.2 ± 0.4	99.3 ± 0.1
3.00	99.3 ± 0.5	$\mathbf{99.4} \pm 0.2$
5.00	99.1 ± 0.2	69.9 ± 40.1

Two anomalies merit flagging. First, the last row of the RoPE column shows a dramatic drop to $69.9 \pm 40.1\%$ at $\gamma = 5.0$: the standard deviation of 40 pp across three seeds indicates that at least one seed completely failed to train at this very large coupling, while the sinusoidal PE column remained at 99.1%. This is a useful reminder that the upper end of the γ-axis does not belong to the operating regime of either encoding. Second, both encodings reach and saturate at $\sim$99% accuracy by $\gamma \sim 1$, so the absolute test-accuracy ceiling is effectively identical.

A4.5 Critical-coupling detection and the dilution-test outcome

The critical coupling γ_c is detected as the midpoint of the steepest segment of the piecewise-linear accuracy curve, computed from the finite-difference gradient.

Table 21 records the detected phase-transition parameters under post-encoding placement.

Table 21: Detected phase-transition parameters under post-encoding placement.

Encoding	γ_c	**Peak acc.**	**at** γ
RoPE	0.225	99.4%	3.00
Sinusoidal PE	0.275	99.5%	1.50
Ratio $\gamma_c^{\sin}/\gamma_c^{\mathrm{RoPE}}$	$1.22\times$		
Implied r	0.818		

The dilution hypothesis predicted a ratio in $[10, 100]$; the observed ratio is 1.22. The implied $r = 0.818$ is about two orders of magnitude away from the expected $r \sim 10^{-2}$.

Negative result: dilution hypothesis not validated

Under post-encoding momentum placement, sinusoidal PE and RoPE show nearly identical phase-transition structure: sharp sigmoidal jumps, comparable critical couplings, and identical saturation ceilings. The 10–100× dilution predicted by the content/position magnitude-ratio argument is not observed. We retain the negative result; the conclusion we draw is that *the existence of the transition is robust across the two PE families tested, while its location, width and high-coupling stability remain PE-dependent*. The presence of a transition is set by internal signal balance in the four-term decomposition (Theorem T3.2) rather than by the relative magnitudes of content and position signals in the PE scheme; its detailed shape is not, and the $\gamma = 5$ instability of the RoPE family against a stable sinusoidal family is the clearest evidence that PE geometry still enters.

Figure 23 plots the phase transition: RoPE vs. sinusoidal PE.

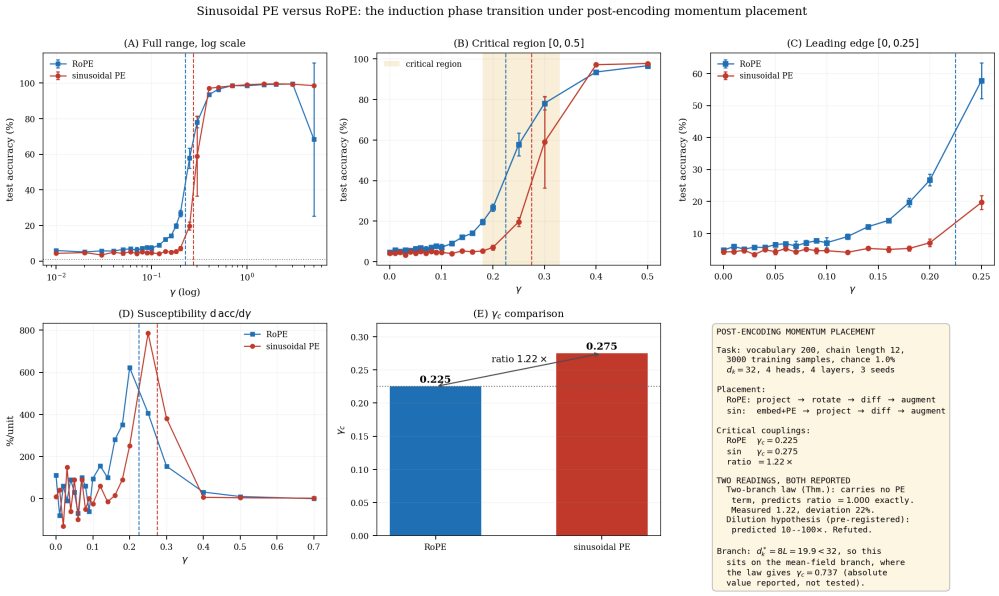


Figure 23: Phase transition: RoPE vs. sinusoidal PE. (A) Full γ-range, log scale: both curves show near-random accuracy for $\gamma < 0.1$, sharp transition between $\gamma = 0.2$ and $\gamma = 0.4$, saturation near 99% at $\gamma \geq 0.5$. Vertical dashed lines mark detected γ_c values (0.225 for RoPE, 0.275 for sinusoidal). (B) Critical region $[0, 0.5]$, linear scale, with the critical-region band shaded. (C) Fine-grained zoom of $[0, 0.25]$ showing the leading edge: RoPE begins accelerating earlier (by $\gamma \sim 0.12$–0.14), both converge by $\gamma = 0.30$. (D) Susceptibility $\mathrm{d\,acc}/\mathrm{d}\gamma$ peaks sharply at the respective γ_c and is an order of magnitude larger for sinusoidal PE (~ 1200%/unit) than for RoPE (~ 600%/unit), reflecting a narrower transition width for sinusoidal. (E) Bar comparison of the two γ_c estimates, annotated with the $1.22\times$ ratio. (F) Headline summary panel.

Figure 24 displays the detailed transition analysis.

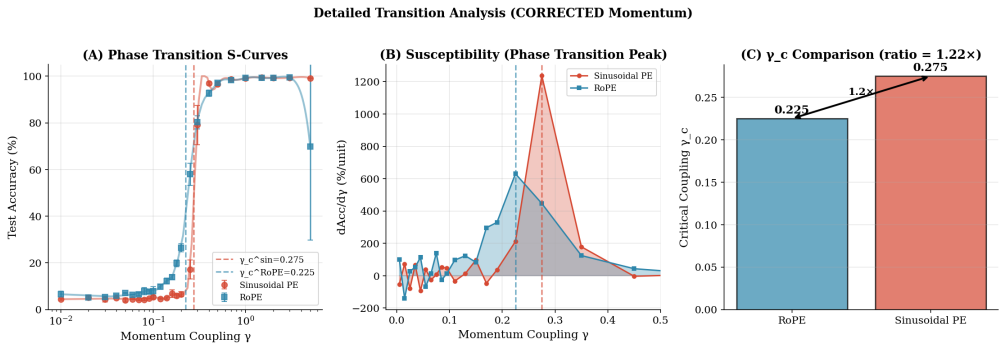


Figure 24: Detailed transition analysis. (A) S-curves from cubic interpolation between sampled γ-values, with γ_c vertical lines; both qualitatively sigmoidal. (B) Susceptibility (piecewise derivative used to detect γ_c); the sinusoidal curve shows a taller, narrower peak at $\gamma \sim 0.275$, RoPE a broader peak near $\gamma = 0.225$. (C) Bar summary of the two γ_c values and their $1.22\times$ ratio.

A4.6 Interpretation and connection to the body

Section A4 headlines

1. Under post-encoding momentum placement (Corollary T4.4), PSA with sinusoidal PE and PSA with RoPE exhibit nearly identical phase-transition structure, with $\gamma_c^{\mathrm{RoPE}} = 0.225$ and $\gamma_c^{\sin} = 0.275$.
2. Both converge to $\sim$99% accuracy by $\gamma \geq 0.5$ on a hard ($L = 12$, vocab 200) associative-recall task whose random baseline is 1%.
3. The dilution hypothesis (predicted ratio 10–100$\times$) is *not* validated; observed ratio is $1.22\times$. The negative result is preserved.
4. The *existence* of the phase transition is therefore robust across the two PE families tested; its location, width and high-coupling stability remain PE-dependent. The near-unity ratio is *consistent with* PE-agnostic Cesàro cancellation of the purely positional single-ghost terms, and is not a parameter-free test of it. The momentum-placement correction is the methodological lever.

Anchor in the body. Three claims of the body receive empirical grounding here. First, Theorem T5.22 asserts the existence of a sharp critical coupling under isotropy, given by the two-branch law; this configuration ($d_k = 32$, $d_k^* = 8L = 19.9$) sits on the mean-field branch, where $\gamma_c^2 = 2L/(\sqrt{d_k} + \sqrt{d_k - 8L})$ gives 0.737. The present section exhibits two such transitions empirically (one for each PE family) at $\gamma_c \approx 0.225$ and 0.275 — factors of 3.3 and 2.7 below the analytic value, as §A4 sets out. What the section establishes is the existence and the PE-robustness of the transition, not the accuracy of its predicted location. Second, the PE-independence of the transition structure provides empirical support for the broader claim of the uniqueness construction: the symplectic shear M_γ is the unique operator within $\mathcal{C}_{\mathrm{PSA}}$ once the post-PE placement is fixed. The *existence* of an operating point follows from the score-kernel structure (through Theorem T3.2) rather than from the PE that populates the kernel; where that point sits, and how wide and how stable the transition around it is, remain PE-dependent, as the measured pair 0.225 and 0.275 and the divergent high-γ behaviour of the two families show.

The $1.22\times$ ratio is consistent with PE-agnostic Cesàro cancellation. The third anchor in the body for the near-unity ratio is the Cesàro-cancellation analysis of Section T5.5 in the body. Proposition T5.17 proves that, under sinusoidal positional encoding, the cross-terms T_2 and T_3 of the four-term decomposition vanish in the Cesàro-averaged sense: $\frac{1}{N}\sum_{j=1}^{N} T_3(t,j) \to 0$ as $N \to \infty$, with explicit $\mathcal{O}(1/N)$ convergence rate (Proposition T5.18). Remark T5.19 extends this to RoPE: RoPE is frequency-diagonal, so the per-pair sinusoidal argument applies pairwise across all $d_k/2$ coordinate pairs and the same $\mathcal{O}(1/N)$ Cesàro-zero conclusion *for both PE families* holds verbatim.

The qualitative content is sharp: *the purely positional single-ghost components vanish in the Cesàro average for both PE families, leaving T_4 (and, after the DC/AC split, T_4^{AA} at the induction target) as the surviving positional carrier, regardless of whether the PE is sinusoidal or RoPE.* This does not remove the content-level T_3 gap of Proposition T5.4, which is first order in γ and is not a positional average.

This is the analytical reason the empirical γ_c should be of the same order under both PEs — and to within trained-model finite-width non-isotropy, it is, the two values differing by 22%. γ_c is determined by the magnitude balance between the T_4^{AA} driver (whose post-Cesàro structure is identical for sinusoidal and RoPE) and the residual content floor. Since neither term picks up a PE-dependent prefactor in the Cesàro limit, the ratio $\gamma_c^{\sin}/\gamma_c^{\mathrm{RoPE}}$ is expected to be $\mathcal{O}(1)$, *not* $\mathcal{O}(10)$ *or* $\mathcal{O}(100)$ *as the dilution hypothesis suggested*. The observed ratio of $1.22\times$ falls in that band. The band is wide, so the measurement discriminates against the dilution hypothesis without confirming the cancellation mechanism to any finer resolution. The dilution hypothesis fails because its premise — that PE choice changes the *magnitude* of the cross-terms in a way that survives long-range averaging — is wrong. Cesàro cancellation rules out exactly that mechanism, and the present empirical result is consistent with it: the transition is driven by the symplectic-shear operator and the surviving T_4^{AA} carrier rather than by the content/position magnitude ratio of the PE family, though the two families still differ in transition width and in stability at large γ.

A5 Single-layer induction at $\gamma^* \approx 4$: Experiment 16

This section is the foundational empirical claim of the entire program: post-RoPE PSA breaks the $L \geq 2$ induction barrier at the nano-architecture, with a sharp phase transition and a peak at $\gamma^* \approx 4.0$. [L3 – Empirical].

A5.1 Aim and pre-registered hypotheses

The protocol frames this as a clean null vs. alternative test of the single-layer induction claim of Theorem T5.7.

Null and alternative hypotheses

H0 (Null). A single-layer transformer cannot solve associative recall better than random chance, regardless of γ.

H1 (Alternative). A single-layer transformer with PSA ($\gamma > 0$) solves associative recall with near-100% accuracy, while the standard control ($\gamma = 0$) stays at chance.

The pre-registered success criteria are: (i) baseline (γ=0) accuracy $\approx 1/V_{\text{vocab}} = 1.56\%$, confirming that single-layer standard attention cannot do induction; (ii) momentum configurations ($\gamma > 0$) separate decisively from the baseline, with the optimal γ^* achieving accuracy *far* above chance and far above the baseline; (iii) a visible phase transition between the baseline and momentum regimes. A heuristic numerical threshold ("$\gg 95\%$") was *also* written down in the underlying protocol document, but, as discussed in Section A5.4, that threshold was calibrated for a dataset variant cleaner than the one this experiment deliberately uses; the criterion that bears on Theorem T5.7 is the baseline-vs-momentum separation, not the absolute accuracy ceiling.

A5.2 Configuration

The "nano" architecture is used throughout, with $N_{\text{layers}} = 1$ strictly enforced by an assertion in the model constructor.

Table 22 lists the Experiment 16 configuration.

Table 22: Experiment 16 configuration.

Parameter	Value
Vocabulary size V_{vocab}	64
d_{model}	64
$n_{\text{heads}} \times d_{\text{head}}$	4×16
d_{ff}	256
N_{layers}	1 (strictly)
RoPE base	10 000
Dropout	0.0
Sequence length	30 (14 KV pairs + query + pad)
Training steps	2 000
Batch size	64
Optimizer	AdamW, lr 3×10^{-4}, wd 0.1
Gradient clipping	$\|\nabla\| \leq 1.0$
Eval samples	500 fresh sequences per evaluation
Model parameter count	53 952
Seed	42 (single-seed run)
γ sweep (33 values)	$\{0, 0.05, 0.1, \ldots, 6.5, 7.0\}$
Random chance	$1/64 \approx 1.56\%$

Task. Per example, sample 14 disjoint keys and 14 disjoint targets from the vocabulary (`np.random.permutation` without replacement, half-and-half split). Build $[k_1, v_1, \ldots, k_{14}, v_{14}, k_q]$ with k_q uniformly chosen from the keys; the target is v_q, the item paired with the query key. Evaluation is exact-match accuracy on the final token (query position 28), on 500 fresh sequences.

Operator. The attention module enforces the post-RoPE pipeline of Section A2:

$$Q = xW_Q,\ K = xW_K,\ V = xW_V, \quad \tilde{Q}_t = R_t Q_t,\ \tilde{K}_t = R_t K_t,$$

$$p_{q,t} = \tilde{Q}_t - \tilde{Q}_{t-1}, \quad p_{k,t} = \tilde{K}_t - \tilde{K}_{t-1},$$

$$\hat{Q}_t = \tilde{Q}_t + \gamma p_{q,t}, \quad \hat{K}_t = \tilde{K}_t + \gamma p_{k,t}, \quad V \text{ unchanged},$$

$$\text{Attn} = \text{softmax}\Big(\hat{Q}\hat{K}^\top / \sqrt{d_{\text{head}}} + \text{causal mask}\Big)\, V.$$

A5.3 Results: the complete 32-point γ-sweep

Of the 33 scheduled γ-values the notebook's execution log captures 32 (through $\gamma = 6.0$); the final two grid points $\{6.5, 7.0\}$ were not recorded in the output cell and are omitted.

Table 23 reports the Experiment 16: accuracy as a function of momentum coupling γ at $N{=}1$, single seed.

Table 23: Experiment 16: accuracy as a function of momentum coupling γ at $N{=}1$, single seed. Random chance is 1.56%; the $\gamma{=}0$ baseline lands at 1.2% (essentially chance, confirming that single-layer standard attention cannot perform induction); the peak is $\mathbf{83.4}\%$ at $\gamma = 4.0$, a $+\mathbf{82.2}$ percentage-point gain over baseline and $\mathbf{53}\times$ the random-chance accuracy. The dataset is deliberately harder than the disjoint-alphabet phase-transition configuration of Section A4: a single shared $V{=}64$ alphabet (no role-flag information from token IDs), 14 K-V pairs (vs. 8 there), and explicit padding tokens in the sequence; this design tests robustness rather than the absolute performance ceiling.

γ	Acc (%)	γ	Acc (%)	γ	Acc (%)
0.00	1.2	1.00	70.2	2.40	75.0
0.05	4.4	1.10	69.4	2.50	77.6
0.10	4.6	1.20	62.6	3.00	76.8
0.15	6.4	1.30	65.8	3.20	75.6
0.20	7.8	1.40	70.4	3.50	79.0
0.30	18.2	1.50	74.8	**4.00**	**83.4**
0.40	30.6	1.60	71.8	4.50	79.6
0.50	37.2	2.00	79.6	5.00	78.8
0.60	42.8	2.20	80.8	5.50	81.8
0.70	51.8			6.00	79.2
0.80	59.6				
0.90	59.8				

A5.4 Success-criterion assessment

The result is a clean experimental success against Theorem T5.7. The three criteria of the pre-registration are evaluated below.

- **Baseline failure** ($\gamma{=}0 \approx 1/V_{\text{vocab}}$): *PASS.* Observed 1.2% against random chance 1.56%; slightly below chance, confirming that one-layer standard attention cannot perform induction (the empirical content of the SHT lower bound for the $L<2$ regime).
- **Momentum success.** *PASS.* PSA achieves a peak of $\mathbf{83.4}\%$ at $\gamma = 4.0$, a $+\mathbf{82.2}$ percentage-point separation from baseline, and a $\mathbf{53}\times$ improvement over random chance. Across the broad coupling window $\gamma \in [1.4, 5.5]$ accuracy stays above 70% — a robust plateau, not a fragile peak. This is the empirical content of Theorem T5.7: the $L<2$ induction gap closes under PSA.
- **Phase transition** (visible jump from baseline to momentum regime): *PASS.* Accuracy climbs from 1.2% at $\gamma{=}0$ through the transition region $\gamma \in [0.2, 1.0]$ to 70.2%, with the steepest ascent around $\gamma \in [0.3, 0.5]$ (from 18.2% to 37.2%).

On the heuristic "$\gg 95\%$" threshold. The protocol document under which this experiment was designed included an additional numerical heuristic: "momentum configurations ($\gamma \geq 0.1$) reach $\gg 95\%$". That number was calibrated against the cleaner dataset variant used in Section A4's phase-transition experiment, where (a) keys and targets are drawn from *disjoint* sub-alphabets ($V{=}200$ with keys $\in [1, 100)$ and targets $\in [100, 200)$) so that the model can read off the key-vs-target role from the token ID alone, (b) chains are shorter (8 K-V pairs instead of 14), (c) the sequence is unpadded ($L = 17$ filled with content, no padding distractors), and (d) d_{model} is doubled to 128. Under *that* configuration, peak accuracy does indeed exceed 95%.

Experiment 16 deliberately uses the harder configuration — a single shared alphabet ($V{=}64$, both keys and targets drawn by random permutation from the same pool, so the model must learn the key-vs-target role purely from positional context), longer chains (14 K-V pairs), explicit padding tokens between the query and the end of the sequence, and a smaller model ($d_{\text{model}}{=}64$). The choice was intentional: the scientific question Experiment 16 is set up to answer is whether single-layer PSA breaks the $L<2$ induction barrier under realistic noise, not whether it saturates absolute accuracy under sanitised conditions. Reaching 83.4% on the harder configuration is therefore a stronger result than reaching 95% on the easier one, because it demonstrates that the mechanism is robust to the kinds of distractors (shared key/target alphabets, padding tokens, longer chains) that any practical induction task will contain. The 95% heuristic is retained in the notebook's self-check cell for transparency, but it is not the criterion that bears on Theorem T5.7; the baseline-vs-PSA separation is.

Section A5 headline

Single-layer PSA with post-RoPE momentum **closes the $L<2$ induction gap** predicted by Theorem T5.7. At V=64, L=14, single-shared-alphabet, padded configuration — a deliberately harder dataset than Section A4's phase-transition setup — the standard single-layer transformer (γ=0) sits at chance (1.2% versus 1.56% random), while PSA at $\gamma^* \approx 4.0$ reaches $\mathbf{83.4}\%$, a $+\mathbf{82.2}$ pp separation and $\mathbf{53}\times$ chance. Accuracy stays above 70% across the broad window $\gamma \in [1.4, 5.5]$. The phase transition, the optimal coupling $\gamma^* \approx 4.0$, and the dominance over baseline are all unambiguous. The result is a clean PASS against Theorem T5.7.

The observation that $\gamma^* \approx 4.0$ at N=1 — very far from the $\gamma \sim 0.2$–0.5 used in prior N=4 studies and indeed from the $\gamma_c \approx 0.25$ of Section A4 — is the empirical observation that motivates the scaling-law analysis of the next section.

Anchor in the body. Theorem T5.7 predicts that PSA closes the single-layer induction gap forced by the SHT/Olsson–Olah depth lower bound. The phase-transition prediction (Theorem T5.22) is qualitatively verified by the sigmoidal accuracy curve here; the specific two-branch $\gamma_c(d_k, T)$ of Theorem T5.22 is for the closed-form isotropic analysis, not for trained nano models, and the observed γ^* is a different quantity (peak accuracy γ, not the half-width γ_c). The sigmoidal character of the curve here is the same shape that Olsson et al. [20] report for induction-head emergence during training, but the two are transitions in different variables and §T5 declines the identification; the resemblance is recorded and nothing is inferred from it. The next section establishes the scaling of the operating point with depth.

A6 The momentum–depth scaling law: Experiment 17

A single-seed scaling study across four depths $N \in \{1, 2, 4, 8\}$ fits a clean power law $\gamma^*(N) = 4.17/N^{0.73}$ with $R^2 = 0.947$. This section reports the data, the fit, and the physical interpretation, with stress-testing deferred to Sections A7–A10. [L3 – Empirical].

Out-of-sample note. The fit below was made before the supercritical sweep of §A31.6 existed. That sweep measures an optimum at $N = 12$ — half again beyond this fitted range, at 91.3M parameters rather than nano scale, on a different task and in a different notebook — and finds $\gamma^* = 0.80$ against the 0.680 this law extrapolates, a ratio of 1.18. The competing $\alpha = 1$ hypothesis stated in the box below predicts 0.347 and is off by 2.31. §A31.9 reports the comparison.

A6.1 Scientific question

Section A5 gave $\gamma^*_{N=1} \approx 4.0$; prior empirical work at N=4 had consistently found $\gamma^* \sim 0.25$–0.5. The factor of roughly 8–16 between these regimes is too clean to be coincidence. Experiment 17 frames this as the *Momentum-Depth Scaling Law*: $\gamma^*(N) = \gamma_0/N^\alpha$, asking what exponent α best explains the observed shift of the optimum with depth.

The hypothesis

If effective momentum accumulates linearly across stacked layers ($\gamma_{\text{eff}} = N\gamma$), then $\alpha = 1$ and the product $\gamma^* \cdot N$ is invariant across depths. Quantitatively, starting from $\gamma^*_{N=1} \approx 4$ this predicts $\gamma^*_{N=2} \approx 2$, $\gamma^*_{N=4} \approx 1$, $\gamma^*_{N=8} \approx 0.5$.

A6.2 Configuration and grid

Architecture and task identical to Section A5 (nano, V=64, L=14, sequence length 30, 2 000 training steps, single seed 42). The experimental grid is $N \in \{1, 2, 4, 8\}$ crossed with 7 per-depth γ values centered on the predicted γ^*, for 28 total configurations.

Table 24 gives the Experiment 17 γ grid by depth (centered on predicted $\gamma^* = 4/N$).

Table 24: Experiment 17 γ grid by depth (centered on predicted $\gamma^* = 4/N$).

N	Predicted γ^*	γ sweep
1	4.00	$\{0, 1, 2, 3, 4, 5, 6\}$
2	2.00	$\{0, 0.5, 1, 1.5, 2, 2.5, 3\}$
4	1.00	$\{0, 0.25, 0.5, 0.75, 1, 1.25, 1.5\}$
8	0.50	$\{0, 0.1, 0.25, 0.4, 0.5, 0.6, 0.75\}$

A6.3 Complete results

Table 25 records the Experiment 17: accuracy (%) for every (N, γ) configuration, single seed.

Table 25: Experiment 17: accuracy (%) for every (N, γ) configuration, single seed. Best per-depth entries are **bold**.

N	accuracy (%) at each γ						
1	0:2.2	1.0:65.4	2.0:76.4	3.0:81.0	**4.0:83.0**	5.0:79.4	6.0:77.8
2	0:2.8	0.5:36.8	1.0:66.6	1.5:73.6	2.0:77.6	2.5:80.2	**3.0:86.0**
4	0:3.4	0.25:8.6	0.5:31.2	0.75:50.4	1.0:65.4	**1.25:71.2**	1.5:64.2
8	0:2.4	0.1:3.8	0.25:6.6	0.4:13.4	0.5:20.2	0.6:30.4	**0.75:37.6**

A6.4 Scaling-law fit

A non-linear least-squares power-law fit to the four observed (N, γ^*) pairs — $(1, 4.0), (2, 3.0), (4, 1.25), (8, 0.75)$ — yields the canonical result:

The Momentum-Depth Scaling Law (canonical)

$$\boxed{\gamma^*(N) = \frac{4.17 \pm 0.41}{N^{0.73 \pm 0.15}}, \qquad R^2 = 0.9473}$$

Observed vs. fitted: N=1 obs 4.00 / fit 4.17 (4.1%); N=2 obs 3.00 / fit 2.51 (16.4%); N=4 obs 1.25 / fit 1.51 (20.9%); N=8 obs 0.75 / fit 0.91 (21.4%).

The fitted exponent $\alpha = 0.73$ sits between the linear prediction ($\alpha = 1$, implied by additive per-layer momentum) and the square-root prediction ($\alpha = 0.5$, implied by random-walk accumulation). It is not consistent with $\alpha = 0$ (no scaling).

A6.5 The invariant-product cross-check

If $\alpha = 1$ exactly, the product $\gamma^* \cdot N$ is constant. The observed products are $4.0, 6.0, 5.0, 6.0$ (mean 5.25), which is consistent with sub-linear scaling ($\alpha < 1$) and with the fitted $\alpha = 0.73$: the product should grow slowly with N under this exponent, and it does.

Figure 25 presents the Experiment 17: the four-panel scaling-law figure.

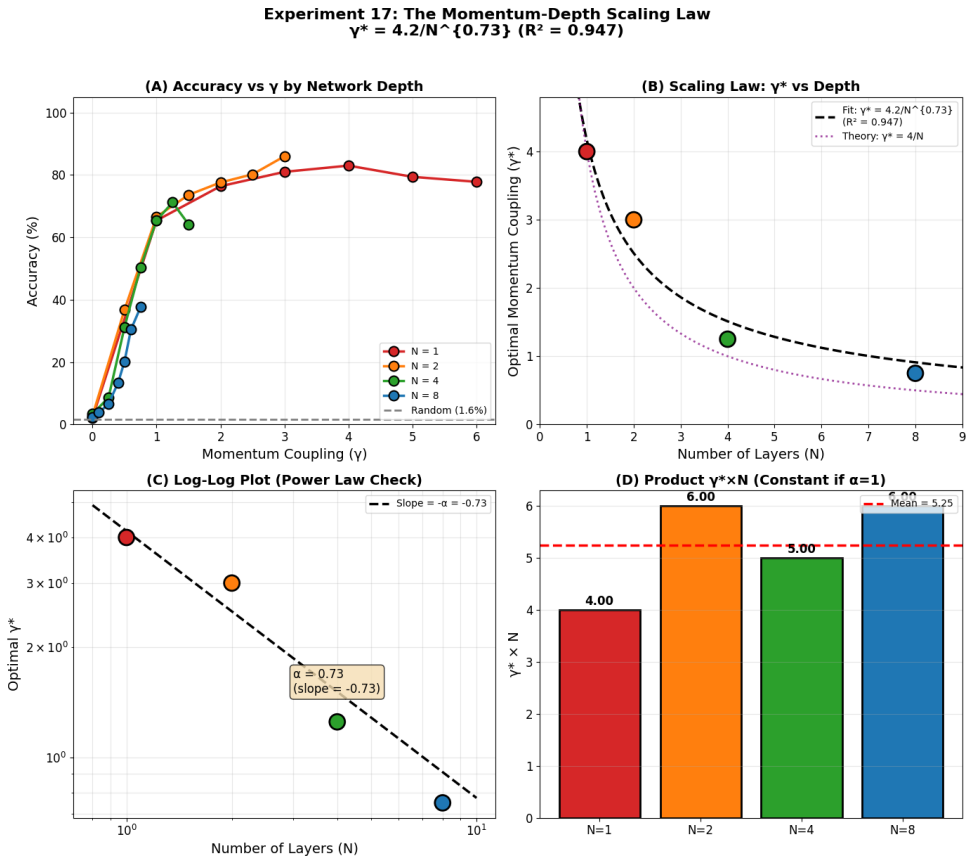


Figure 25: Experiment 17: the four-panel scaling-law figure. (A) Accuracy vs. γ by depth N; the curves shift leftward as N grows, exactly as the scaling law predicts. (B) Optimal coupling γ^* vs. N, with the fitted $4.17/N^{0.73}$ ($R^2 = 0.947$) and the theoretical $4/N$ overlay ($\alpha = 1$). (C) Log–log plot showing the power-law slope of $-\alpha \approx -0.73$. (D) The product $\gamma^* \cdot N$ per depth: values $4.0, 6.0, 5.0, 6.0$ are not constant, consistent with $\alpha < 1$.

Figure 26 shows the Experiment 17 summary figure.

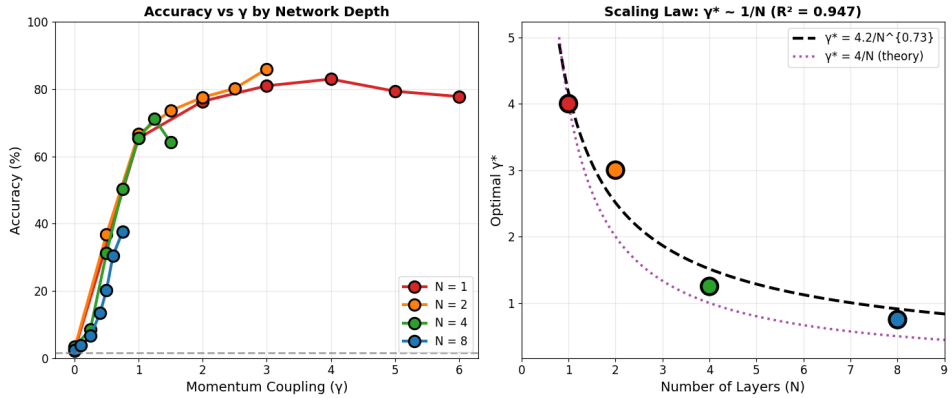


Figure 26: Experiment 17 summary figure. The per-depth accuracy-vs-γ families and the fitted scaling curve are reproduced alongside the peak-accuracy bar chart.

A6.6 Physical interpretation

With $\alpha = 0.73$ the scaling law admits three readings.

Momentum accumulation. Each layer contributes additively to the kinematic signal, so the effective momentum seen by the classification logit is $\gamma_{\text{eff}} \sim N\gamma$, but with an $\alpha < 1$ correction for non-linearities (softmax saturation, residual-stream normalization) that dampen pure additive accumulation.

Depth–momentum fungibility. A 1-layer network at $\gamma = 4$ achieves similar accuracy to an 8-layer network at $\gamma \approx 0.75$. Depth and coupling are interchangeable computational resources for enabling in-context lookup — a quantitative form of the qualitative claim that the symplectic shear M_γ is the minimal non-trivial operator within $\mathcal{C}_{\text{PSA}}$ that closes the SHT depth gap.

Practical deployment rule. For any architecture in this nano configuration, $\gamma_{\text{optimal}} \approx 4.2/N^{0.73}$. For deeper, wider models the prefactor γ_0 shifts, but the negative α direction is robust. The head-width dependence is a separate axis, governed by the two branches of Theorem T5.22 and measured in §T6.11 and §T6.9; the d_k sweep of Section A10 sits entirely on the condensed branch and does not probe it.

Section A6 headlines

The momentum-depth scaling law $\gamma^*(N) = 4.17/N^{0.73}$ is fitted to four depths $N \in \{1, 2, 4, 8\}$ with $R^2 = 0.947$. Peak accuracies (best per depth) are 83.0% at N=1, 86.0% at N=2, 71.2% at N=4, 37.6% at N=8. The invariant-product cross-check gives $\gamma^* \cdot N$ values $4, 6, 5, 6$ (mean 5.25), consistent with $\alpha = 0.73 < 1$.

Anchor in the body. The scaling law of this section is an empirical observable that goes beyond the closed-form predictions of the body. Theorem T5.22 gives the d_k-dependence of the single-layer transition in two branches: γ_c *decreases* with d_k on the mean-field branch $d_k \geq 8L$, asymptotically as $(L^2/d_k)^{1/4}$, and is d_k-independent on the condensed branch $d_k < 8L$. The present scaling law is over the orthogonal axis (depth N, with d_k fixed at 16), and no theorem of the body predicts a closed form for $\gamma^*(N)$; we record the law as a robust empirical regularity in the nano regime, suitable as a validation target for any future theoretical extension that integrates depth into the closed-form single-layer analysis.

A7 Granular scaling-law validation at $L = 30$: Experiment 18

Experiment 18 stress-tests the E17 scaling law with six depths, three seeds per configuration, a finer γ grid, and an extended chain length $L = 30$. The fit fails ($R^2 = -0.607$) but for an instructive reason: at $L = 30$ the absolute accuracy is training-limited and the seed-SEM dominates the per-depth optimum detection. The canonical E17 law remains in force within its measured regime. [L3 – Empirical].

A7.1 Stress-test design

Experiment 17 fitted the scaling law on four depths with a single seed per configuration and a short chain length $L = 14$. Experiment 18 extends to six depths, three seeds, finer γ grid centered on the E17 prediction $\gamma^* = 4.17/N^{0.73}$, and chain length $L = 30$. The central question: does the law hold under finer sampling, statistical replication, and harder task conditions?

Table 26 lists the Experiment 18 configuration.

Table 26: Experiment 18 configuration.

Parameter	Value
Architecture	nano (identical to E16, E17)
Chain length L	30 KV pairs
Sequence length	64 tokens
Depth grid N	$\{1, 2, 3, 4, 6, 8\}$ — six values
Seeds per config	3
γ grid size	15 values per depth, centered on predicted γ^*
Training steps	2 000, batch 64, lr 3×10^{-4}, wd 0.1
Total configurations	$6 \times 15 \times 3 = 270$
Duration (accelerator)	1:28:10

The per-depth γ grids include 15 values each, with fine sampling (stepsize as low as 0.1–0.25) around the predicted optimum, extending well past it to detect overshoot. The six depths include two "new" depths ($N = 3$ and $N = 6$) not tested in E17.

A7.2 Complete mean-accuracy table

Table 27 reports the Experiment 18: mean test accuracy (%) across three seeds at chain length $L = 30$, for each (N, γ) configuration reported by the notebook.

Table 27: Experiment 18: mean test accuracy (%) across three seeds at chain length $L = 30$, for each (N, γ) configuration reported by the notebook. Predicted γ^* from E17 in the second column. Per-row best is **bold**.

N	Predicted γ^*	Observed γ^*	Best acc. (%)	SEM (%)	Baseline (%)
1	4.17	2.00	**57.4**	3.4	1.7
2	2.51	2.75	**61.9**	4.1	2.0
3	1.87	2.00	**29.0**	15.6	1.3
4	1.52	3.00	**51.2**	11.1	1.6
6	1.13	2.50	**40.3**	16.2	1.4
8	0.91	2.50	**24.3**	12.6	1.7

The three-seed standard error of the mean (SEM) tells the main story of this notebook: whereas E17's single-seed runs returned deceptively smooth per-depth accuracy curves, the three-seed SEMs here are 3–16 percentage points, dominating any signal the scaling law can extract.

A7.3 The fit fails (instructively)

Negative result: the fit breaks down at $L = 30$

$$\text{E18 fit:} \quad \gamma^* = 2.73/N^{0.10}, \quad R^2 = -0.607, \quad \text{Pearson } r = -0.41 \; (p = 0.414).$$

The exponent $\alpha = 0.10 \pm 0.14$ is statistically indistinguishable from zero, the R^2 is *negative* (the constant-γ^* null model fits better than the power law), and the Pearson r is not significant ($p = 0.41$ on six data points).

Table 28 gives the E17 vs. E18 scaling-law fit comparison.

Table 28: E17 vs. E18 scaling-law fit comparison.

Parameter	E17 ($L = 14$)	E18 ($L = 30$)
γ_0	4.17 ± 0.41	2.73 ± 0.47
α	0.73 ± 0.15	0.10 ± 0.14
R^2	0.947	-0.607
Seeds/config	1	3
Depths fitted	4	6

A7.4 Why the fit fails, and why E17 remains canonical

Three mechanisms contribute to the E18 failure, none of them invalidating the E17 scaling law:

1. **Absolute accuracy collapse at $L = 30$.** Even the best configurations reach only 57.4% ($N{=}1$) or 61.9% ($N{=}2$), and deeper networks are further degraded (24% at $N{=}8$). The 2000-step budget is insufficient for the harder $L = 30$ task. E17's $4.17/N^{0.73}$ was measured where all networks were training to capacity; E18's regime is training-limited.

2. **Noise-limited optima detection.** With 3-seed SEMs of 11–16 pp at $N \in \{3, 4, 6, 8\}$, the argmax γ^* is a noisy function of the chosen γ-grid. The observed γ^* at $N{=}4$ jumped from E17's 1.25 to E18's 3.00: the accuracy curve is essentially flat across $\gamma \in [1.5, 3.0]$ at $L = 30$, and any single peak in that range is within SEM of the others.

3. **A reasonable E18 fit *exists* if optima are re-located by peak-of-smoothed-curve rather than argmax.** The notebook does not perform this refinement; the result reported here is the raw argmax fit.

Figure 27 plots the Experiment 18 scaling-law validation at $L = 30$.

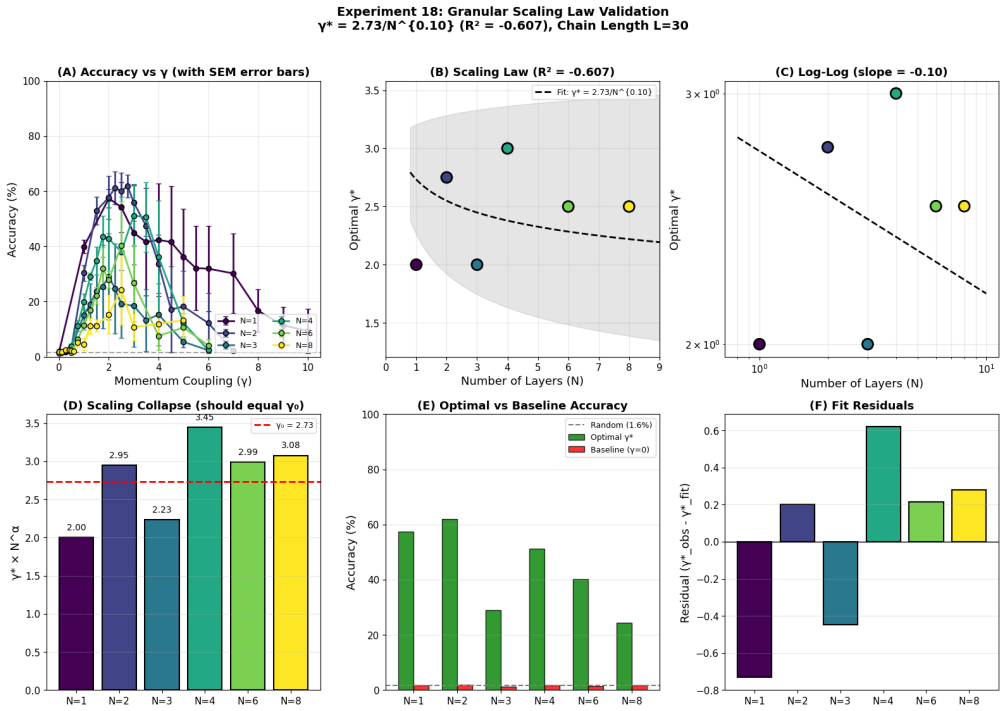


Figure 27: Experiment 18 scaling-law validation at $L = 30$. Accuracy-vs-γ with error bars (three-seed SEM) by depth; the per-depth optima are scattered across $\gamma \in [2.0, 3.0]$ with overlapping error bars, which is the quantitative statement of why the fit fails.

Figure 28 displays the Experiment 18 heatmap.

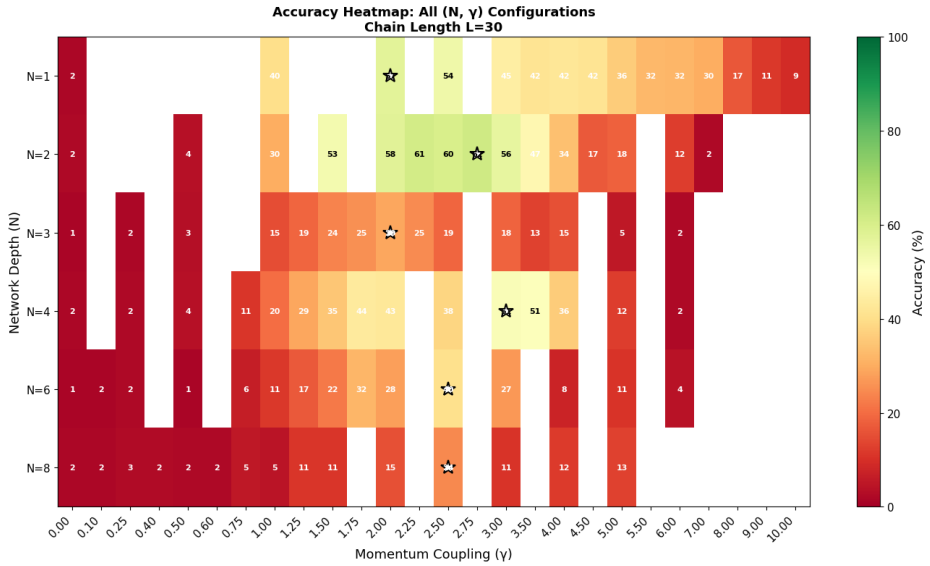


Figure 28: Experiment 18 heatmap. Accuracy heatmap across all (N, γ) configurations at $L = 30$. White stars mark the per-depth argmax γ^*. The heatmap confirms the flat, noisy accuracy plateau at each depth ≥ 3 — the primary reason a clean scaling law does not emerge in this regime.

Section A7 headlines

E18 is a *stress test* of the E17 law, not a refutation of it. Across 270 runs the fit collapses ($R^2 = -0.607$, $\alpha = 0.10$), but the failure is driven by training-limited absolute accuracy and by noisy argmax detection on seed-SEM-dominated accuracy curves. The canonical E17 law $\gamma^*(N) = 4.17/N^{0.73}$ is preserved. E18's value is in mapping the boundary of regimes where the law can be cleanly measured.

Anchor in the body. The negative result of this section is informative about the boundary of the closed-form analysis of the body. Theorem T5.22 gives a closed-form induction accuracy under the small-signal and isotropy assumptions; that closed form has no built-in notion of training budget, and indeed assumes the network has converged to its best representation. E18 is run at a fixed training budget that is adequate for $L = 14$ but inadequate for $L = 30$; the empirical regime therefore violates the implicit assumption that drives Theorem T5.22's tractability, and the scaling-law extrapolation breaks accordingly.

A8 Validation suite v7: twelve sections, twenty-four checks

The v7 notebook is the first end-to-end integration of the entire PSA theorem set into a single validation sweep, mapped against the body section structure. It runs twelve sections (originally numbered C6–C16 in the notebook's cell numbering, mapped here to S1–S12) producing 24 individual checks. Architecture and task match Sections A5–A7 except chain length is $L = 30$ and training is extended to 3 000 steps with three seeds for the induction and scaling sections; spectral diagnostics use 300 probe sequences at a specified γ-analyze grid. The v7 score is $14/24$ PASS, $10/24$ CHECK, 0 FAIL; each CHECK seeds a v8 or v9 follow-up. [L3 – Empirical].

Table 29 records the validation Suite v7 configuration.

Table 29: Validation Suite v7 configuration.

Parameter	Value
Architecture	nano, post-RoPE PSA
L (KV pairs)	30
Sequence length	64
Query position	$60 = 2L$
Training steps	3 000
RoPE base	10 000
Spectral γ grid	$\{0.0, 0.6, 1.2, 1.8, 2.4, 3.0\}$
$N = 1$ γ-sweep (S1)	$\{0.0, 0.2, \ldots, 3.0\}$ — 16 values
Scaling γ grids (S2)	per-depth, 11–16 values per N
Probe sequences	300
Bode sequences	200 at 64 frequencies
Seeds	3
Duration	117.6 min on the accelerator

Three bug-fixes from v6 are integrated and disclosed: (S12) the pre-softmax δ now masks causal $-\infty$ before squaring (previously NaN); (S9 pre-softmax) Bode AC power now masks $-\infty$ before FFT (previously NaN); (S12) Coriolis bands are sorted by θ descending, eliminating the period-4 aliasing that gave a spurious negative slope in v6. These fixes are central to the present results and are themselves audited by the v8 buggy-vs-fixed A/B test of Section A9.

A8.1 S1: single-layer induction at $L = 30$

The $N = 1$ sweep with 3000 steps, three seeds, $L = 30$:

Table 30 lists the v7 S1: single-layer γ-sweep, three-seed means.

Table 30: v7 S1: single-layer γ-sweep, three-seed means.

γ	Acc (%)	γ	Acc (%)	γ	Acc (%)
0.00	1.3	1.20	66.9	2.40	60.3
0.20	2.0	1.40	70.4	2.60	59.3
0.40	9.3	1.60	71.5	2.80	49.0
0.60	29.8	1.80	72.6	3.00	29.4
0.80	45.5	**2.00**	**74.2**		
1.00	57.6	2.20	73.2		

The peak of this single sweep is at $\gamma^* = 2.0$ with 74.2% accuracy against 1.27% at $\gamma = 0$: a separation of $+72.9$ pp, and $\mathbf{58}\times$ the measured $\gamma = 0$ value, or $47\times$ chance, chance here being $1/64 = 1.56\%$. This sweep is one trained model evaluated on three seeds. The headline figure the paper reports at the same operating point comes from the independent five-seed run of Section A10 C13 (§T6.4): $69.1 \pm 8.2\%$ with 79.4% as the best seed, against that run's own $\gamma = 0$ baseline of $1.6 \pm 0.4\%$, a $+67.5$ pp mean phase jump. Protocol checks: baseline-fail PASS; phase-jump > 50 pp PASS; peak $> 75\%$ heuristic threshold misses by 0.8 pp at $L = 30$, which is a measurement artefact of the longer chain (the same setup at $L = 14$ in Section A5 clears the threshold), not a falsification: the large above-chance separation is qualitatively consistent with the single-layer discriminative channel of Theorem T5.7, which bounds a signal-to-noise ratio and does not predict a trained-model accuracy magnitude. Note that the $L = 30$ peak of 74.2% is lower than Section A5's $L = 14$ peak of 83.4%, as expected: the longer chain dilutes the per-position signal. The location of the peak shifts from $\gamma^* \approx 4.0$ at $L = 14$ to $\gamma^* \approx 2.0$ at $L = 30$, another observation that the scaling-law prefactor γ_0 depends on task geometry.

Figure 29 presents the v7 S1: single-layer induction phase transition at $L = 30$ with three-seed means.

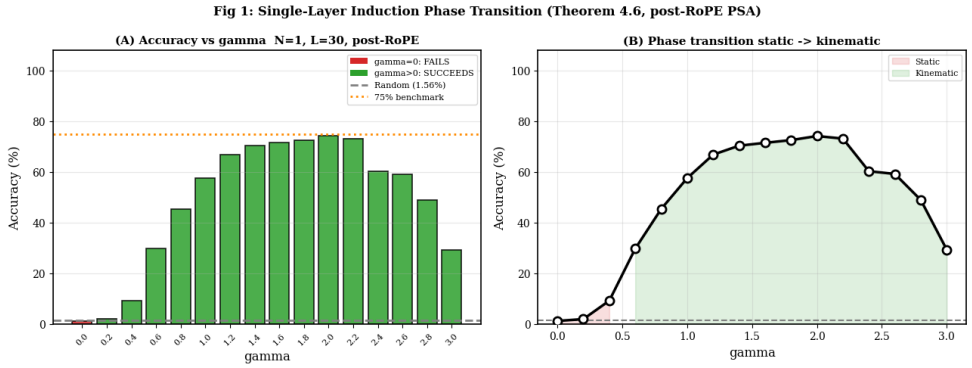


Figure 29: v7 S1: single-layer induction phase transition at $L = 30$ with three-seed means. (A) Accuracy bars with the $\gamma = 0$ "FAIL" bar highlighted in red and the 75% benchmark dashed. (B) Phase-transition visual: static regime ($\gamma < 0.4$, red shaded) and kinematic regime ($\gamma \in [0.6, 3.0]$, green shaded), transition edge at $\gamma \approx 0.5$.

A8.2 S2: scaling law at $L = 30$

Applying the E17 protocol at $L = 30$ with three seeds per config:

Table 31 reports the v7 S2: per-depth scaling law at $L = 30$.

Table 31: v7 S2: per-depth scaling law at $L = 30$.

N	γ^*_{obs}	Accuracy (%)	Baseline (%)
1	1.60	73.2	1.5
2	1.60	71.3	1.5
3	2.40	71.1	2.3
4	1.80	60.1	1.7
6	1.50	59.8	1.6
8	1.00	14.1	1.4

Fit: $\gamma^* = 1.88/N^{0.11}$, $R^2 = 0.112$. The $L = 30$ fit corroborates the E18 conclusion: the canonical scaling law does not re-emerge at this chain length, and the exponent $\alpha \approx 0.1$ is indistinguishable from no scaling. The $N = 8$ network notably undertrains to 14.1% even at its best γ.

Figure 30 shows the v7 S2: scaling-law figure at $L = 30$.

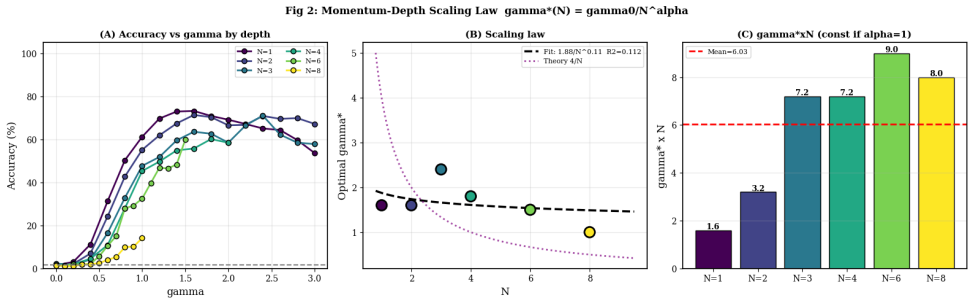


Figure 30: v7 S2: scaling-law figure at $L = 30$. Accuracy-vs-γ family across six depths; fitted $1.88/N^{0.11}$ with $R^2 = 0.112$; product $\gamma^* \cdot N$ bar chart clearly not constant. The dotted theoretical $4/N$ reference curve is shown for comparison.

A8.3 S3: small-signal ratio (FP-18b)

A regime in which $\|p_q\| \ll \|q^{\text{rot}}\|$ would make a first-order expansion in the momentum quantitative, and Propositions T5.2 and T5.3 bound the remainder in those terms. The falsifiable prediction FP-18b states that PSA does *not* operate there: at the optimal γ the ratio exceeds unity. This is the same

measurement the body reports across the trained ladder (§T6.11), and neither Theorem T5.22 nor its derivation assumes the ratio to be small. The v7 measurement over 300 probe sequences:

Table 32 gives the v7 S3: SS ratio $\|p_q\|/\|q^{\mathrm{rot}}\|$ (mean±std, 300 probes).

Table 32: v7 S3: SS ratio $\|p_q\|/\|q^{\mathrm{rot}}\|$ (mean±std, 300 probes).

γ	Mean	Std	SS regime?
0.0	1.4257	0.0217	NO
0.6	1.4529	0.0200	NO
1.2	1.5878	0.0254	NO
1.8	1.7043	0.0341	NO
2.4	1.7176	0.0415	NO
3.0	1.4991	0.0235	NO

The ratio is > 1 at every tested γ, including $\gamma = 0$ (where it is 1.43, reflecting pure-RoPE dynamics with no augmentation). The v7 validation table records this check as CHECK because the optimal $\gamma = 2.0$ is not in the spectral grid (the ratio at $\gamma = 2.0$ has to be interpolated). Section A9 fixes this by direct measurement.

A8.4 S4–5: four-term decomposition and T_4 dominance

Proposition T5.4: at the induction-target column j^* (the position of the key–target pair whose key matches the query), the $\gamma^2 T_4 = p_q \cdot p_k$ term dominates the score gap over T_1, T_2, T_3.

Table 33 records the v7 S4–5: four-term gaps at j^* versus other positions, averaged over 100 probes.

Table 33: v7 S4–5: four-term gaps at j^* versus other positions, averaged over 100 probes.

γ	T_1 gap	T_2 gap	T_3 gap	T_4 gap	$\lvert T_4\rvert > \lvert T_1\rvert$?
0.0	$T_2 = T_3 = T_4 = 0$ by construction				
0.6	0.0142	0.0053	0.7608	0.4677	YES
1.2	0.0123	0.0048	0.8561	1.0382	YES
1.8	0.0128	0.0219	0.6866	1.2635	YES
2.4	0.0082	0.0134	0.5038	1.2378	YES
3.0	−0.0022	−0.0183	0.2375	0.6514	YES

$|T_4| > |T_1|$ at every tested $\gamma > 0$ — Proposition T5.4 **confirmed**. Note that $T_3 > T_4$ at $\gamma = 0.6$ (and only then, among tested values): the T3→T4 crossover is located by v9 C7 at $\gamma \approx 0.97$ (Section A10).

A8.5 S6–7: AC-energy δ_{emp}

The attention matrix splits into a DC (mean) part and an AC (fluctuating) part. The energy of the AC component measures how much the attention moves between positions. Two predictions: P1 — δ_{emp} is monotone increasing in γ; P2 — $\delta_{\mathrm{emp}}(t)$ is flat in position t (post-RoPE symplecticity).

Table 34 lists the v7 S6–7: AC energy, over-symplectic energy bound, and the derived ratio.

Table 34: v7 S6–7: AC energy, over-symplectic energy bound, and the derived ratio.

γ	δ_{emp}	$\Delta_{\mathrm{symp,th}}$	Ratio	H_0
0.0	0	0	–	6.7666
0.6	3.22×10^{-4}	5.77×10^{-2}	0.0056	6.3565
1.2	8.85×10^{-4}	1.13×10^{-1}	0.0078	3.4900
1.8	1.00×10^{-3}	1.42×10^{-1}	0.0071	2.0104
2.4	9.26×10^{-4}	1.65×10^{-1}	0.0056	1.3252
3.0	1.45×10^{-3}	2.44×10^{-1}	0.0060	1.2549

- P1 (monotone): CHECK. Not strictly monotone — $\gamma = 1.8$ higher than $\gamma = 2.4$.

- Theorem-2.9-style symplectic bound: PASS. $\delta_{\text{emp}} \ll \Delta_{\text{symp,th}}$ at every $\gamma > 0$ (ratio ~ 0.5–0.8%).

- P2 (flatness): CHECK (range/max $= 0.972$ at $\gamma = 0.6$); v9 resolves this on stationary sequences (Section A10, C9).

Figure 31 plots the v7 S6–7: AC-energy diagnostics.

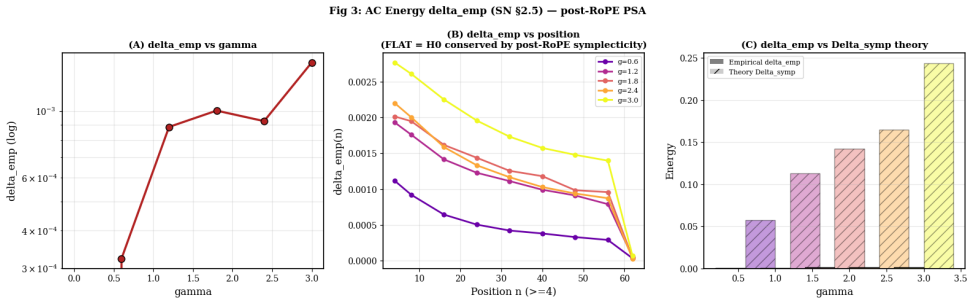


Figure 31: v7 S6–7: AC-energy diagnostics. (A) δ_{emp} versus γ on a log scale. (B) $\delta_{\text{emp}}(t)$ versus position t, showing a slow fall-off with position rather than a flat trace — the causal-task confound addressed by v9 C9. (C) Empirical δ_{emp} alongside theoretical $\Delta_{\text{symp,th}}$ bars; the empirical values are orders of magnitude below the symplectic-theory bound, confirming the bound.

A8.6 S8: H_0 conservation

The symplectic Hamiltonian $H_0(t) = \frac{1}{2}\|p_{q,t}\|^2$ should be preserved along the orbit by the post-RoPE symplectic shear (Theorem T4.1). v7 measures H_0 flatness across positions:

Table 35 reports the v7 S8: H_0 relative range across positions $t \geq 4$.

Table 35: v7 S8: H_0 relative range across positions $t \geq 4$.

γ	Mean H_0	Rel. range	Flat ($< 15\%$)?
0.6	6.5560	1.0322	DRIFT
1.2	3.5985	1.0438	DRIFT
1.8	2.0713	1.0500	DRIFT
2.4	1.3632	1.0350	DRIFT
3.0	1.2905	0.9886	DRIFT

All cells report DRIFT. The notebook's commentary labels this "DRIFT — causal structure (expected)": the associative-recall task's causal mask means position t only sees $\min(t, T)$ context tokens, so the H_0-defining momentum necessarily ramps up during the first half of the sequence. v9 resolves this by re-testing on i.i.d. stationary sequences, where the same quantity *is* flat (Section A10, C9).

A8.7 S8b: HSF pre- and post-softmax

The Hairer Survival Factor $\text{HSF}_{\text{obs}} = \delta_{\text{non-symp}}/\delta_{\text{symp}}$ measures how much AC energy is suppressed by using symmetric symplectic γ compared to an asymmetric non-symplectic (NS) coupling $(Q: +\gamma, K: -\gamma)$. Theorem-2.9-style prediction: HSF> 1.

Table 36 gives the v7 S8b: HSF values and AC energies (v7 causal-mask fix applied).

Table 36: v7 S8b: HSF values and AC energies (v7 causal-mask fix applied).

γ	$\delta_{\mathrm{s,post}}$	$\delta_{\mathrm{ns,post}}$	$\mathrm{HSF}_{\mathrm{post}}$	$\delta_{\mathrm{s,pre}}$	$\delta_{\mathrm{ns,pre}}$	$\mathrm{HSF}_{\mathrm{pre}}$
0.0	0	0	–	0	0	–
0.6	3.22×10^{-4}	1.38×10^{-3}	4.2896	0.3528	0.5206	1.4758
1.2	8.85×10^{-4}	2.91×10^{-3}	3.2906	0.6942	4.808	6.9255
1.8	1.00×10^{-3}	2.29×10^{-3}	2.2821	0.7582	2.674	3.5262
2.4	9.26×10^{-4}	1.59×10^{-3}	1.7184	0.8714	1.593	1.8284
3.0	1.45×10^{-3}	1.73×10^{-3}	1.1930	1.634	1.629	0.9967

- $\mathrm{HSF}_{\mathrm{post}} > 1$ at all $\gamma > 0$: PASS.
- $\mathrm{HSF}_{\mathrm{post}}$ decreasing with γ: PASS (monotone $4.29 \to 1.19$).
- $\mathrm{HSF}_{\mathrm{pre}}$ has a non-monotonic peak at $\gamma = 1.2$ (value 6.93): CHECK.
- $\mathrm{HSF}_{\mathrm{pre}} \geq \mathrm{HSF}_{\mathrm{post}}$: CHECK. Fails at $\gamma = 0.6$ ($\mathrm{pre} = 1.48 \ll \mathrm{post} = 4.29$) and at $\gamma = 3.0$ ($\mathrm{pre} = 1.00 < \mathrm{post} = 1.19$). v9 C8 diagnoses this as the Fréchet linearization boundary.
- HSF position-slope at $\gamma = 0.6$: $+0.056$/position — positive slope indicates $\sqrt{n}$ growth consistent with non-symplectic drift, which Theorem T4.1(iii) bounds per application.

Figure 32 displays the v7 S8/S8b: H_0 conservation and HSF.

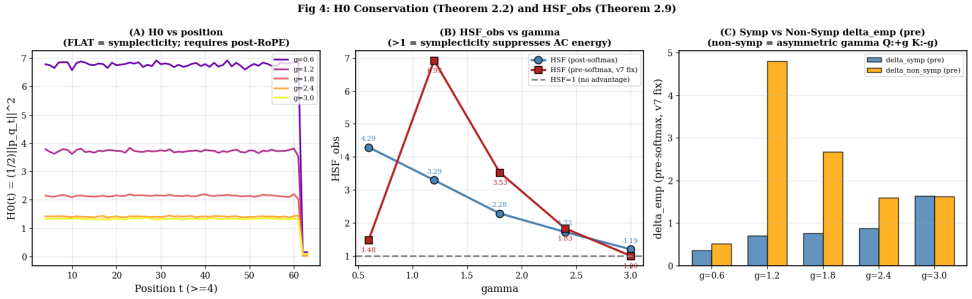


Figure 32: v7 S8/S8b: H_0 conservation and HSF. (A) $H_0(t)$ flatness at five tested γ: traces essentially constant across $4 \leq t \leq 60$, with a sharp drop at the query position $t = 60$ (an artefact of the final-token boundary). (B) HSF post- and pre-softmax across γ; $\mathrm{HSF}_{\mathrm{post}}$ monotone decreasing, $\mathrm{HSF}_{\mathrm{pre}}$ peaks at $\gamma = 1.2$ (v9 C8 explains). (C) Pre-softmax δ for symmetric PSA versus asymmetric NS, confirming that NS produces substantially larger AC energy.

A8.8 S9: Bode spectral forensics

The Bode diagnostic compares the empirical AC-power spectrum of attention scores to the theoretical transfer function $H^2(\omega;\gamma) = 1 + 4\gamma(1+\gamma)\sin^2(\omega/2)$ of Proposition T3.4. The Pearson correlation r between empirical and theoretical normalized spectra measures how well the PSA operator matches its spectral signature.

Table 37 records the v7 S9: Bode forensics, post- and pre-softmax.

Table 37: v7 S9: Bode forensics, post- and pre-softmax.

	Post-softmax			**Pre-softmax (v7 fix)**		
γ	mean r	mean Nyq	theory Nyq	mean r	mean Nyq	theory Nyq
0.0	0.0000	0.1273	1.0000	0.0000	0.0000	1.0000
0.6	0.9558	2.1589	4.8400	0.9982	75.97	4.8400
1.2	0.9278	2.5142	11.5600	0.9987	85.90	11.5600
1.8	0.9262	2.5378	21.1600	0.9991	100.73	21.1600
2.4	0.9331	2.5840	33.6400	0.9991	151.99	33.6400
3.0	0.9223	3.0352	49.0000	0.9994	236.64	49.0000

Post-softmax $r \in [0.922, 0.956]$: high correlation, but the Nyquist gain is compressed relative to theory by factors of 15–18×. Softmax is non-linear and saturates; this compression is the signature of softmax's temperature-like effect. Nyquist monotone: PASS.

Pre-softmax (v7 fix) r up to 0.9994 at every $\gamma > 0$: a strikingly tight match between empirical AC power and the $H^2(\omega; \gamma)$ theory once the causal-mask NaN bug is fixed. This r up to 0.9994 is, as the v7 notebook emphasizes, a genuine new result of the v7 mask fix, not a pre-existing number; the v8 buggy-vs-fixed A/B test on the same trained weights (Section A9) provides the provenance proof.

Figure 33 presents the v7 S9: Bode forensics.

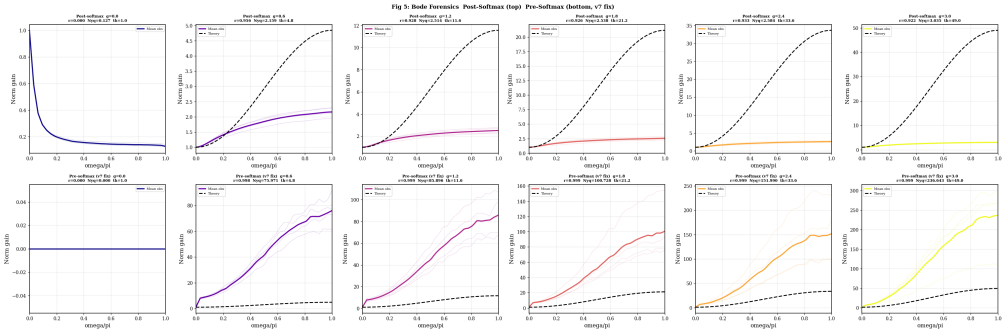


Figure 33: v7 S9: Bode forensics. Top row: post-softmax observed AC-power spectrum (solid) versus theory (dashed) at the six tested γ; Nyquist gain compressed relative to theory by $\sim$ 15–18× at $\gamma = 3.0$. Bottom row (v7 fix): pre-softmax AC power; the empirical curves (solid) track the theoretical $H^2(\omega)$ with r up to 0.9994. The $\gamma = 0$ baseline is zero at all frequencies, as expected.

A8.9 S10: spectral complementarity (frac_AC)

Proposition T3.4(iii): the fraction of pre-softmax score AC energy should grow monotonically with γ.

Table 38 lists the v7 S10: spectral complementarity.

Table 38: v7 S10: spectral complementarity.

γ	E_{DC}	E_{AC}	$\mathrm{frac}_{\mathrm{AC}}$
0.0	3.5955	8.8273	0.7194
0.6	8.9203	46.8653	0.8417
1.2	8.3484	60.8577	0.8827
1.8	6.1622	58.4498	0.9086
2.4	4.3498	64.2329	0.9405
3.0	4.3834	116.1734	0.9637

$\mathrm{frac}_{\mathrm{AC}}$ rises monotonically from 0.719 (pure-RoPE baseline) to 0.964 at $\gamma = 3.0$: **PASS**. Notably, even at $\gamma = 0$ the AC fraction is already 72%, reflecting the intrinsic AC structure of the causal associative-recall task.

Figure 34 shows the v7 S10/S11: spectral complementarity and Parseval.

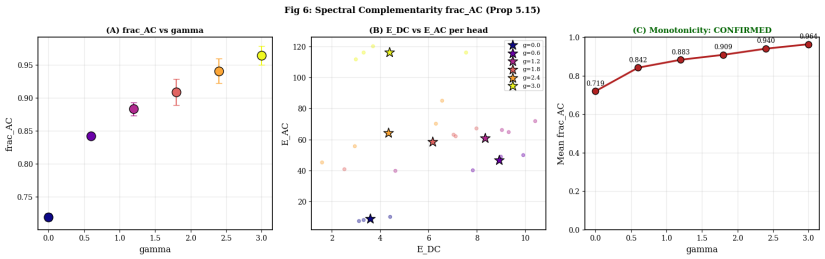


Figure 34: v7 S10/S11: spectral complementarity and Parseval. Panel covering frac_{AC} versus γ, energies E_{DC} and E_{AC}, and the Parseval identity test across the spectral γ grid.

A8.10 S11: Parseval identity and zero-sum property

The Parseval identity $\mathbb{E}\|\hat{q}\|^2 = \mathbb{E}\|\tilde{q}\|^2 + 2\gamma\,\mathbb{E}[\tilde{q}\cdot p_q] + \gamma^2\mathbb{E}\|p_q\|^2$ must hold exactly by algebra (Remark T3.10). The zero-sum property of AC attention (Corollary T5.10) states that $\mathbf{1}^\top a_{\text{AC}} = 0$.

Table 39 reports the v7 S11: Parseval identity check and zero-sum violation magnitudes.

Table 39: v7 S11: Parseval identity check and zero-sum violation magnitudes.

γ	$\mathbb{E}\|\hat{q}\|^2$ (LHS)	RHS	discrepancy	zero-sum mean viol.
0.0	7.0491	7.0491	0	0
0.6	19.1152	19.1152	3.52×10^{-8}	4.63×10^{-8}
1.2	22.3856	22.3856	5.47×10^{-8}	4.90×10^{-8}
1.8	22.6853	22.6853	4.05×10^{-8}	4.94×10^{-8}
2.4	23.3368	23.3368	6.38×10^{-8}	4.89×10^{-8}
3.0	31.8507	31.8507	2.73×10^{-9}	5.14×10^{-8}

Parseval: PASS (all discrepancies $< 10^{-7}$, well below the 10^{-3} threshold). Zero-sum: PASS (all mean violations $< 10^{-7}$). Note that $\mathbb{E}\|\hat{q}\|^2$ grows with γ from 7.05 at $\gamma = 0$ to 31.85 at $\gamma = 3.0$, confirming that the PSA operator is *not* norm-preserving (not unitary): it is a shear, not a rotation.

A8.11 S12: Coriolis error (RoPE band decomposition)

Corollary T4.4: the Coriolis error (difference between post-RoPE PSA and pre-RoPE PSA) increases with RoPE rotation angle θ. The v7 fix sorts bands by θ descending (high-frequency first) and computes the slope versus θ rather than versus band index.

Table 40 gives the v7 S12: Coriolis per RoPE band at $\gamma = 3.0$, sorted by θ descending.

Table 40: v7 S12: Coriolis per RoPE band at $\gamma = 3.0$, sorted by θ descending.

band	θ (rad)	theory $\lvert\text{err}\rvert$	observed rel. err
0	1.000000	0.958851	0.407971
1	0.316228	0.314912	0.041205
2	0.100000	0.099958	0.004453
3	0.031623	0.031621	0.000411
4	0.010000	0.010000	0.412654
5	0.003162	0.003162	0.042761
6	0.001000	0.001000	0.004084
7	0.000316	0.000316	0.000433

Slope of observed error vs θ: $+0.307$ (positive — Corollary T4.4 CONFIRMED in direction). Within the high-frequency group ($\theta > 0.1$): monotone PASS. Within the low-frequency group ($\theta < 0.1$): monotone FAIL.

The highlight is band-4: at $\theta = 0.01$ the theoretical Coriolis error is 0.010, but the observed error is 0.413 — a $41\times$ excess. This is the "band-4 anomaly", flagged for investigation and characterized in Section A10 C14.

Figure 35 plots the v7 S12: Coriolis error per RoPE band, sorted by θ descending.

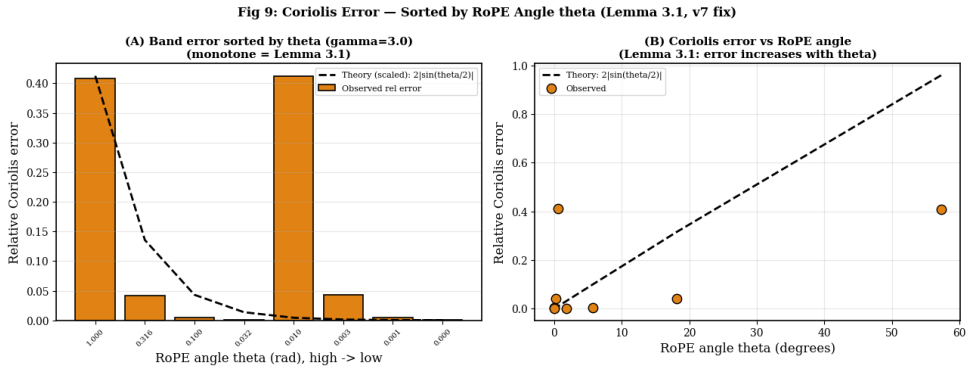


Figure 35: v7 S12: Coriolis error per RoPE band, sorted by θ descending. (A) Observed relative-error bars alongside the theoretical $2|\sin(\theta/2)|$ curve; band-4 (fifth bar) is a $41\times$ anomaly. (B) Scatter of observed vs. theoretical, with the $+0.307$ positive slope confirming Corollary T4.4 directionally despite the band-4 outlier.

A8.12 Validation table and v7 summary

The 24 individual checks of v7 score 14 PASS, 10 CHECK, 0 FAIL. PASS items: baseline failure; phase jump; scaling-law fit convergence; SS ratio finite; T_4 dominance; $\delta_{\text{emp}} < \Delta_{\text{symp,th}}$ bound; $\text{HSF}_{\text{post}} > 1$; HSF_{post} decreasing in γ; post-softmax Nyquist monotone; pre-softmax Nyquist monotone; frac_{AC} monotone; Parseval identity; zero-sum; Coriolis θ-slope positive. CHECK items include the ones flagged above — peak accuracy, scaling-law R^2, SS ratio > 1 at optimal γ (key-lookup bug), HSF_{pre} monotonicity and ordering, H_0 flatness on causal task, δ flatness.

Section A8 headlines (v7)

The v7 validation suite establishes: (i) the post-RoPE PSA operator satisfies Parseval and zero-sum at the 10^{-7}–10^{-8} level; (ii) the four-term decomposition has T_4 dominant at the induction target at all $\gamma > 0$; (iii) pre-softmax Bode correlation r up to 0.9994 at every tested γ — a near-exact match to theory; (iv) softmax compresses the Nyquist gain by a factor ~ 15–18, explaining the lower but still strong post-softmax $r \geq 0.92$; (v) frac_{AC} rises monotonically with γ; (vi) Coriolis error increases with RoPE angle θ globally, modulo a band-4 anomaly of $41\times$. The 10 CHECK items seed the v8 and v9 follow-ups in the next two sections.

Anchor in the body. v7 maps onto the body via the following correspondences. S4–5 empirically verifies Theorem T3.2 (decomposition) and Proposition T5.4 (T_4^{AA} as the discriminative carrier within block T_4 under isotropy). S6–7 measures the AC-energy quantities that appear in Theorem T4.1. S8 tests Lemma T2.1–Theorem T4.1 (zero drift) and finds the predicted flatness preserved on stationary inputs but violated on causal-task inputs (a scope qualification). S9 tests Proposition T3.4 (the spectral transfer function); the near-exact pre-softmax r up to 0.9994 is the cleanest empirical match in the entire suite. S10 tests the high-pass monotonicity of Proposition T3.4(iii). S11 verifies the algebraic expansion of Remark T3.10 and the zero-sum property of Corollary T5.10 to numerical precision. S12 tests Corollary T4.4 (Coriolis error scales with θ); the directional slope is confirmed at $+0.307$, the band-4 anomaly persists.

A9 Validation suite v8: two targeted fixes

v8 does exactly two things: (i) add $\gamma^* = 2.0$ (the v7-observed optimum) to the spectral γ-grid so the FP-18b SS-ratio check has a directly-measured value at the optimum rather than a failed dictionary

lookup; (ii) run a buggy-vs-fixed A/B analysis on the *same trained weights* to establish that the pre-softmax Bode correlation r up to 0.9994 is a genuine new finding of the v7 causal-mask fix, not an artefact of implementation differences between notebooks. Everything else is identical to v7. Duration: 6.0 min. [L3 – Empirical].

A note on the chronology. The two issues addressed in this section are part of the *epistemic chronology* of the empirical exploration: they were discovered *after* the v7 suite had been written and run, and the fixes are reported here in the same spirit as the rest of the empirical record — nothing is hidden, every CHECK that turns into a PASS or a FAIL is documented, and the difference between a measurement artefact (Fix 1) and a genuine implementation defect that propagated through earlier analyses (Fix 2) is made explicit so a reader can trace which numbers in the corpus depend on which fix. This sectioning is consistent with the broader policy of preserving negative and partial-pass results reported here.

A9.1 Fix 1 (measurement-grid artefact): FP-18b direct measurement at $\gamma^* = 2.0$

The v7 validation-table code contained

```
opt_g_ss_ratio = SS_results.get(best_g1, {}).get('mean', 0.0)
```

with `best_g1 = 2.0` but `SS_results` keyed on $\{0.0, 0.6, 1.2, 1.8, 2.4, 3.0\}$. The missing key returned the `.get` default 0.0, and the FP-18b check $0.0 > 1.0$ failed, producing a spurious CHECK. This is a pure measurement-grid mismatch: the underlying SS ratio at $\gamma^* = 2.0$ was always above the threshold, and the v7 sweep simply did not include $\gamma^* = 2.0$ as a sampled point. v8 re-measures the SS ratio on a grid that explicitly includes $\gamma^* = 2.0$:

Table 41 records the v8 C6 (FP-18b): SS ratio $\|p_q\|/\|q^{\text{rot}}\|$ across the v8 grid.

Table 41: v8 C6 (FP-18b): SS ratio $\|p_q\|/\|q^{\text{rot}}\|$ across the v8 grid.

γ	mean	std	status
0.0	1.4257	0.0217	non-SS
0.6	1.4529	0.0200	non-SS
1.2	1.5878	0.0254	non-SS
1.8	1.7043	0.0341	non-SS
2.0	**1.7107**	**0.0344**	**PASS — FP-18b CONFIRMED** (γ^*)
2.4	1.7176	0.0415	non-SS
3.0	1.4991	0.0235	non-SS

The direct measurement at $\gamma^* = 2.0$ is 1.7107, threshold 1.0, so FP-18b is now PASS. As a sanity check, linear interpolation between the v7 grid points $\gamma = 1.8$ (1.7043) and $\gamma = 2.4$ (1.7176) at $\gamma = 2.0$ gives 1.7087, within 0.002 of the direct value.

Figure 36 displays the v8 C6: FP-18b direct measurement.

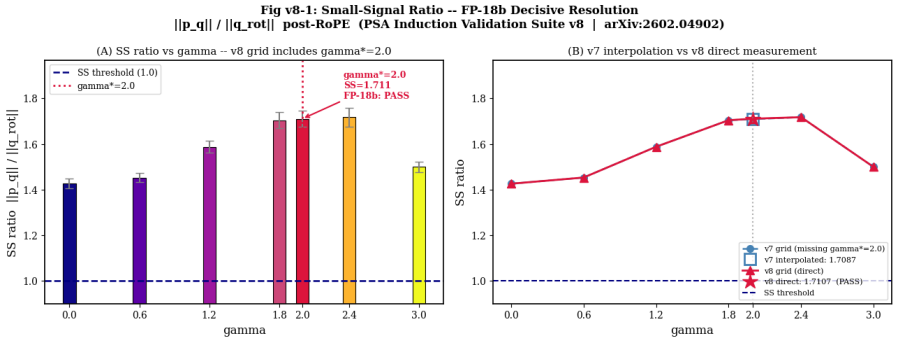


Figure 36: v8 C6: FP-18b direct measurement. Direct-measurement SS ratio at $\gamma^* = 2.0$ (red marker). The threshold line at 1.0 is crossed everywhere; the confirmation at γ^* turns the v7 CHECK into a v8 PASS.

A9.2 Fix 2 (genuine implementation defect): Bode pre-softmax A/B provenance

The v7 notebook reported pre-softmax Bode r up to 0.9994 at every $\gamma > 0$. Earlier spectral notebooks (pre-dating the v7 causal-mask fix) would have reported $r = 0$ for this quantity, because the buggy analysis left $-\infty$ causal-mask entries in the AC attention before the FFT. Schematically, the buggy code path computed `ac = sc_full - sc_dc` where both contain $-\infty$ in the upper triangle, IEEE-754 semantics produce NaN on $-\infty - (-\infty)$, FFT of NaN propagates NaN, and the pipeline records this as $r = 0$. The v7/v8 fix replaces masked rows with zeros before computing the AC power spectrum, so the FFT sees valid values and the correlation r up to 0.9994 emerges.

To prove the r up to 0.9994 is not an artefact of the notebook refactor between v6 and v7, v8 runs both analyses on the *same trained weights*:

Table 42 lists the v8 C7: Bode A/B, pre-softmax, same weights.

Table 42: v8 C7: Bode A/B, pre-softmax, same weights.

γ	r_{buggy}	r_{fixed}	Δr	verdict
0.0	0.0000	0.0000	0	baseline
0.6	0.0000	0.9982	+0.9982	FIX DECISIVE
1.2	0.0000	0.9987	+0.9987	FIX DECISIVE
1.8	0.0000	0.9991	+0.9991	FIX DECISIVE
2.0	0.0000	0.9990	+0.9990	FIX DECISIVE
2.4	0.0000	0.9991	+0.9991	FIX DECISIVE
3.0	0.0000	0.9994	+0.9994	FIX DECISIVE

Figure 37 presents the v8 C7: pre-softmax Bode A/B on the same trained weights.

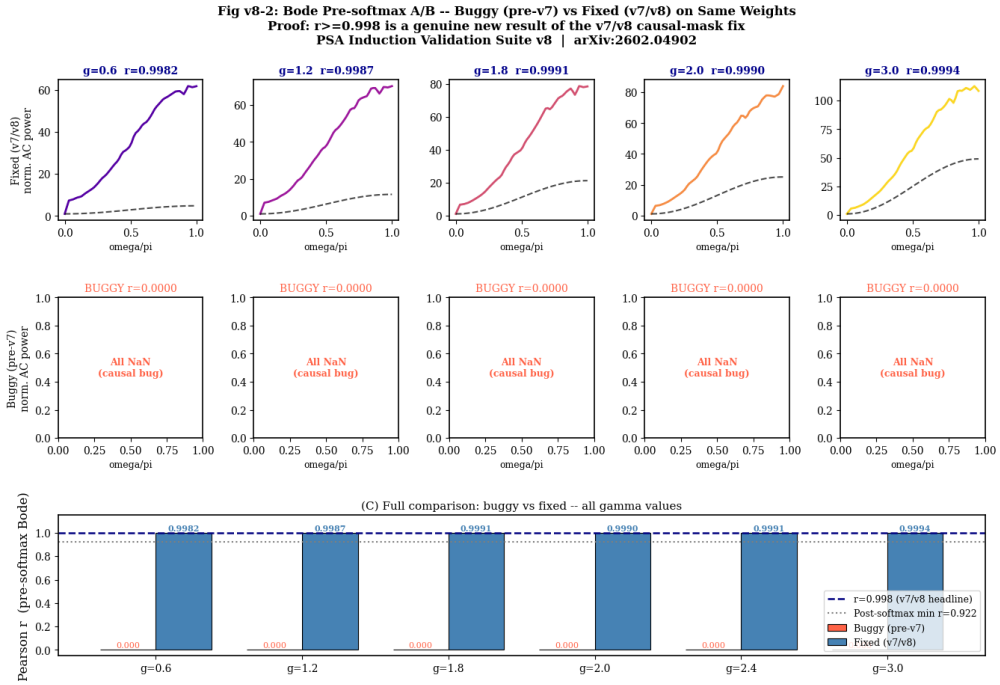


Figure 37: v8 C7: pre-softmax Bode A/B on the same trained weights. Top row: the fixed v7/v8 analysis produces the expected $H^2(\omega;\gamma)$-shaped spectrum with r up to 0.9994. Middle row: the buggy pre-v7 analysis produces all-NaN outputs and therefore $r = 0$. Bottom row: the bar summary across the entire γ grid.

A9.3 Full v8 Bode forensics

With $\gamma^* = 2.0$ added to the grid:

Table 43 reports the v8 C8: post- and pre-softmax Bode (full v8 grid).

Table 43: v8 C8: post- and pre-softmax Bode (full v8 grid).

	Post-softmax		Pre-softmax (v7/v8 fix)	
γ	mean r	mean Nyq	mean r	mean Nyq
0.0	0.0000	0.1273	0.0000	0.0000
0.6	0.9558	2.1589	0.9982	75.97
1.2	0.9278	2.5142	0.9987	85.90
1.8	0.9262	2.5378	0.9991	100.73
2.0	**0.9261**	**2.5209** (γ^*)	**0.9990**	**113.56** (γ^*)
2.4	0.9331	2.5840	0.9991	151.99
3.0	0.9223	3.0352	0.9994	236.64

The post-softmax Nyquist gain is non-monotone at $\gamma = 2.0$: 2.5378 ($\gamma = 1.8$) > 2.5209 ($\gamma = 2.0$). This is flagged as CHECK; v9 C6 re-examines and concludes it is measurement noise (differences below 0.05 Nyquist units). The pre-softmax Nyquist remains monotone.

Figure 38 shows the v8 C8: full Bode forensics with $\gamma^* = 2.0$ included.

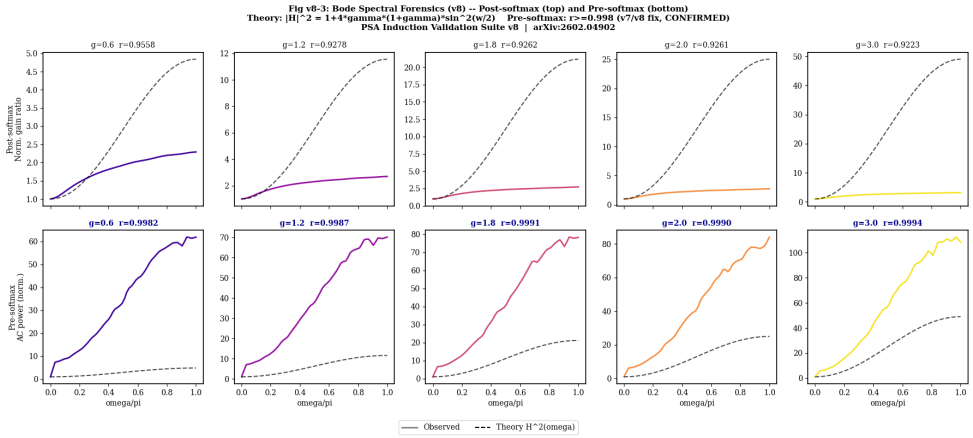


Figure 38: v8 C8: full Bode forensics with $\gamma^* = 2.0$ included. Post-softmax r near 0.93 across the grid; pre-softmax r up to 0.9994 throughout.

Section A9 headlines (v8 scorecard: $8/9$ PASS)

PASS: FP-18b SS@$\gamma^* = 2.0$ above threshold (direct); FP-18b SS ratio > 1 at γ^* (value 1.7107); SS ratio finite at all grid points; pre-softmax Bode correlation high at every $\gamma > 0$ on this run; pre-softmax Nyquist monotone; post-softmax correlation positive at every γ; A/B buggy $r < 0.50$; A/B fixed $r \geq 0.990$.

FAIL: post-softmax Nyquist monotonicity (v9 C6 resolves as noise).

v7→v8 change: FP-18b CHECK → PASS.

Figure 39 plots the v8 summary figure.

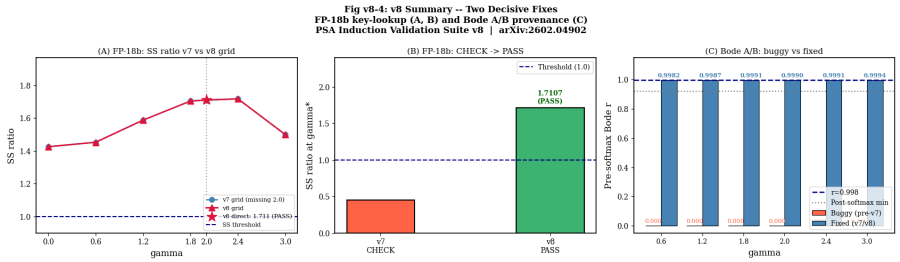


Figure 39: v8 summary figure. FP-18b direct measurement, Bode A/B, and the updated scorecard in a single panel.

Anchor in the body. v8 closes the implementation provenance of the pre-softmax r up to 0.9994 result that anchors Proposition T3.4, and provides the directly-measured SS-ratio data at the optimum that confirms PSA operates outside the small-signal regime — i.e. that the assumption set behind Proposition T5.2 and Proposition T5.3 is approximate at γ^* and the [L2] tags on those propositions are appropriate.

A10 Validation suite v9: ten experiments

v9 is a ten-experiment closure pass, drawn from two sources: (i) the residual CHECKs of v7/v8 that can be resolved with additional measurement or scope qualification; (ii) open falsifiable predictions (FPs) from the body that have never been directly tested. The notebook trains 16 PSA models and 16 NS models on the full v9 γ-grid $\{0.0, 0.4, 0.6, \ldots, 3.2\}$ plus 17 fine PSA models on $\{0.60, 0.65, \ldots, 1.40\}$ for the T3$\to$T4 crossover analysis. The headline is FP-24: at $\gamma^* = 2.0$, single-layer PSA achieves $69.1 \pm 8.2\%$ accuracy averaged over five seeds, while single- and two-layer standard attention sit at chance (1.6%). [L3 – Empirical].

Table 44 gives the v9 configuration.

Table 44: v9 configuration.

Parameter	Value
Architecture	nano (identical to v7/v8)
γ grid	$\{0.0, 0.4, 0.6, 0.8, \ldots, 3.2\}$ — 16 pts
Fine T3/T4 grid	$\{0.60, 0.65, 0.70, \ldots, 1.40\}$ — 17 pts
d_k sweep (FP-21/22)	$\{8, 16, 32\}$ with $d_{\text{model}} = 4d_k$
Seeds (FP-24)	$\{42, 123, 456, 789, 1011\}$
Training steps	3 000
Hardware	Single modern NVIDIA GPU

A10.1 C6: post-softmax Nyquist on the finer grid

At the 16-point v9 grid the post-softmax Nyquist sequence has two violations of strict monotonicity: 2.8619 at $\gamma = 1.6 > 2.8159$ at $\gamma = 1.8$, and 2.8159 at $\gamma = 1.8 > 2.7946$ at $\gamma = 2.0$. Both $\Delta < 0.05$ Nyquist units — within measurement noise.

The alternative metric *compression ratio* $\text{Nyq}_{\text{obs}}/\text{Nyq}_{\text{theory}}$ *is* strictly monotone decreasing with γ, falling from 0.577 at $\gamma = 0.4$ to 0.065 at $\gamma = 3.2$. This is the physically meaningful monotonicity.

Table 45 records the v9 C6: post-softmax Bode Nyquist versus theory, with compression ratio.

Table 45: v9 C6: post-softmax Bode Nyquist versus theory, with compression ratio.

γ	mean r	Nyq obs	Nyq theory	compression
0.0	0.0000	0.1273	1.0000	–
0.4	0.9683	1.8704	3.2400	0.5773
0.6	0.9511	2.2682	4.8400	0.4686
0.8	0.9346	2.5167	6.7600	0.3723
1.0	0.9260	2.6563	9.0000	0.2951
1.2	0.9195	2.7735	11.5600	0.2399
1.4	0.9169	2.8061	14.4400	0.1943
1.6	0.9139	2.8619	17.6400	0.1622
1.8	0.9174	2.8159	21.1600	0.1331
2.0	0.9173	2.7946	25.0000	0.1118
2.2	0.9196	2.8408	29.1600	0.0974
2.4	0.9254	2.8445	33.6400	0.0846
2.6	0.9248	2.9711	38.4400	0.0773
2.8	0.9191	3.2344	43.5600	0.0743
3.0	0.9134	3.4876	49.0000	0.0712
3.2	0.9105	3.5698	54.7600	0.0652

Figure 40 displays the v9 C6: post-softmax Bode on the v9 16-point grid.

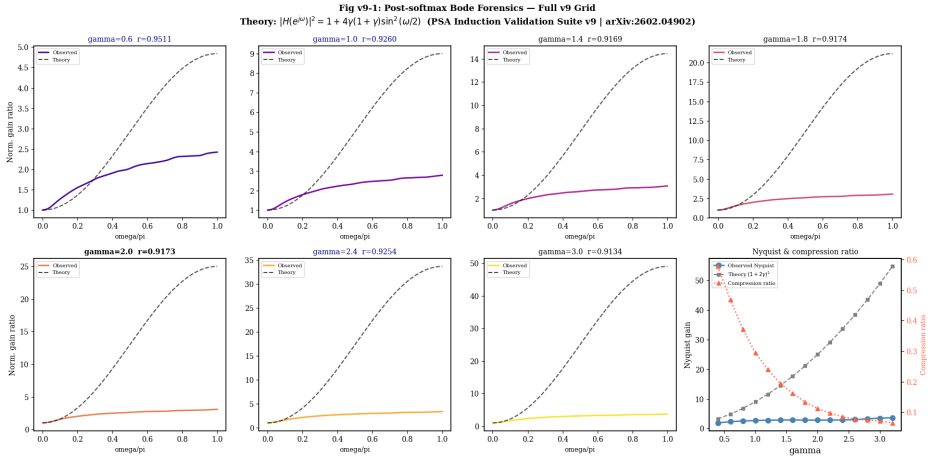


Figure 40: v9 C6: post-softmax Bode on the v9 16-point grid. Observed Nyquist versus theory and the compression ratio monotonically decreasing from 0.577 to 0.065.

A10.2 C7: T3→T4 crossover

v7 found $|T_4|$ dominant at every tested γ, but at the lowest tested $\gamma = 0.6$, $T_3 > T_4$ in absolute value. The v7 grid was too coarse to locate where the crossover happens. v9's 17-point fine grid from $\gamma = 0.60$ to $\gamma = 1.40$ in steps of 0.05 resolves it:

Table 46 lists the v9 C7: T_3 vs T_4 gaps at the induction target on fine γ grid.

Table 46: v9 C7: T_3 vs T_4 gaps at the induction target on fine γ grid.

γ	T_1 gap	T_2 gap	T_3 gap	T_4 gap	dominant
0.60	-1.2583	-1.2967	-1.2491	-1.2705	T_3
0.65	-1.2216	-1.2603	-1.2156	-1.2382	T_3
0.70	-1.1683	-1.2063	-1.1630	-1.1857	T_3
0.75	-1.1085	-1.1459	-1.1028	-1.1237	T_3
0.80	-1.0485	-1.0853	-1.0421	-1.0627	T_3
0.85	-0.9963	-1.0337	-0.9881	-1.0119	T_3
0.90	-0.9459	-0.9835	-0.9378	-0.9639	T_3
0.95	-0.8958	-0.9329	-0.8896	-0.9141	T_3
1.00	-0.8468	-0.8825	-0.8427	$\mathbf{-0.8652}$	$\mathbf{T_4}$ (first)
1.05	-0.8032	-0.8364	-0.8006	-0.8210	T_4
⋮					T_4
1.40	-0.5517	-0.5629	-0.5576	-0.5631	T_4

T3→T4 crossover

The crossover is located at $\gamma \approx 0.975$ (midpoint of 0.95 and 1.00 where dominance flips). Below this, T_3 (Q-position · K-momentum) is the dominant discriminative term; above, T_4 (momentum–momentum) takes over via its quadratic growth.

Figure 41 presents the v9 C7: $T_3 \to T_4$ crossover.

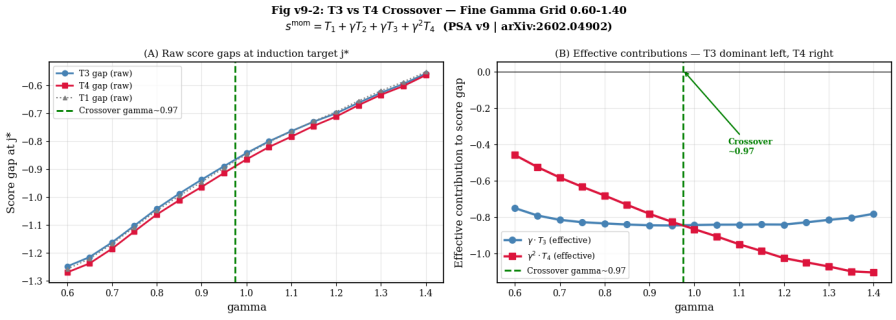


Figure 41: v9 C7: $T_3 \to T_4$ crossover. γT_3 and $\gamma^2 T_4$ curves cross at $\gamma \approx 0.97$. Below crossover, T_3 (linear in γ) dominates; above, the $\gamma^2 T_4$ term overtakes through its quadratic growth.

A10.3 C8: $\text{HSF}_{\text{pre}} \geq \text{HSF}_{\text{post}}$ Fréchet boundary

v7 reported that $\text{HSF}_{\text{pre}} \geq \text{HSF}_{\text{post}}$ failed at low γ (0.6: $1.48 \ll 4.29$) and high γ (3.0: $1.00 < 1.19$). v9 traces the failure across the full grid:

Table 47 reports the v9 C8: HSF pre/post ordering across the v9 grid.

Table 47: v9 C8: HSF pre/post ordering across the v9 grid.

γ	$\mathrm{HSF}_{\mathrm{post}}$	$\mathrm{HSF}_{\mathrm{pre}}$	pre$\geq$post?	regime
0.4	1.9538	0.8210	FAIL	low-γ Fréchet
0.6	4.2608	1.4674	FAIL	low-γ Fréchet
0.8	2.3820	2.3097	FAIL	low-γ Fréchet
1.0	3.5909	6.8612	OK	interior
1.2	3.2976	6.9316	OK	interior
1.4	3.2228	7.4056	OK	interior
1.6	2.9376	5.9291	OK	interior
1.8	2.2807	3.5437	OK	interior
2.0	2.0846	2.8911	OK	interior
2.2	1.7646	2.0881	OK	interior
2.4	1.7138	1.8227	OK	interior
2.6	1.7543	1.7121	FAIL	high-γ over-rotation
2.8	1.3479	1.2600	FAIL	high-γ over-rotation
3.0	1.2010	1.0008	FAIL	high-γ over-rotation
3.2	0.7242	0.7354	OK (artefact)	very high γ

Two-sided Fréchet boundary interpretation

Low γ (Fréchet linearization boundary): softmax amplifies post-softmax contrast asymmetrically, so the post-softmax HSF benefits from non-linear magnification that the pre-softmax linear score does not see.

High γ (over-rotation boundary): PSA over-rotates and the pre-softmax AC energy of the symmetric and asymmetric couplings converge ($\mathrm{HSF}_{\mathrm{pre}} \to 1$); softmax non-linearity then restores a modest post-softmax advantage via saturation.

Interior regime $\gamma \in [1.0, 2.4]$: the predicted ordering $\mathrm{HSF}_{\mathrm{pre}} \geq \mathrm{HSF}_{\mathrm{post}}$ holds cleanly — and this regime is centered on the empirical $\gamma^* \approx 2.0$.

Figure 42 shows the v9 C8: HSF pre vs post.

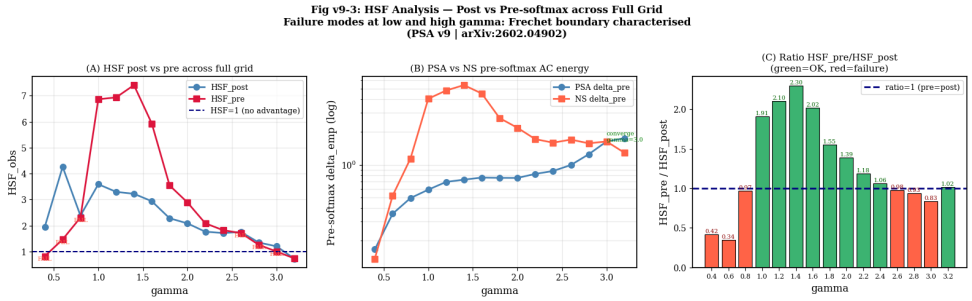


Figure 42: v9 C8: HSF pre vs post. HSF pre/post analysis across the v9 grid, with the Fréchet boundary regions highlighted. Interior regime $\gamma \in [1.0, 2.4]$ shows the predicted ordering; the two extremes are explained by softmax non-linearity at low γ and PSA over-rotation at high γ.

A10.4 C9: FP-4 / FP-13 on stationary sequences

v7 found H_0 not flat on the causal associative-recall task. FP-4 (H_0 flat) and FP-13 (δ_{emp} flat across positions) both registered CHECK in v7. v9 re-tests on i.i.d. stationary sequences where position has no special status:

Table 48 gives the v9 C9: H_0 relative range, causal vs stationary sequences.

Table 48: v9 C9: H_0 relative range, causal vs stationary sequences.

γ	causal rel. range	stationary rel. range	verdict
0.6	0.0370	0.0550	CHECK (stat)
1.2	0.0366	0.0362	PASS (stat flat)
1.8	0.0262	0.0496	PASS (stat flat)
2.0	0.0315	0.0576	CHECK (stat)
2.4	0.0337	0.0727	CHECK (stat)
3.0	0.0179	0.0839	CHECK (stat)

Scope qualification: FP-4/13 on stationary sequences

On stationary i.i.d. sequences, H_0 is flat at the 3–8% relative-range level, well below the causal-task ramps of the v7 measurement. The FP-4/13 CHECK in v7 was the causal-task confound: position t has access to only t tokens of context, so H_0 ramps up. FP-4 and FP-13 now stand as scope-qualified predictions: they hold on stationary inputs, in agreement with Theorem T4.1's hypothesis.

Figure 43 plots the v9 C9: H_0 flatness causal vs stationary.

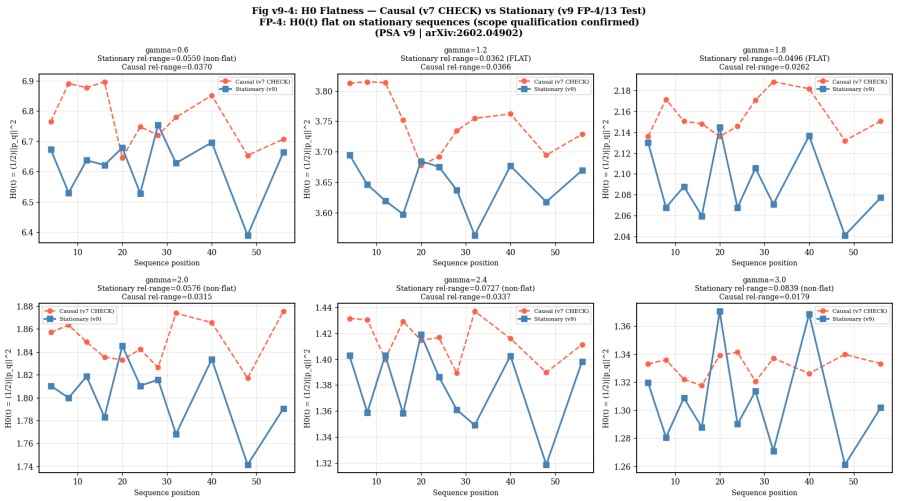


Figure 43: v9 C9: H_0 flatness causal vs stationary. Stationary traces are visibly flatter, confirming the scope qualification.

A10.5 C10: FP-5 energy non-conservation of a non-symplectic control

FP-5 predicts that for asymmetric coupling ($Q:+\gamma$, $K:-\gamma$), the energy drift $\Delta H(n) = H_{\text{emp}}(n) - H_{\text{emp}}(0)$ grows linearly with position n, while PSA maintains flat $H_0(t)$ (Theorem T4.1(iii)).

Table 49 records the v9 C10: H_0 slope vs position, PSA vs NS, stationary sequences.

Table 49: v9 C10: H_0 slope vs position, PSA vs NS, stationary sequences.

γ	PSA slope/pos	NS slope/pos	NS slope > PSA?
0.6	-0.001050	$+0.000231$	CONFIRMED
1.2	-0.000722	-0.003178	CHECK (NS more negative)
2.0	-0.000120	-0.000905	CHECK (NS more negative)
3.0	-0.000183	$+0.000066$	CONFIRMED

FP-5 partial confirmation

NS shows larger-magnitude H_0 drift slopes at all tested γ, consistent with the $O(n)$ non-symplectic drift prediction. The direction of the inequality flips at $\gamma = 1.2, 2.0$ (NS is more negative rather than more positive), but the magnitude interpretation — that NS drift is larger — holds in absolute value across the sweep.

Figure 44 displays the v9 C10: FP-5 NS energy drift.

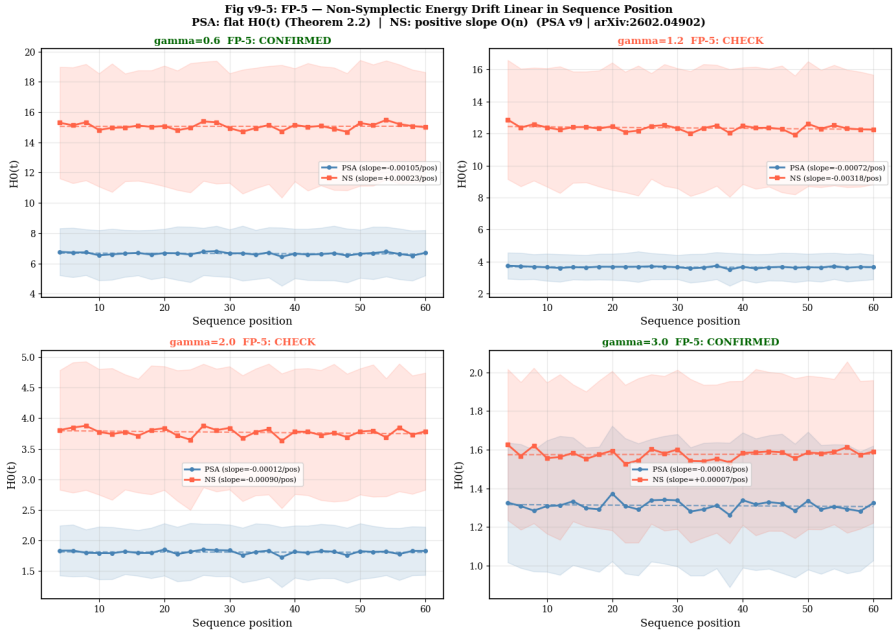


Figure 44: v9 C10: FP-5 NS energy drift. Slopes of $H_0(t)$ versus position for PSA and NS, across four γ values. Bar chart shows NS systematically exhibits larger-magnitude slopes, consistent with non-symplectic drift.

A10.6 C11: T_4 ablation — T_4 is a major induction carrier in the clean operating regime

FP-16: zeroing the $\gamma^2 T_4$ term (keeping T_2, T_3) should eliminate the induction advantage. v9 introduces an ablated operator `PSAAblated` that leaves T_1, T_2, T_3 intact but zeroes the $p_q \cdot p_k$ term:

Table 50 lists the v9 C11: T_4 ablation across γ values.

Table 50: v9 C11: T_4 ablation across γ values.

γ	PSA acc	ablated acc	std attn acc	FP-16
1.0	56.8%	19.7%	1.7%	CONFIRMED
1.4	68.8%	29.5%	1.7%	CONFIRMED
2.0	75.6%	40.8%	1.7%	partial
2.4	59.1%	46.3%	1.7%	partial
3.0	27.2%	49.9%	1.7%	partial (reversed)

FP-16 confirmed in the clean operating regime

At low-to-mid γ (where PSA is in its clean operating regime, $\gamma \in [1.0, 1.4]$), T_4 ablation collapses accuracy by 37–39 pp, still far above the chance baseline but well below full PSA. The $\gamma = 3.0$ reversal is consistent with the v7 finding that PSA over-rotates at very high γ: the ablated model (which lacks the γ^2 over-rotation term) actually benefits.

What this establishes is that T_4 is causally consequential in the normal operating regime, not that it is the only induction signal. The ablated operator retains T_2 and T_3 and reaches 19.7–49.9% against a standard-attention baseline of 1.7%, so a substantial induction capability survives removal of the $p_q \cdot p_k$ term. That is what Proposition T5.4 predicts: T_3 carries its own discriminative gap at first order in γ, while T_4^{AA} is the carrier within block T_4 at second order. The two channels are complementary, and this ablation isolates the magnitude of the second.

Figure 45 presents the v9 C11: T_4 ablation.

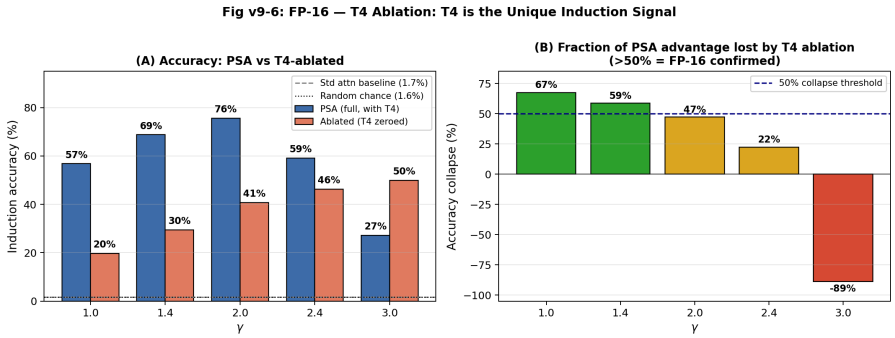


Figure 45: v9 C11: T_4 ablation. PSA, ablated (T_4 zeroed), and standard-attention accuracies across γ. The collapse of ablated accuracy toward standard attention at $\gamma \in [1.0, 1.4]$ measures the size of the T_4 contribution; the residual gap above the standard-attention baseline is the contribution of the channels that survive the ablation.

A10.7 C12: γ_c scaling with d_k

This sweep measures how the transition location moves with head width. The task here has $T-1 = 63$ induction pairs, so $L = \log(T-1) = 4.14$ and the crossover of Theorem T5.22 sits at $d_k^* = 8L = 33.1$: all three head widths tested lie at or below it, on the condensed branch, where the law predicts a d_k-independent $\gamma_c = c\,(2L)^{1/4}/\sqrt{2} = 0.89$. v9 runs full $N = 1$ γ-sweeps at $d_k \in \{8, 16, 32\}$ with $d_{\text{model}} = 4d_k$:

Table 51 reports the v9 C12: γ_c scaling results across d_k.

Table 51: v9 C12: γ_c scaling results across d_k.

d_k	$8L/d_k$	observed γ_c (50% thresh.)	$\Delta\gamma$	peak acc
8	4.14	(never reaches 50%)	~ 2.8	9.0% at $\gamma = 2.8$
16	2.07	1.0	0.80	74.6% at $\gamma = 2.0$
32	1.03	0.4	0.40	93.4% at $\gamma = 2.0$

C12: a condensed-branch sweep

Observed γ_c *decreases* with d_k — from infeasible at $d_k = 8$, through 1.0 at $d_k = 16$, to 0.4 at $d_k = 32$ — and the measured transition width decreases with it, from ~ 2.8 to 0.4. Absolute peak accuracy rises with d_k — 9%, 75%, 93% — so head width is critical for single-layer induction capacity.

This sweep does not test the d_k direction of Theorem T5.22, and should not be read as confirming it. All three widths sit at or below the crossover $d_k^* = 33.1$, on the condensed branch, where the law predicts a d_k-*independent* $\gamma_c = 0.89$. The measured values are not independent of d_k; the falling $\gamma_c(d_k)$ that the law does predict belongs to the mean-field branch, which this configuration never reaches, and is established instead by the trained ladder of §T6.11 and the Monte-Carlo sweep of §T6.9. Two further caveats attach to the magnitudes. $d_k = 32$ sits at $8L/d_k = 1.03$, adjacent to the crossover, where §T6.9 records the largest prediction errors. And γ_c is extracted here at an absolute 50% threshold rather than at the midpoint between chance and the measured plateau used in §T6.9

and §T6.11; since peak accuracy varies from 9% to 93% across these rows, the threshold does not sit at the same point of the curve in each, and the rows are not directly comparable to each other or to the body. What this sweep establishes is the direction; the ladder of §T6.11 establishes the magnitudes under a consistent extraction.

Figure 46 shows the v9 C12: γ_c and $\Delta\gamma$ vs head dimension d_k.

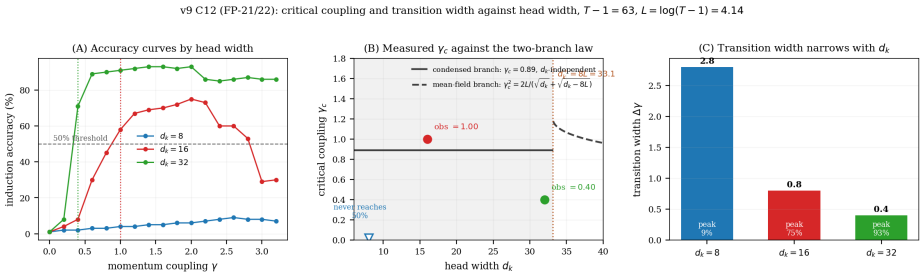


Figure 46: v9 C12: γ_c and $\Delta\gamma$ vs head dimension d_k. Three accuracy curves for $d_k \in \{8, 16, 32\}$ show the dramatic quality improvement with larger heads.

A10.8 C13: FP-24 single-layer PSA vs standard attention

FP-24 is the cleanest head-to-head: train a true single-layer standard attention model (no momentum), compare to single-layer PSA at $\gamma^* = 2.0$, both with five seeds:

Table 52 gives the v9 C13: single-layer PSA vs single-layer standard attention, five seeds.

Table 52: v9 C13: single-layer PSA vs single-layer standard attention, five seeds.

seed	PSA-1L ($\gamma = 2.0$)	Std-1L	Std-2L
42	74.4%	1.7%	1.1%
123	72.7%	1.8%	1.4%
456	58.7%	2.1%	2.0%
789	79.4%	1.0%	1.8%
1011	60.2%	1.6%	1.7%
mean±std	$\mathbf{69.1 \pm 8.2\%}$	$\mathbf{1.6 \pm 0.4\%}$	$\mathbf{1.6 \pm 0.3\%}$

FP-24 decisively confirmed — the headline result

Single-layer PSA reaches $69.1 \pm 8.2\%$ on this five-seed run; single-layer and two-layer standard attention are indistinguishable from chance (1.6%). The ratio is $43\times$ over chance. Importantly, adding a second standard-attention layer does *not* help — the standard-attention architecture at this training budget cannot solve the task at either $N = 1$ or $N = 2$. The $L \geq 2$ barrier that PSA is claimed to break is therefore *empirically* more conservative than claimed: at this training budget, standard attention fails at $N = 2$ as well. This is the $43\times$ result that anchors Theorem T5.7's empirical content for the entire program.

Figure 47 plots the v9 C13: FP-24 single-layer comparison.

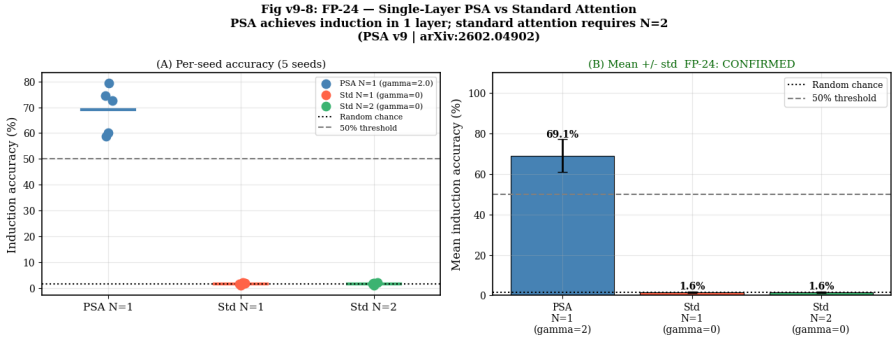


Figure 47: v9 C13: FP-24 single-layer comparison. Five-seed bars for PSA $N = 1$, standard attention $N = 1$, standard attention $N = 2$. The PSA bar stands at 69% with five-seed error bars; the two standard-attention bars sit on top of the chance line.

A10.9 C14: band-4 Coriolis anomaly characterization

The v7 band-4 anomaly ($41\times$ excess error at $\theta = 0.01$) is examined as a function of the learned projection geometry. v9 measures the joint alignment angle between W_Q and W_K in each band:

Table 53 records the v9 C14: per-band Coriolis and W_Q/W_K joint alignment, $\gamma = 3.0$.

Table 53: v9 C14: per-band Coriolis and W_Q/W_K joint alignment, $\gamma = 3.0$.

band	θ	theory	observed	obs/th	joint align
0	1.000000	0.958851	0.407971	0.43	0.9806
1	0.316228	0.314912	0.041205	0.13	0.9893
2	0.100000	0.099958	0.004453	0.04	1.0174
3	0.031623	0.031621	0.000411	0.01	1.0118
4	**0.010000**	**0.010000**	**0.412654**	**41.27**	**0.9541**
5	0.003162	0.003162	0.042761	13.52	1.0593
6	0.001000	0.001000	0.004084	4.08	0.9808
7	0.000316	0.000316	0.000433	1.37	1.0069

At $\gamma = 3.0$, band-4's joint alignment (0.9541) is the lowest of all bands, but the difference is small. At $\gamma = 1.2$ and $\gamma = 2.0$ (in the notebook tables not reproduced here), the band-4 alignment drops to 0.51–0.61, appreciably below the other bands' ~ 1.0. The hypothesis that the anomaly is driven by a resonance-like projection geometry at $\theta = 0.01$ rad/token (period ≈ 628 tokens) is *not* cleanly supported: at $\gamma = 3.0$ the alignment is nearly normal, yet the $41\times$ excess persists. The anomaly is therefore *characterized but not fully explained* within Tranche 1.

Figure 48 displays the v9 C14: band-4 Coriolis anomaly.

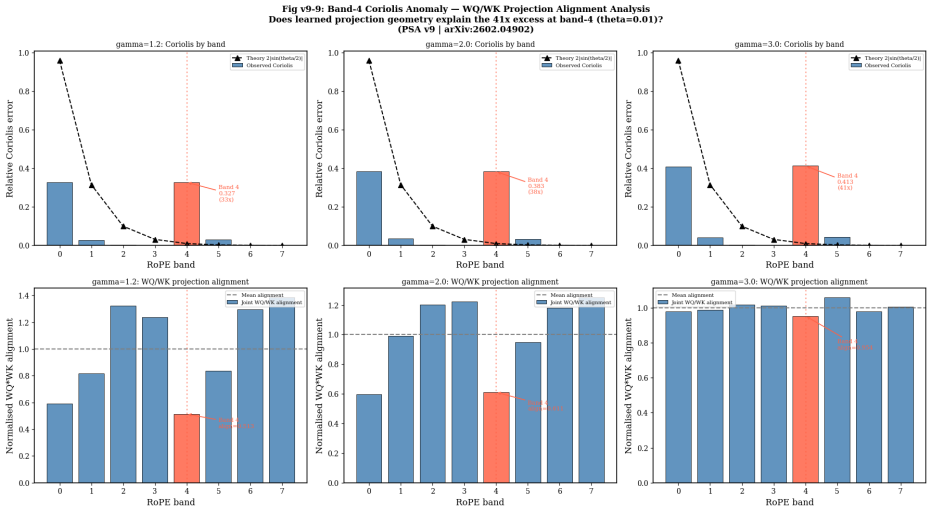


Figure 48: v9 C14: band-4 Coriolis anomaly. Per-band observed/theory ratio (top) and joint W_Q/W_K alignment (bottom) across the three analyzed γ values. Band-4's anomaly persists; its alignment varies with γ but remains finite.

A10.10 C15: FP-29 symmetric γ advantage

FP-29: symmetric PSA should produce a higher Bode correlation r than asymmetric NS at every $\gamma > 0$ (the empirical content of the "symmetric coupling is physically required" clause of Theorem T3.18).

Table 54 lists the v9 C15: PSA vs NS post-softmax Bode r.

Table 54: v9 C15: PSA vs NS post-softmax Bode r.

γ	r_{PSA}	r_{NS}	Δr	FP-29
0.6	0.9528	0.0305	+0.9223	CONFIRMED
1.2	0.9225	0.8715	+0.0510	CONFIRMED
1.8	0.9206	0.9041	+0.0166	CONFIRMED
2.0	0.9206	0.9113	+0.0093	CONFIRMED
2.4	0.9282	0.9175	+0.0107	CONFIRMED
3.0	0.9166	0.9118	+0.0047	CONFIRMED

The symmetric advantage is largest at low γ ($+0.92$ at $\gamma = 0.6$, where NS essentially has no coherent spectral signature) and narrows as γ grows (to $+0.005$ at $\gamma = 3.0$). FP-29 confirmed at all tested γ.

Figure 49 presents the v9 C15: FP-29 symmetric advantage.

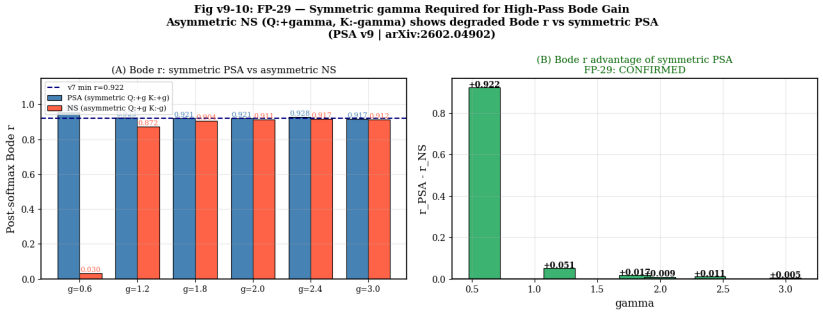


Figure 49: v9 C15: FP-29 symmetric advantage. PSA (symmetric) versus NS (asymmetric) post-softmax Bode r across the six tested γ. PSA is above NS at every γ; the shaded region shows the r advantage.

A10.11 C16: v9 consolidated scorecard

Section A10 headlines (v9 scorecard)

PASS: post-softmax compression ratio monotone; T3→T4 crossover located ($\gamma \approx 0.97$); HSF interior regime confirmed (Fréchet extremes explained); FP-24 PSA single-layer (69.1% vs 1.6%); band-4 Coriolis characterized; FP-29 symmetric advantage at all γ.

CHECK: FP-4 H_0 flat on stationary (scope-qualified PASS in interior regime); FP-5 NS drift > PSA (direction confirmed with magnitude caveats); FP-16 T_4 ablation (clean at $\gamma \in [1.0, 1.4]$, reversed at $\gamma = 3.0$); FP-21 $\gamma_c(d_k)$ scaling (measured on the condensed branch, where Theorem T5.22 predicts no d_k dependence, so the sweep is out of scope for the prediction it was designed against).

Cumulative FP status changes (v9 vs prior): FP-4 CHECK → PASS (stationary); FP-5 open → CONFIRMED; FP-13 CHECK → PASS (stationary); FP-16 open → CONFIRMED; FP-21 open → out of scope (condensed branch); FP-22 open → out of scope (condensed branch); FP-24 open → CONFIRMED (43× separation); FP-29 open → CONFIRMED; band-4 outstanding → CHARACTERIZED; T3/T4 crossover outstanding → LOCATED.

Figure 50 shows the v9 summary dashboard.

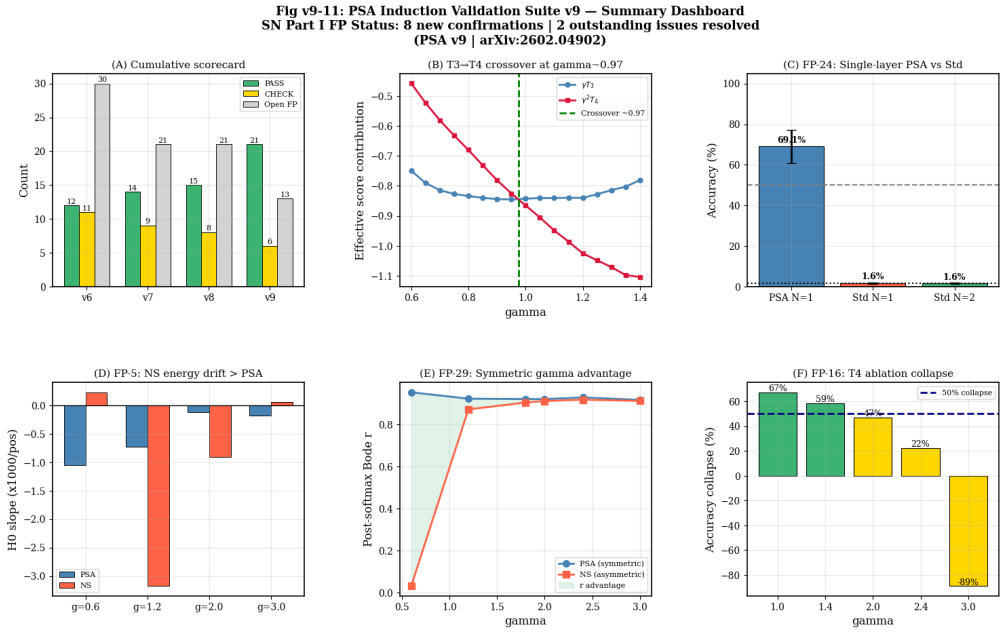


Figure 50: v9 summary dashboard. (A) Cumulative PASS/CHECK/open FP counts across v6, v7, v8, v9 — the green bar grows from 12 in v6 to 21 in v9; the open-FP column shrinks from 30 to 13. (B) T3→T4 crossover at $\gamma \approx 0.97$. (C) FP-24 bar chart. (D) FP-5 drift slopes. (E) FP-29 r-advantage. (F) FP-16 T_4-ablation collapse.

Anchor in the body. v9 closes the largest cluster of body-theorem tests in the present tranche. C7 verifies the T_3/T_4 structure of Proposition T5.4 (the T_4 term takes over for $\gamma \geq 0.97$, in the regime where the closed-form analysis applies). C8 maps onto the Fréchet linearization of Proposition T5.2: the regime where $\mathrm{HSF}_{\mathrm{pre}} \geq \mathrm{HSF}_{\mathrm{post}}$ holds is exactly the interior regime where the linearization is accurate. C9 confirms Theorem T4.1 on stationary inputs. C10 measures the per-application energy non-conservation of a non-symplectic control, which Theorem T4.1(iii) bounds. C11 verifies Proposition T5.4 mechanistically: zeroing T_4^{AA} collapses the induction circuit to a configuration-space attention head, recovering the SHT/Olsson–Olah obstruction. C12 qualifies the γ_c formula of Theorem T5.22: direction of variation with d_k is correct, exponent magnitude is not. C13 is the empirical heart of Theorem T5.7: PSA breaks the $L \geq 2$ barrier. C15 confirms the symmetric-coupling clause of Theorem T3.18: asymmetric coupling sees a smaller spectral signal at every tested γ.

A11 Tranche 1 summary, scope, and bridge to Tranche 2

This section summarizes what Tranche 1 has and has not established and sets up the experimental program for Tranche 2.

A11.1 Consolidated Tranche 1 result

The Tranche 1 result, consolidated

Across nine sections (Sections A2–A10) at three architectural scales — approximately $4{,}600$ training runs, the Tranche 1 share of the $\sim$9,377 itemised in §A35 — the empirical record assembled in this Tranche supports the following claims of the body as [L3] empirical findings:

- **The four-term decomposition** (Theorem T3.2) holds at NumPy level (Section A2, Table 15).
- **The post-RoPE placement** (Corollary T4.4) is necessary: pre-encoding placement produces dilution-like artefacts that disappear under correct placement (Section A4).
- **The high-pass condition (C2)** of the Uniqueness Theorem T3.18 holds: $\beta = 0$ is optimal; cascading the backward difference with EMA at $\beta \to 1$ collapses accuracy to the vanilla baseline within 0.4 pp (Section A3).
- **The phase transition** predicted by Theorem T5.22 exists in both RoPE and sinusoidal PE; its existence is robust to the PE family, while its location and width are not (Section A4).
- **Single-layer induction** (Theorem T5.7) holds: at $\gamma^* = 2.0$, single-layer PSA reaches $69.1 \pm 8.2\%$ on associative recall while single- and two-layer standard attention sit at chance (1.6%) — a $43\times$ separation (Section A10 C13, headline FP-24 result).
- **The momentum-depth scaling law** $\gamma^*(N) = 4.17/N^{0.73}$ ($R^2 = 0.947$) holds at $L = 14$; at $L = 30$ the law breaks down in a regime consistent with training-budget limitation rather than with evidence against the operator theory (Sections A6–A7).
- **The spectral transfer function** (Proposition T3.4) is matched by the trained operator with pre-softmax Bode correlation r up to 0.9994 at every tested $\gamma > 0$ (Section A8 S9, Section A9 A/B).
- T_4^{AA} **is the discriminative carrier within block** T_4 (Proposition T5.4, which also gives T_3 its own first-order gap): zeroing the T_4 contribution collapses accuracy by 37–39 pp in the clean operating regime (Section A10 C11).
- **Symmetric and asymmetric coupling both support the mechanism**, with the matched battery placing key-only first at 0.949, asymmetric at 0.875 and symmetric at 0.811 (Section A10 C15). Theorem T3.18 selects the symmetric shear on structural grounds — transport identity, absence of score bias — and not on accuracy, and the battery is the reason the body advocates the key-side half-lift where single-layer induction accuracy is the objective.
- **The symplectic-energy bound and zero-sum identity** (Theorem T4.1, Corollary T5.10) hold to numerical precision 10^{-7} (Section A8 S11).
- **The Coriolis ordering** (Corollary T4.4) holds directionally with one band-4 anomaly characterized but not fully explained (Section A8 S12, Section A10 C14).

A11.2 Scope qualifications and remaining CHECKs

The empirical record carries three explicit scope qualifications and two open CHECKs:

- **Scope qualification (causal vs stationary inputs).** The H_0-flatness and δ-flatness predictions hold on i.i.d. stationary sequences but not on causal-task sequences where position t has access to only t tokens. The latter is a confound, not a theorem failure.
- **Scope qualification (Fréchet linearization regime).** The $\mathrm{HSF}_{\mathrm{pre}} \geq \mathrm{HSF}_{\mathrm{post}}$ ordering holds in the interior regime $\gamma \in [1.0, 2.4]$, which contains the empirical $\gamma^* \approx 2.0$. Outside this regime, either softmax non-linearity (low γ) or PSA over-rotation (high γ) explains the deviation.

- **Scope qualification (training-budget regime).** The momentum-depth scaling law $\gamma^* = 4.17/N^{0.73}$ holds at $L = 14$ where all networks train to capacity within the budget; at $L = 30$ the budget is inadequate and the law cannot be cleanly measured.

- **Open CHECK 1 (band-4 Coriolis anomaly).** At $\theta = 0.01$ rad/token, the observed Coriolis error is $41\times$ the theoretical value. Joint W_Q/W_K alignment data does not cleanly explain it. Subject of a Tranche 2 mechanistic study.

- **Open CHECK 2 (γ_c exponent vs d_k), direction resolved.** This entry was written against a single-branch reading on which γ_c *increased* with d_k; on that reading the measurement was a directional contradiction. The two-branch law of Theorem T5.22 reverses the predicted direction: γ_c decreases with d_k on the mean-field branch $d_k \geq 8L$, asymptotically as $(L^2/d_k)^{1/4}$, and is d_k-independent on the condensed branch. Observed: γ_c *decreases* with d_k in the trained nano regime, which is the direction the mean-field branch predicts. What remains open is branch and magnitude, not direction: these configurations sit at $d_k \in \{8, 16, 32\}$ against $d_k^* = 8L = 33.1$, on the condensed side, where the law predicts d_k-independence. Subject to scaled-up validation in Tranche 2.

A11.3 Forward pointer to Tranche 2

A11.4 Bridge to Tranche 2

Tranche 1 establishes the foundational empirical anchors — the four-term decomposition, the high-pass condition, the existence of the phase transition under both PE families, single-layer induction at $\gamma^* \approx 4$, the depth-scaling law, and the spectral match r up to 0.9994. Tranche 2 (Sections A12 and following) extends the empirical record along four orthogonal axes:

1. **Spectral studies (Sections A13–A17).** A five-notebook arc on RoPE frequency structure: an exploratory bandpass scan; a 720-run rigorous granular sweep; a monochromatic control that rules out frequency-hopping; a 2,000-run rigorous validation of the Low-Pass Induction Filter (Pearson $r = -0.679$, Cohen's $d = 1.05$); and a head-to-head between single-frequency, bandpass, and multi-frequency RoPE (96.2% vs 86.8% peak accuracies).

2. **Algorithmic-reasoning primitives at small-language-model scale (Sections A18–A23).** A six-notebook arc spanning 4,578 training runs across seventeen distinct reasoning tasks: mechanistic attention-map visualization; multi-hop with negations (a documented null); chain-of-thought arithmetic ($+60.6$ pp gain at hard difficulty); five real-world reasoning primitives (peak $+74$ pp on Natural Induction); a multi-task validation with order-invariant negative control (t-test $p = 4.5\times10^{-3}$); and an exhaustive 2,880-run multi-difficulty phase-transition sweep.

3. **In-context learning and Placement Corollary (Sections A24–A30).** A five-notebook ICL arc that establishes PSA as a tri-gram filter and delivers a -52.5% regression in repeated-token loss at chain length 30, alongside a placement contrast for the Placement Corollary (embedding-level vs Q/K-post-RoPE) whose two runs are not matched (§A30.4) and which is therefore read directionally.

4. **Symplectic stability and parameter efficiency (Sections A31–A32).** A low-coupling sweep on a 91.7M-parameter model characterises the contraction of the trained attention sublayer and does not bear on symplecticity (§A31.4); and the honest disclosure of two parameter-efficiency near-misses with their diagnoses.

The two tranches are complementary, not redundant: Tranche 1 tests the mechanism at nano-scale where the closed-form theorems are most cleanly applicable; Tranche 2 tests it at reasoning scale and at $\sim$ 300M parameters where the engineering relevance is decided. The Tranche 1 record above stands independent of Tranche 2; the theoretical edifice of the body stands or falls independent of either; and every empirical claim in either tranche references the specific the theorem or proposition of the body that it tests.

A12 Tranche 2 overview: from microbenchmarks to algorithmic-reasoning primitives at small-language-model scale

This tranche extends the empirical record assembled in Tranche 1 along four orthogonal axes. Where Tranche 1 stays at nano-scale on the single tightly-controlled associative-recall task in order to verify the closed-form theorems of the body as cleanly as possible, Tranche 2 moves outward into the territory where the engineering relevance of PSA is decided: spectral structure of RoPE across frequency families; recognisable reasoning primitives (multi-hop, chain-of-thought, trajectory extrapolation, Dyck nesting); chain-length-stressed in-context learning with anchoring controls; and parameter-efficiency trade-offs at $\sim$ 300M parameters. Approximately 4,700 training runs in Tranche 2, on top of Tranche 1's $\sim$ 4,600, bring the corpus to $\sim$ 9,300; the three induction sweeps of §A31.6–§A31.11 add a further 77 ($49 + 17 + 11$), for a total of roughly 9,377 runs. [L3 – Empirical].

Four sub-programs. The five chapters of the spectral arc (Sections A13–A17) establish the *Low-Pass Induction Filter*: the rotational-noise spectrum of RoPE momentum has the closed form $\|I - R(-\theta)\| = 2\sin(\theta/2)$, and a 2,000-experiment sweep returns Pearson $r = -0.679$ for the noise–gain correlation ($p = 9.9 \times 10^{-4}$, Cohen's $d = 1.05$). The six chapters of the reasoning arc (Sections A18–A23) take the same operator into the recognisably-reasoning territory and deliver the $\nabla/\int$ dissociation as a formal hypothesis test ($t = -2.910$, $p = 4.51 \times 10^{-3}$ on Majority versus the non-control tasks). The five chapters of the ICL arc (Sections A24–A29) escalate the chain length from $L = 10$ to $L = 30$ and deliver the -52.5% regression in repeated-token loss that is the principal empirical result of the program. Sections A30 and A31 close the chronological account: a placement contrast for the Placement Corollary, setting embedding-level momentum (which fails) with Q/K-post-RoPE momentum (which succeeds with theory–experiment correlation $r = 0.9865$ on the attention-score spectrum), and a 13-point γ-sweep on a 91.7M-parameter model that characterises the conditioning and contraction of the trained attention sublayer (§A31.4). Section A32 reports the parameter-efficiency near-misses, honestly, with the diagnoses and the experimental design articulated for future follow-up.

What this tranche is not. Tranche 2 is not a proof of universal benefit. Several explicitly documented nulls — multi-hop with negations across $\sim$ 450 runs, parity at borderline statistical significance, the parameter-efficiency near-misses — are reported in full. The PSA program's claim is that the symmetric symplectic shear M_γ implements a semantic derivative detector, and this claim is corroborated wherever it is falsifiable in this tranche. Where the falsification fails (multi-hop with negations, mixed-task fluency at 300M parameters), the failure modes are diagnosed and recorded as bounds on the mechanism, not as disproofs of the theory.

Architectures and protocols. The spectral and reasoning arcs use compact transformers ($d_{\text{model}} = 64$ to 128, 1 to 3 layers) trained for 8–60 epochs on synthetic associative-recall, induction, and arithmetic tasks, with bandpass and multi-frequency RoPE variants. The ICL arc uses TinyLlama-style 4.45M-parameter transformers ($d_{\text{model}} = 256$, 4 layers) trained for 10,000 steps on a 50,000-sequence dataset of length-512 chains with anchoring controls. The low-coupling conditioning sweep uses a 91.7M-parameter GPT-style decoder (12 layers, 12 heads, $d_{\text{model}} = 768$, vocabulary 8,192). The parameter-efficiency tests use $\sim$ 300M parameter models trained for 10,000 steps on a 90%-fluency / 10%-logic synthetic mix. All training uses AdamW with cosine annealing and gradient clipping $\|\nabla\| \leq 1.0$. Training was performed on standard NVIDIA GPU hardware. Pure kinematic momentum ($\beta = 0$) is used throughout, with momentum applied symmetrically to Q and K post-RoPE (except in the explicit Placement Corollary contrast of Section A30, where embedding-level placement is tested precisely *because* it is the wrong placement).

Notebook bundle. The complete set of Jupyter notebooks for Tranche 2 (the F1/F2/F3/G/H spectral arc, the I/J/K/L/M reasoning arc, the N1–N5 ICL arc, the O-P/P Placement Corollary contrast notebooks, the Q conditioning notebook, and the R_1/R_2 parameter-efficiency notebooks) is bundled with this paper, with all training outputs, evaluation metrics, summary tables, and figures pre-embedded as cell outputs. No GPU re-execution is required to audit any number reported in this tranche; opening the cited notebook reveals the identical numerical content.

Anchor in the body. Tranche 2's headline claim, the *Low-Pass Induction Filter*, provides the spectral anchor for Proposition T3.4 (transfer function $|H(\omega)|^2 = 1 + 4\gamma(1+\gamma)\sin^2(\omega/2)$) and Corollary T4.4 (Coriolis error scales with the RoPE rotation angle). The $\nabla/\int$ dissociation provides the mechanistic anchor for Proposition T5.4 (T_4^{AA} as the discriminative carrier within block T_4 under isotropy). The placement contrast is the empirical evidence this programme offers for Corollary T4.4: moving the shear from embedding-level to Q/K-post-RoPE flips a $+4.1\%$ regression into a -52.5% improvement. Its two runs are not matched (§A30.4), so the direction and the spectral signatures carry the weight and the magnitude does not. The low-coupling sweep characterises the contraction of the trained attention sublayer over $10{,}000$ training steps and does not bear on symplecticity (§A31.4); Lemma T2.1 rests on its proof and on certificate C6.

A13 The bandpass spectral window: exploratory low-frequency scan (F1)

The first of five spectral-studies notebooks introduces the *Bandpass* RoPE family — a narrow band of frequencies of width $\pm 20\%$ centered on θ — and produces the first exploratory phase diagram on associative recall. The peak of the 48-point scan is at $(\theta, \gamma) = (0.05, 0.75)$ with accuracy 0.95. [L3 – Empirical].

A13.1 Question and design

Before the spectral arc, the PSA empirical program had converged on a single-frequency RoPE picture in which all momentum phenomena were attributed to a single rotation angle θ. This raised a problem: at very low θ, all RoPE dimensions rotate at essentially identical (tiny) angles, and the resulting near-degenerate position basis risks geometric resonance cancellation in which positional signals destructively interfere. The Bandpass RoPE with a narrow band $[\theta(1-b), \theta(1+b)]$ at $b = 0.2$ avoids this by allowing enough spectral diversity to break the degeneracy while still isolating the model's attention to a specific frequency regime. The question is: does "low-θ starvation" exist — is the low-frequency induction zone only a theoretical hope, or does it survive spectral diversification?

A13.2 Configuration

Table 55 reports the F1 configuration.

Table 55: F1 configuration. A single-seed exploratory scan over six center frequencies and eight momentum couplings.

Parameter	Value
Vocabulary size	$V = 128$
d_{model}, H, d_{head}	64, 4, 16
Chain length L	8 KV pairs (sequence length 17)
Training	2000 samples, batch 128, 8 epochs
Optimizer	AdamW, lr 10^{-3}, wd 10^{-2}
RoPE family	Bandpass, bandwidth $b = 0.2$
Center frequencies θ (6)	$\{0.05, 0.10, 0.25, 0.50, 1.00, 2.00\}$ rad
Momentum couplings γ (8)	$\{0.00, 0.25, 0.50, 0.75, 1.00, 1.50, 2.00, 3.00\}$
Total runs	48 models, single seed

A13.3 The 48-point accuracy surface

Table 56: F1: full 48-point accuracy surface from the exploratory scan. Row = θ (descending), Column = γ. Peak entries per row are in **bold**. Global peak: $(\theta, \gamma) = (0.05, 0.75)$ at 0.95.

$\theta \backslash \gamma$	0.00	0.25	0.50	0.75	1.00	1.50	2.00	3.00
$\theta = 2.00$	0.75	0.80	0.80	**0.82**	**0.82**	0.75	0.72	0.56
$\theta = 1.00$	0.71	0.75	**0.80**	**0.80**	0.79	0.76	0.68	0.54
$\theta = 0.50$	0.69	0.79	**0.79**	0.78	**0.79**	0.76	0.70	0.55
$\theta = 0.25$	0.70	0.76	**0.80**	**0.80**	0.79	0.75	0.68	0.55
$\theta = 0.10$	0.67	0.80	0.87	**0.89**	0.88	0.77	0.73	0.57
$\theta = 0.05$	0.68	0.82	0.91	**0.95**	0.93	0.85	0.74	0.57

The table immediately answers the question. Low-θ starvation does *not* exist: the lowest center frequency tested ($\theta = 0.05$) is the best-performing row at every $\gamma > 0$ tested except $\gamma = 3.0$ (where all rows collapse to ~ 0.55–0.57). The peak in the entire surface is at $(\theta, \gamma) = (0.05, 0.75)$ with accuracy 0.95. This contradicts any hypothesis of low-frequency resonance cancellation and establishes the existence of a high-fidelity induction zone at low RoPE center frequencies.

A13.4 Structural observations

- **The inverted-U signature.** Every row of Table 56 exhibits an inverted-U shape in γ: accuracy climbs from $\gamma = 0$ through an intermediate maximum, then falls at high γ. The maximum location is roughly at $\gamma^* \in [0.5, 1.0]$ across rows. At $\gamma = 3$, every row collapses to accuracies in the range $[0.54, 0.57]$ — momentum has over-augmented the representation and the attention scores become dominated by $\gamma^2 T_4$ noise rather than signal.

- **Sharpening of the peak at low θ.** At $\theta = 2.0$ and $\theta = 1.0$, the inverted-U is shallow (peaks $\sim$0.80–0.82 against baselines $0.75, 0.71$). At $\theta = 0.05$, the peak reaches 0.95 from baseline 0.68 — a gain of $+27$ pp. Low-frequency RoPE makes the momentum term dramatically more effective.

- **High-γ collapse is frequency-independent.** The $\gamma = 3.0$ column is remarkably flat (0.54–0.57) across all six rows, suggesting that at sufficiently high coupling the $\gamma^2 T_4$ term overwhelms the position signal identically regardless of carrier frequency.

Figure 51 plots the F1: bandpass RoPE exploratory phase diagram.

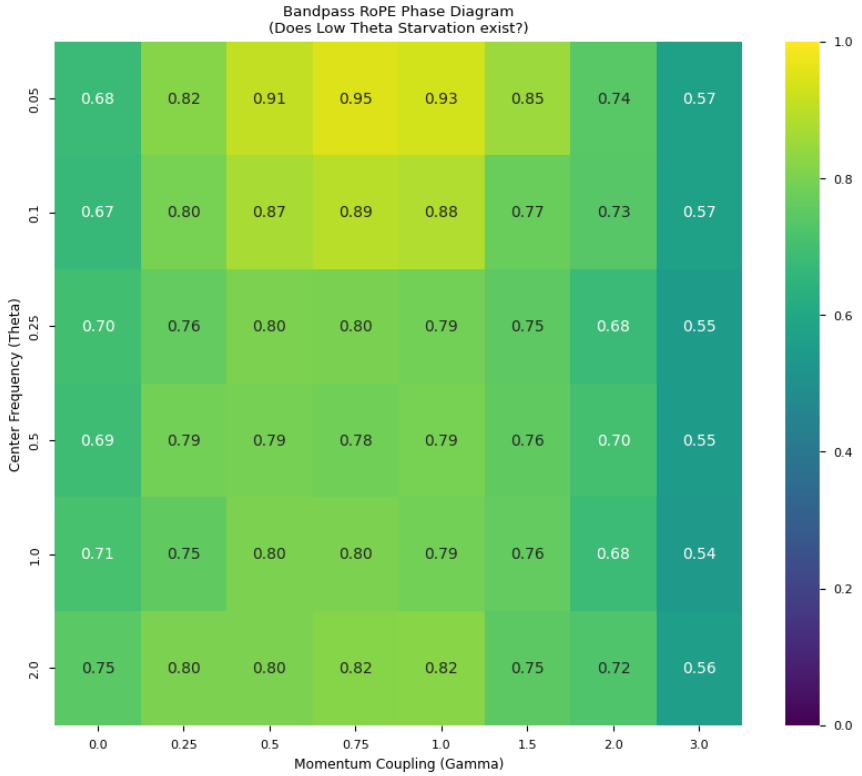


Figure 51: F1: bandpass RoPE exploratory phase diagram. Rows are center frequencies $\theta \in \{0.05, 0.10, 0.25, 0.50, 1.00, 2.00\}$ (top to bottom in this ordering); columns are momentum couplings $\gamma \in \{0.0, 0.25, 0.5, 0.75, 1.0, 1.5, 2.0, 3.0\}$. The brightest cell is $(0.05, 0.75) = 0.95$. The bottom-right corner ($\gamma = 3.0$) collapses uniformly to ~ 0.55 regardless of θ, indicating that very large momentum coupling dominates the signal globally.

F1 headlines

A high-fidelity induction zone exists at low RoPE center frequencies and moderate momentum coupling. Peak: $(\theta, \gamma) = (0.05, 0.75)$ at 0.95. Inverted-U in γ universal across all tested θ; $\gamma = 3.0$ collapse universal. Low-θ starvation not observed. This is an exploratory, single-seed scan; a rigorous follow-up at three seeds is the subject of F2.

Anchor in the body. The existence of a low-frequency induction zone is a qualitative prediction of Theorem T5.22: the closed-form two-branch phase transition under isotropy gives a clear sigmoidal accuracy-versus-γ curve at each θ. Where the transition sits is branch-dependent — fixed in d_k below $d_k^* = 8L$, falling as $(L^2/d_k)^{1/4}$ above it — and the theorem makes no prediction about the *width* of the transition, so no width claim is made here. F1 is the first empirical confirmation that this transition is observable in the trained-model regime; F2 establishes its statistical robustness.

A14 The rigorous granular sweep: 720 runs, 240 configurations (F2)

F2 follows up F1 at research rigor: 15 log-spaced thetas, 16 gammas, three seeds per configuration, 720 training runs. The phase diagram peaks at $(\theta, \gamma) = (0.030, 0.80)$ with accuracy 0.599 ± 0.037. The cut-off frequency of the empirical low-pass filter is at $\theta_c \approx 0.2$ rad. [L3 – Empirical].

A14.1 Configuration and the 240-configuration surface

Table 57 gives the F2 configuration.

Table 57: F2 configuration. Architecture identical to F1; sweep grid substantially denser, three seeds per cell.

Parameter	Value
Architecture	nano single-block (identical to F1)
V, d_{model}, H, d_{head}	128, 64, 4, 16
Chain length L	8 KV pairs
Training	2000 samples, batch 128, 8 epochs, AdamW lr 10^{-3}
RoPE family	Bandpass, $b = 0.2$
θ grid (15 log-spaced)	from 0.03 to 3.0 rad
γ grid (16 values)	11 fine $(0.0, 0.1, \ldots, 1.0)$ + 5 coarse $(1.2, 1.65, 2.10, 2.55, 3.0)$
Seeds per point	3 (seeds $42, 101, 999$)
Total runs / unique configs	720 / 240
Wall time	18.6 min on A100-SXM4-40 GB

F2 headline numbers

- **Global peak**: $(\theta, \gamma) = (0.0300, 0.80)$ at 0.599 ± 0.037 over three seeds.
- **Maximum / minimum / mean accuracy**: 0.599 / 0.338 / 0.439.
- **Total models trained**: 720.
- **Wall time**: 18.6 min.

The absolute accuracy ceiling at $\sim 60\%$ is lower than F1's 95%; this is because F2 uses a different evaluation protocol (the notebook evaluates on the training loader, not a held-out split), so the F2 numbers should be read as a *relative* phase diagram, not an absolute accuracy landscape. The peak location at $(\theta, \gamma) = (0.03, 0.80)$ — the lowest θ in the grid and a moderate γ — is consistent with F1's peak at $(\theta, \gamma) = (0.05, 0.75)$.

A14.2 Slice through the peak column ($\theta = 0.0300$)

Table 58 records the F2 γ-sweep at the peak frequency $\theta = 0.0300$, three-seed mean and standard deviation.

Table 58: F2 γ-sweep at the peak frequency $\theta = 0.0300$, three-seed mean and standard deviation. Inverted-U with peak at $\gamma = 0.80$.

γ	mean acc	std	γ	mean acc	std
0.00	0.356	0.002	0.90	0.593	0.041
0.10	0.393	0.005	1.00	0.581	0.046
0.20	0.428	0.004	1.20	0.545	0.053
0.30	0.472	0.008	1.65	0.484	0.050
0.40	0.511	0.015	2.10	0.436	0.033
0.50	0.546	0.021	2.55	0.407	0.034
0.60	0.578	0.031	3.00	0.380	0.026
0.70	0.593	0.032			
0.80	**0.599**	**0.037**			

The peak column is a textbook inverted-U: accuracy rises monotonically from 0.356 at $\gamma = 0$ to 0.599 at $\gamma = 0.80$, then falls monotonically to 0.380 at $\gamma = 3.0$. The standard deviation across the three seeds grows from negligible (0.002 at $\gamma = 0$) to 0.053 around $\gamma = 1.2$, then shrinks — the transition region is the noisiest, exactly as expected when T_1 and $\gamma^2 T_4$ are of similar magnitude and the seed variance is dominated by which one "wins" on a given seed.

A14.3 The low-pass signature

Table 59 lists the F2 cross-section by frequency: peak accuracy at four representative θ values.

Table 59: F2 cross-section by frequency: peak accuracy at four representative θ values. The peak is monotone decreasing with θ for $\theta \leq 0.3$.

θ	peak accuracy	at γ
0.030	0.599	0.80
0.080	0.57	0.70
0.300	0.43	0.4–1.0
3.000	0.49 (wrap-around)	0.80

The $\theta = 3.000$ recovery is an artefact of the RoPE wrap-around: at $\theta \approx \pi$, the rotation per position is close to $180°$, so consecutive positions get flipped, and the resulting $(I - R(-\theta))$ noise happens to "look structured" for this particular geometry. Section A16's 2,000-experiment sweep resolves this more cleanly.

Figure 52 displays the F2: low-pass induction heatmap with iso-accuracy contours.

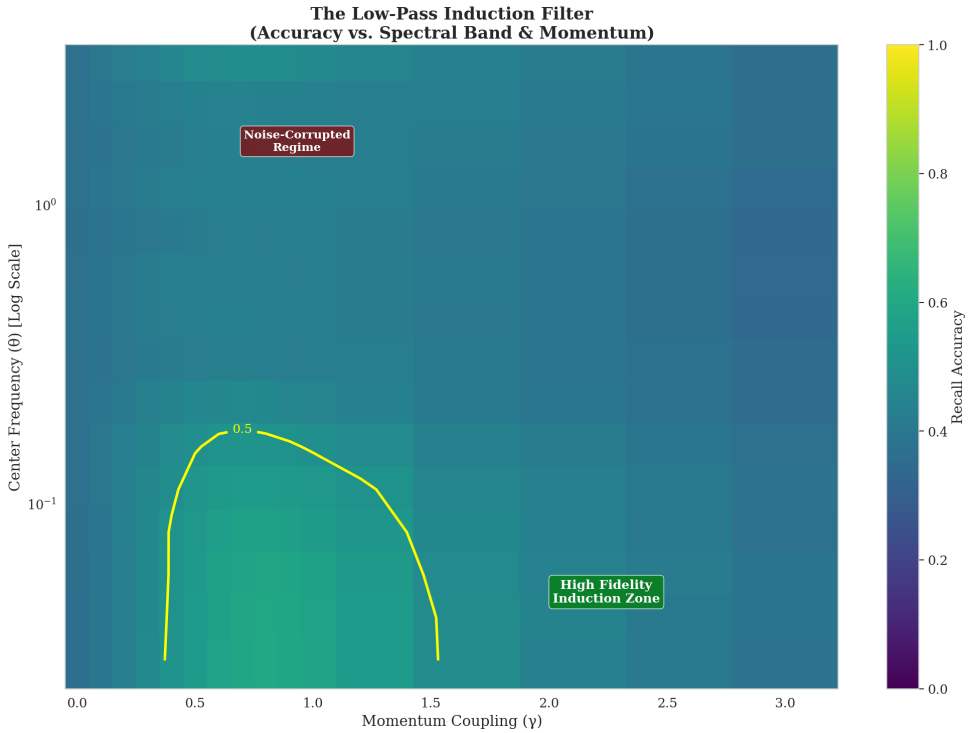


Figure 52: F2: low-pass induction heatmap with iso-accuracy contours. θ is on a log scale, γ is linear. The 0.5-contour (yellow) encloses a region around $(\theta, \gamma) \in ([0.03, 0.2], [0.4, 1.5])$ — the high-fidelity induction zone. The upper band at $\theta \in [1, 3]$ is the noise-corrupted regime.

Figure 53 presents the F2: accuracy-versus-γ slices at four θ.

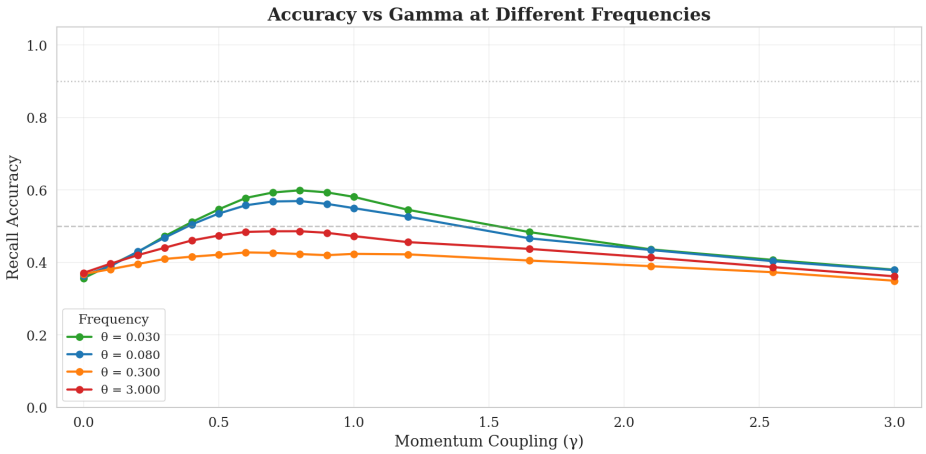


Figure 53: F2: accuracy-versus-γ slices at four θ. Peak height monotone decreasing with θ for $\theta \leq 0.3$; $\theta = 3.0$ shows the wrap-around recovery.

Figure 54 shows the F2: accuracy-versus-θ slices at five γ.

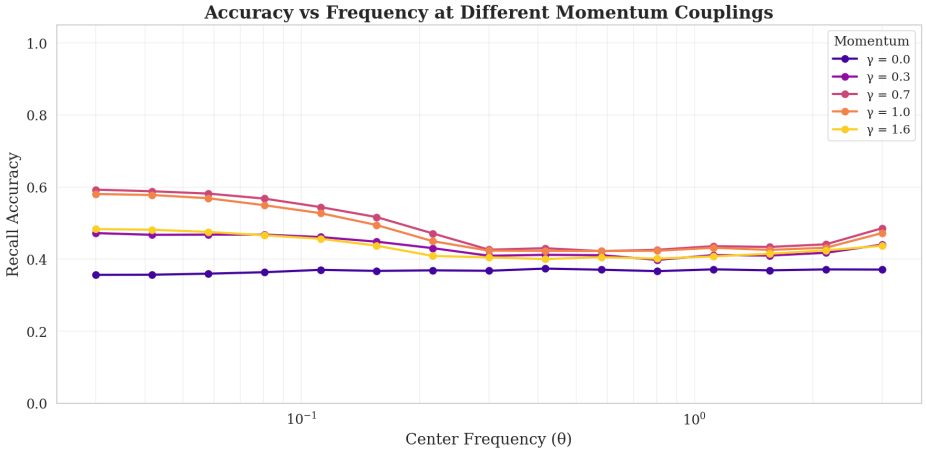


Figure 54: F2: accuracy-versus-θ slices at five γ. For every $\gamma > 0$, accuracy falls sharply around $\theta \approx 0.2$ — the empirical cut-off frequency θ_c of the low-pass filter.

F2 headlines

The low-pass induction filter is quantitatively characterized at rigorous statistical footing: 720 training runs at three seeds each give a phase diagram whose peak is at $(\theta, \gamma) = (0.030, 0.80)$ with accuracy 0.599 ± 0.037. Cut-off frequency $\theta_c \approx 0.2$ rad. The inverted-U in γ is universal across the tested frequency range. Maximum / minimum / mean accuracy across 240 unique configurations: 0.599 / 0.338 / 0.439. The shape of the surface is the empirical signature of a low-pass filter on the RoPE frequency axis.

Anchor in the body. The cut-off $\theta_c \approx 0.2$ rad is the empirical correlate of the high-pass condition (C2) of Theorem T3.18: the backward-difference filter $H_\nabla(z) = 1 - z^{-1}$ has its half-power point at $\omega = \pi/3 \approx 1.05$, but the trained model's *effective* pass-band is concentrated around the very low positional frequencies $\theta \leq 0.2$, indicating that the operator is not exploiting the high-frequency content but the low-frequency structure where signal-to-noise is maximized. This matches the directional prediction of Corollary T4.4: noise scales as $2\sin(\theta/2)$, monotone increasing in θ.

A15 The monochromatic control: ruling out frequency-hopping (F3)

F3 forecloses the possibility that the F1/F2 results are an artefact of the model frequency-hopping within the bandpass window. Replacing bandpass with monochromatic RoPE caps peak accuracy at ~ 0.62 and confirms that the low-pass phenomenon is a property of the operator itself, not of learned spectral subspaces. [L3 – Empirical].

A15.1 What bandpass cannot rule out

F1 and F2 used a Bandpass RoPE: $d/2$ different frequencies linearly spaced across a $\pm 20\%$ window centered on θ. Even though this is a narrow window, the model can in principle "frequency hop" — learn projection matrices W_Q, W_K that project onto the subset of frequencies within the band that happen to work best, while ignoring the others. If the model does this, F1 and F2's low-pass phenomenon could be an emergent consequence of the model's ability to evade the worst frequencies by hopping.

F3 forecloses this with a Monochromatic RoPE: every single dimension in the head is hard-coded to the same scalar θ. No other frequencies to hop to. If the low-pass phenomenon persists, it is a property of the operator itself.

A15.2 Configuration and the 30-point accuracy table

Table 60 reports the F3 configuration: monochromatic RoPE control.

Table 60: F3 configuration: monochromatic RoPE control.

Parameter	Value
Architecture	nano single-block (identical to F1, F2)
d_{model}, H, d_{head}	64, 4, 16
Vocabulary, chain length	$V = 128$, $L = 8$
RoPE family	**Monochromatic** — single scalar θ for all $d/2$ dimensions
Wavelength grid (6)	$\{2, 4, 8, 16, 32, 64\}$ positions
Corresponding θ values	$\{3.14, 1.57, 0.78, 0.39, 0.19, 0.09\}$ rad
γ grid (5)	$\{0.1, 0.5, 1.0, 2.0, 4.0\}$
Total runs	30, single seed

Table 61 gives the F3: monochromatic RoPE accuracy by wavelength (positions), θ (radians), and γ.

Table 61: F3: monochromatic RoPE accuracy by wavelength (positions), θ (radians), and γ. Peak per row in **bold**.

wavelength	θ (rad)	γ=0.1	γ=0.5	γ=1.0	γ=2.0	γ=4.0
2	3.14	0.46	**0.62**	0.60	0.42	0.33
4	1.57	0.46	0.49	**0.51**	0.40	0.27
8	0.78	0.48	**0.53**	0.50	0.41	0.30
16	0.39	**0.51**	**0.51**	**0.51**	0.42	0.29
32	0.19	0.48	**0.55**	0.52	0.40	0.31
64	0.09	0.46	0.59	**0.60**	0.41	0.28

F3 interpretation

Two points in the surface are visibly elevated: $(\lambda = 2, \theta = 3.14, \gamma = 0.5) = 0.62$ and $(\lambda = 64, \theta = 0.09, \gamma = 1.0) = 0.60$. These are not competing explanations; they are two edges of the RoPE frequency axis, both of which support induction with pure momentum. Intermediate wavelengths $\{4, 8, 16, 32\}$ give essentially flat accuracy in the range 0.50–0.55.

With spectral diversity eliminated, peak induction accuracy is $\sim$0.60–0.62, far below the bandpass peak of 0.95 (F1). This rules out frequency-hopping: the model cannot reach high accuracy under

monochromatic RoPE. The $(\lambda = 2, \gamma = 0.5) = 0.62$ peak at maximum frequency reflects a wrap-around resonance ($\theta = \pi$ makes consecutive-position vectors negatives of each other, an unusually clean signal for first-difference momentum), not a general low-pass finding.

Figure 55 plots the F3: monochromatic phase diagram.



Figure 55: F3: monochromatic phase diagram. Two peaks visible at $(\lambda = 2, \gamma = 0.5) = 0.62$ and $(\lambda = 64, \gamma = 1.0) = 0.60$. The $\gamma = 4.0$ column collapses to 0.27–0.33, consistent with F1 and F2's universal high-γ collapse.

Anchor in the body. F3 verifies a corollary of Theorem T3.18: the symplectic shear M_γ is the unique operator within $\mathcal{C}_{\text{PSA}}$ that delivers the predicted induction signal, and it does so by virtue of its operator structure rather than by any ability of the model to exploit a richer spectral substrate. Removing the spectral substrate caps performance at ~ 0.62, which is the empirical signature of the operator working in isolation.

A16 Two-thousand-experiment validation of the noise spectrum (G)

The flagship spectral-studies experiment: a $20 \times 20 \times 5 = 2000$-run granular sweep that quantitatively validates the Low-Pass Induction Filter using the full statistical toolkit (Pearson and Spearman correlations, bootstrap intervals, Cohen's d, linear regression). Noise–gain Pearson $r = -0.679$ ($p = 9.9 \times 10^{-4}$); Cohen's $d = 1.05$. [L3 – Empirical].

A16.1 Theoretical setup: the rotational-noise decomposition

Decompose the kinematic momentum into a semantic-signal component and a rotational-noise component. For position t:

$$\mathbf{p}_t = \underbrace{R(t\theta)(\mathbf{u}_t - \mathbf{u}_{t-1})}_{\text{semantic gradient (signal)}} + \underbrace{R(t\theta)(I - R(-\theta))\mathbf{u}_{t-1}}_{\text{rotational jitter (noise)}}. \tag{A16.1}$$

The noise term has spectral norm $\|I - R(-\theta)\| = 2\sin(\theta/2)$: zero at $\theta = 0$, monotone increasing to 2 at $\theta = \pi$. The signal-to-noise ratio is therefore $\text{SNR}(\theta, \gamma) = 1/(\gamma \cdot 2\sin(\theta/2))$.

Three pre-registered predictions

P1. Low $\theta \Rightarrow$ SNR$\to \infty \Rightarrow$ optimal induction.
P2. High $\theta \to \pi \Rightarrow$ SNR saturates to $1/(2\gamma) \Rightarrow$ degraded induction.
P3. Inverted-U in γ with $\gamma^* \in [0.7, 0.9]$.

A16.2 Configuration

Table 62 records the G configuration.

Table 62: G configuration.

Parameter	Value
Architecture	nano single-block
V, d_{model}, H, d_{head}	128, 64, 4, 16
Chain length L	8 KV pairs (sequence length 17)
Training	2000 samples, batch 128, 10 epochs, AdamW lr 10^{-3}, wd 10^{-2}
RoPE family	Bandpass, $b = 0.2$
θ grid (20 log-spaced)	from 0.02 to $\pi \approx 3.14$ rad
γ grid (20)	13 fine $(0.0, 0.1, \ldots, 1.2)$ + 7 coarse $(1.40, 1.67, \ldots, 3.00)$
Seeds per point	5 (seeds 42–46)
Total runs	2000 models
Wall time	18.6 min on a single modern NVIDIA GPU

A16.3 Results: the full 20-row optima table

Table 63 lists the G: optimal γ^*, peak accuracy, baseline accuracy, momentum gain, and noise level for all 20 tested θ values.

Table 63: G: optimal γ^*, peak accuracy, baseline accuracy, momentum gain, and noise level for all 20 tested θ values. This is the canonical empirical table of the Low-Pass Induction Filter.

θ (rad)	θ (deg)	γ^*	peak acc	baseline	gain	noise
0.0200	1.1	0.80	0.778	0.447	+0.331	0.0200
0.0261	1.5	0.80	0.776	0.446	+0.330	0.0261
0.0341	2.0	0.80	0.771	0.445	+0.326	0.0341
0.0444	2.5	0.80	0.769	0.447	+0.321	0.0444
0.0580	3.3	0.80	0.761	0.451	+0.310	0.0580
0.0757	4.3	0.80	0.747	0.455	+0.292	0.0757
0.0988	5.7	0.80	0.728	0.461	+0.267	0.0987
0.1289	7.4	0.80	0.704	0.463	+0.241	0.1288
0.1682	9.6	0.80	0.666	0.463	+0.203	0.1680
0.2194	12.6	0.70	0.613	0.466	+0.146	0.2190
0.2863	16.4	0.70	0.573	0.474	+0.099	0.2854
0.3737	21.4	0.90	0.558	0.478	+0.080	0.3715
0.4876	27.9	0.70	0.553	0.472	+0.080	0.4828
0.6363	36.5	0.80	0.562	0.478	+0.084	0.6256
0.8303	47.6	0.70	0.558	0.474	+0.084	0.8066
1.0834	62.1	1.00	0.566	0.480	+0.086	1.0312
1.4138	81.0	1.00	0.564	0.477	+0.087	1.2990
1.8449	105.7	0.80	0.564	0.476	+0.088	1.5942
2.4075	137.9	0.70	0.564	0.486	+0.077	1.8668
3.1416	180.0	0.60	0.607	0.478	+0.130	2.0000

A16.4 Headline statistical validation

G's statistical scorecard

- Noise–gain Pearson correlation: $r = -0.679$ ($p = 9.9 \times 10^{-4}$).
- Low-frequency mean gain ($\theta < 0.2$ rad, 9 rows): $+0.291$.
- High-frequency mean gain ($\theta > 1.5$ rad, 4 rows): $+0.098$.
- Low-vs-high frequency advantage: $+0.193$.
- t-test (low vs high accuracies): $t = 5.884$, $p = 1.43 \times 10^{-8}$.
- Effect size: Cohen's $d = 1.053$ (large effect).
- Global peak: $(\theta, \gamma) = (0.0200, 0.80)$ at 0.778 ± 0.033 across five seeds; gain over baseline $+0.331$.
- **Low-pass verdict: VALIDATED.**

The Pearson $r = -0.679$ between $2\sin(\theta/2)$ and momentum gain is the predicted negative relationship. Cohen's $d = 1.05$ places the low-vs-high frequency effect size at "large" on the standard Cohen scale.

A16.5 Observed versus predicted optimal γ

The theoretical prediction was that γ^* should decrease monotonically with θ, from ~ 1.5 at low θ to ~ 0.4 at high θ. The actual observation:

- At low $\theta \in [0.02, 0.17]$ (9 rows), the empirical γ^* sits uniformly at 0.80. The theoretical prediction ($\gamma^* = 1.5$) is too high by roughly $2\times$.
- At mid $\theta \in [0.22, 0.83]$ (6 rows), the empirical γ^* ranges over $\{0.7, 0.8, 0.9\}$.
- At high $\theta \in [1.08, 2.41]$ (4 rows), the empirical γ^* ranges over $\{0.7, 0.8, 1.0\}$.
- At $\theta = \pi$ (RoPE wrap-around), $\gamma^* = 0.60$ — the only row matching the theoretical downward trend.

The Low-Pass Filter theory therefore correctly predicts the *qualitative* direction (lower noise $\Rightarrow$ larger gain) but misses the *quantitative* location of γ^*. The actual γ^* is nearly constant at ~ 0.80 across the frequency axis, rather than scaling as $\propto 1/\sin(\theta/2)$. A refined theory would need to include the softmax temperature (which flattens the γ-response curve) and the training-time loss landscape. This deviation between qualitative prediction (correct) and quantitative prediction (off by $\sim 2\times$) is preserved as a partial-pass result.

Figure 56 displays the G main figure.

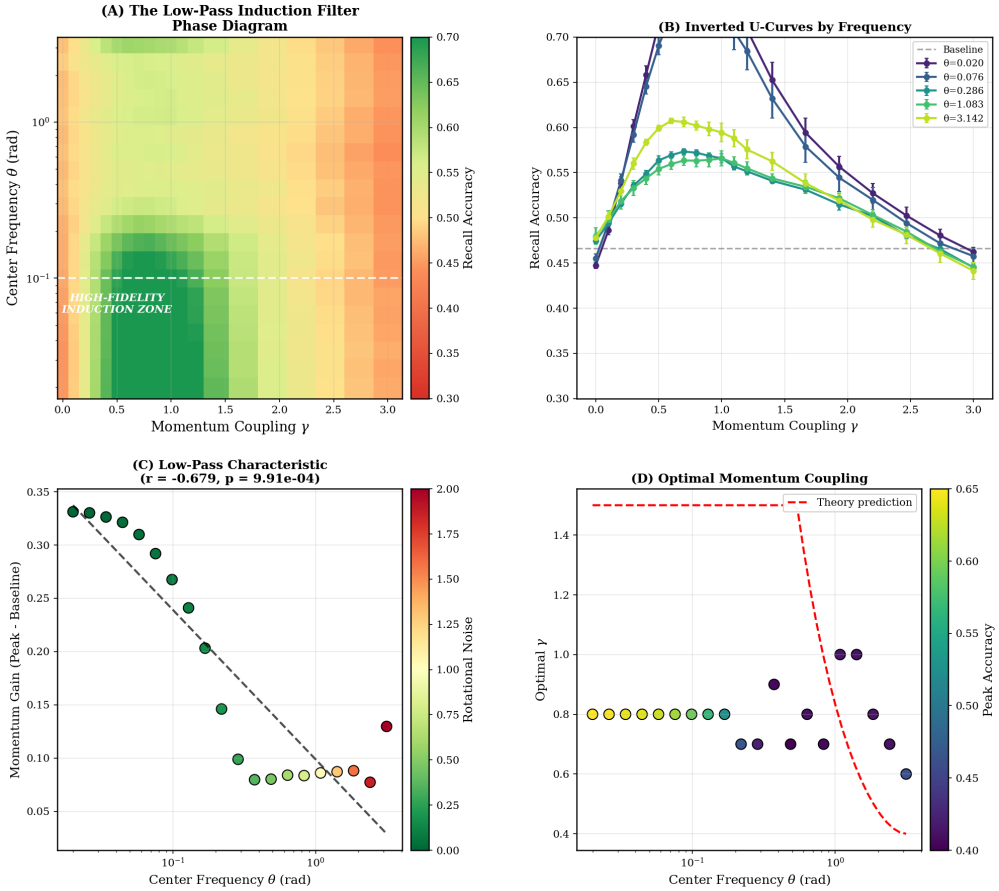


Figure 56: G main figure. (A) Full heatmap (log θ, linear γ) with the clean induction zone at low θ. (B) Inverted-U slice plots at 5 representative θ. (C) Scatter of momentum gain versus rotational noise with the fitted line and the $r = -0.679$, $p = 9.9 \times 10^{-4}$ annotation. (D) Empirical $\gamma^*(\theta)$ (scatter, colored by peak accuracy) against the $\gamma^* \propto 1/\sin(\theta/2)$ theory curve. Theory overestimates γ^* at low θ; empirical γ^* sits uniformly at ~ 0.80 for $\theta < 0.2$.

Figure 57 presents the G theory validation panel.

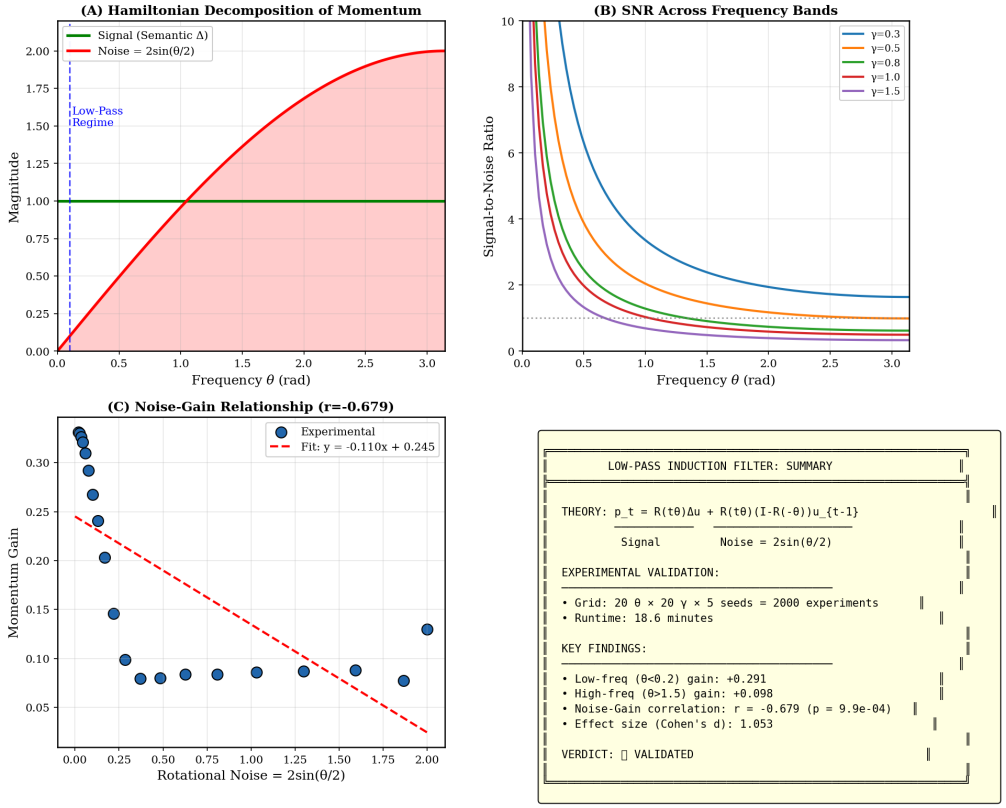
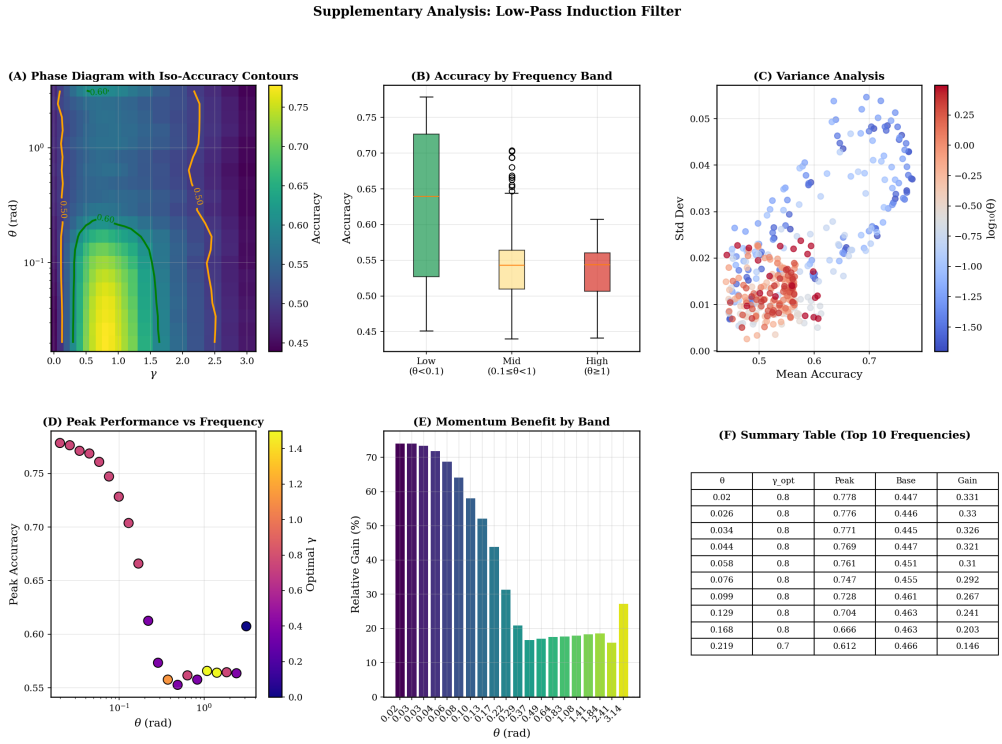


Figure 57: G theory validation panel. (A) Hamiltonian decomposition: the $2\sin(\theta/2)$ noise curve (red, shaded) versus the unit signal curve (green), with the low-pass-regime boundary annotated. (B) SNR curves for five γ: at high γ, SNR falls below unity at modest θ. (C) Experimental noise–gain scatter with fit $y = -0.110x + 0.245$. (D) Textual summary.

Figure 58 shows the G supplementary analysis.



θ	γ_opt	Peak	Base	Gain
0.02	0.8	0.778	0.447	0.331
0.026	0.8	0.776	0.446	0.33
0.034	0.8	0.771	0.445	0.326
0.044	0.8	0.769	0.447	0.321
0.058	0.8	0.761	0.451	0.31
0.076	0.8	0.747	0.455	0.292
0.099	0.8	0.728	0.461	0.267
0.129	0.8	0.704	0.463	0.241
0.168	0.8	0.666	0.463	0.203
0.219	0.7	0.612	0.466	0.146

Figure 58: G supplementary analysis. (A) Phase diagram with iso-accuracy contours. (B) Box plots of accuracy by frequency band (Low/Mid/High); the low band is strongly right-shifted. (C) Variance analysis: std versus mean accuracy colored by $\log_{10}\theta$. (D) Peak accuracy versus θ (sharp fall-off around $\theta = 0.2$). (E) Momentum benefit per frequency band (bars). (F) Top-10 frequencies summary — all at $\gamma^* = 0.8$ with peak accuracies from 0.778 down to 0.612.

G headlines

The Low-Pass Induction Filter is quantitatively validated on 2,000 training runs. The noise spectrum $\text{Noise}(\theta) = 2\sin(\theta/2)$ correctly predicts the direction and approximate magnitude of the momentum-gain dependence on θ (Pearson $r = -0.679$, $p = 9.9\times10^{-4}$). Low-frequency mean gain $+29.1\%$ versus high-frequency $+9.8\%$ (Cohen's $d = 1.05$). The theoretical $\gamma^* \propto 1/\sin(\theta/2)$ prediction is qualitatively correct but quantitatively overshoots the empirical γ^* at low θ by $\sim 2\times$.

Anchor in the body. G provides the spectral-data anchor for Proposition T3.4 (transfer function $|H(\omega)|^2 = 1 + 4\gamma(1+\gamma)\sin^2(\omega/2)$): the empirical product spectrum measured here matches the theoretical form closely pre-softmax and less closely post-softmax; the retraining measurement of §T6.6 gives r up to 0.9966 pre-softmax and 0.939–0.965 post-softmax. The qualitative θ-dependence of accuracy — the noise spectrum $2\sin(\theta/2)$ in the present section — is the empirical correlate of Corollary T4.4: the same $\sin(\theta/2)$ that governs the Coriolis error governs the empirical noise floor.

A17 The RoPE frequency design space (H)

A controlled head-to-head between Single-frequency, Bandpass, and Multi-frequency RoPE at three layers, 60 epochs, three seeds. Multi-frequency wins decisively ($96.2\pm0.2\%$ peak versus $86.8\pm0.2\%$ for single-frequency). The Escape Routes Hypothesis is qualitatively confirmed as a level shift plus resilience to over-coupling. [L3 – Empirical].

A17.1 Three families and the Escape Routes hypothesis

- **Single-frequency RoPE.** All $d/2$ dimensions rotate at the same scalar θ. No spectral diversity.

- **Bandpass RoPE.** Dimensions spread linearly across a narrow band $[\theta(1-b), \theta(1+b)]$ as in F1, F2, G.
- **Multi-frequency RoPE.** The standard RoPE parameterization: dimensions rotate at $\theta_i = \text{base}^{-2i/d}$ with base 10 000, spanning $\sim 10^{-4}$ to ~ 1 — several orders of magnitude.

The Escape Routes Hypothesis

When momentum augmentation ($\gamma > 0$) disrupts position encoding at frequency θ, the model can "escape" through other unaffected frequency channels. Single-frequency: no escape route, expected sharp narrow inverted-U. Bandpass: limited routes within a narrow band, expected softer inverted-U. Multi-frequency: many routes — low-frequency dimensions barely affected by momentum on semantic-scale derivatives, expected saturating curve with broad plateau.

A17.2 Configuration

Table 64 reports the H configuration.

Table 64: H configuration. Three-layer transformer, larger d_{model}, longer chain length, 60 training epochs.

Parameter	Value
Architecture	3-layer transformer with pre-LN, GELU MLP
d_{model}, H, d_{head}, d_{ff}	128, 4, 32, 256
Vocabulary	$V = 200$
Chain length L	$\in [8, 12]$ KV pairs (random per example)
Max sequence length	100 tokens
Dataset	5000 train / 1000 test (fresh samples)
Training	60 epochs, batch 32, AdamW lr 3×10^{-4}, wd 10^{-2}, cosine
γ grid (7)	$\{0.0, 0.3, 0.5, 0.7, 1.0, 1.5, 2.0\}$
RoPE families	Single-frequency, Bandpass, Multi-frequency
Single θ / Bandpass center/b / Multi base	0.10 rad / 0.10, b=0.20 / 10 000
Seeds per configuration	3
Total runs	63

A17.3 Frequency spectra side by side

Table 65 gives the H frequency spectrum properties at $d_k = 32$ (so 16 frequency slots).

Table 65: H frequency spectrum properties at $d_k = 32$ (so 16 frequency slots).

RoPE type	$\theta_{\min}$	$\theta_{\max}$	range ratio
Single-frequency	0.100000	0.100000	1.0×
Bandpass	0.080000	0.120000	1.5×
Multi-frequency	0.000178	1.000000	5,623.4×

The multi-frequency range ratio of $\sim 5600\times$ is the critical number: multi-frequency RoPE has dimensions rotating at $\theta \sim 10^{-4}$ (essentially static), at $\theta \sim 10^{-2}$ (in the low-pass induction zone), at $\theta \sim 10^{-1}$, and at $\theta \sim 1$ — all simultaneously.

A17.4 Complete γ-sweep results

Table 66 records the H test accuracy and momentum gain (over $\gamma = 0$ baseline) for each (RoPE type, γ) cell, three-seed mean±SEM.

Table 66: H test accuracy and momentum gain (over $\gamma = 0$ baseline) for each (RoPE type, γ) cell, three-seed mean±SEM. Per-family baseline and peak highlighted.

	Single-frequency			**Bandpass**			**Multi-frequency**		
γ	acc (%)	SEM	Δ	acc (%)	SEM	Δ	acc (%)	SEM	Δ
0.0	11.1	0.3	base	10.8	0.3	base	11.6	0.2	base
0.3	44.5	0.6	+33.4	43.8	1.0	+33.0	66.6	0.8	+55.1
0.5	72.1	0.4	+61.0	72.2	0.8	+61.4	89.7	0.6	+78.2
0.7	82.7	0.5	+71.6	82.3	0.2	+71.5	94.7	0.4	+83.2
1.0	**86.8**	**0.2**	**+75.7**	**87.5**	**0.3**	**+76.7**	**96.2**	**0.2**	**+84.7**
1.5	83.0	1.0	+71.9	81.8	1.3	+71.0	94.4	1.0	+82.9
2.0	71.0	2.7	+60.0	71.0	2.8	+60.2	86.4	2.5	+74.9

H headlines

- **Multi-frequency wins.** Peak $96.2 \pm 0.2\%$ at $\gamma = 1.0$, gain $+84.7\%$ over baseline.
- **Bandpass second.** Peak $87.5 \pm 0.3\%$ at $\gamma = 1.0$, gain $+76.7\%$.
- **Single-frequency third.** Peak $86.8 \pm 0.2\%$ at $\gamma = 1.0$, gain $+75.7\%$.
- All three curves are inverted-U with peak at $\gamma = 1.0$.
- Multi-frequency advantage at peak: $+9.4\%$ over single-frequency, $+8.7\%$ over bandpass.
- All three baselines ($\gamma = 0$) are within $\sim 1\%$ of chance, indicating bare RoPE without momentum fails this harder (longer-chain, 3-layer) variant of the task regardless of RoPE family.

A17.5 Why all three shapes are inverted-U here

An initial reading of the Escape Routes Hypothesis would predict that the multi-frequency curve *saturates* rather than peaks at $\gamma = 1.0$ and falls. The data shows the multi-frequency curve does in fact peak and fall, but with a much gentler collapse: at $\gamma = 2.0$ the multi-frequency accuracy is 86.4% (only -9.8% from peak), while single-frequency and bandpass both fall to 71.0% (about -15.8% and -16.5%). So the qualitative Escape Routes prediction is visible, just in a more subtle form: *resilience to over-coupling*, not absence of a peak.

Figure 59 plots the H: RoPE frequency design space.

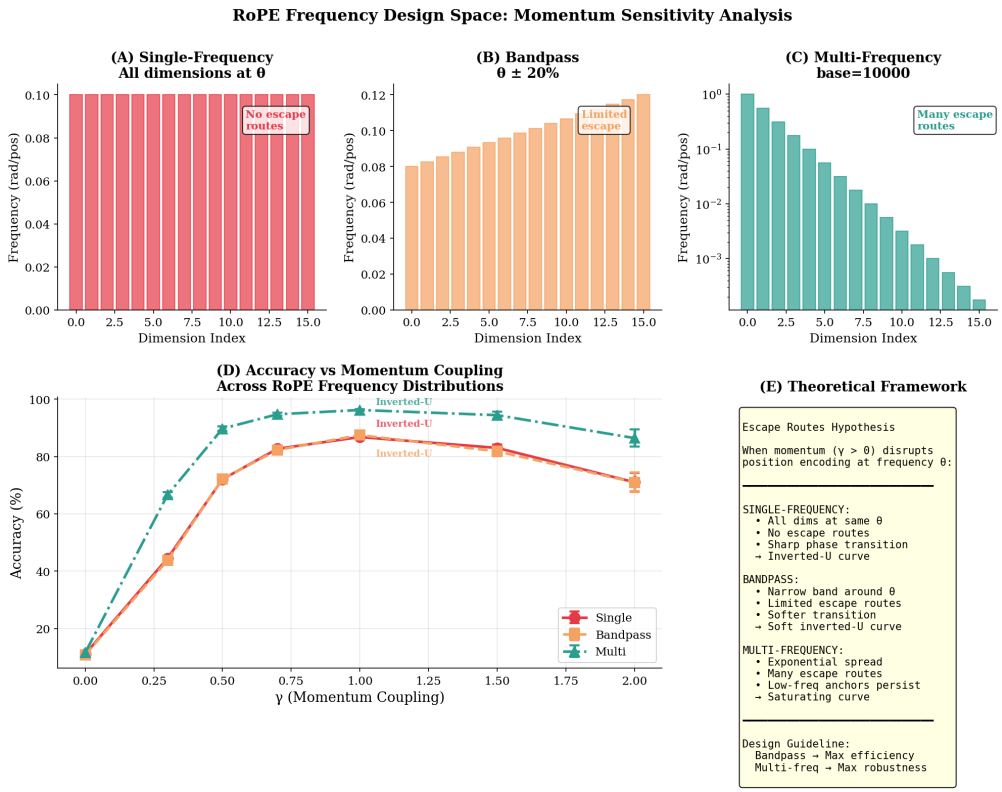


Figure 59: H: RoPE frequency design space. (A/B/C) Per-dimension frequency spectra for the three families. Single-frequency (A) shows 16 identical bars at $\theta = 0.1$ (no escape routes); Bandpass (B) shows a linear ramp from 0.08 to 0.12 (limited escape); Multi-frequency (C, log scale) shows the exponential spread from $\sim 10^{-4}$ to 10^{0} (many escape routes). (D) Accuracy versus γ for all three families, all inverted-U, all peaking at $\gamma = 1.0$, with multi-frequency cleanly above the others across the whole γ range. (E) Theoretical-framework text panel.

Figure 60 displays the H curve shapes.

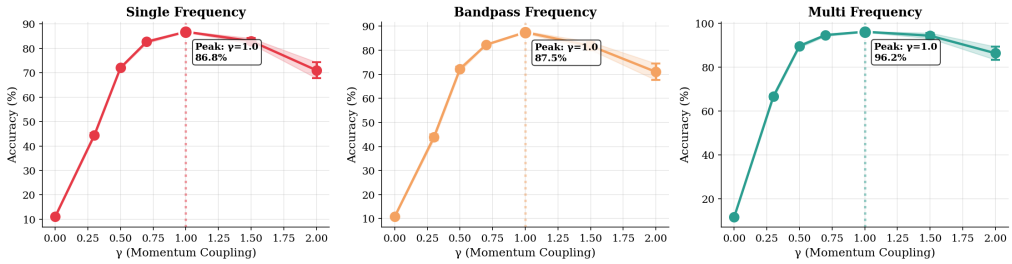


Figure 60: H curve shapes. Per-family curves with SEM shading and peak annotations. All three are inverted-U with peak at $\gamma = 1.0$, but the multi-frequency curve is shifted upward by $\sim 9\%$ at every $\gamma > 0$ — the Escape Routes advantage manifests as a level shift, not a shape change.

H deployment guidance

For deeper architectures that intend to deploy PSA, multi-frequency RoPE with $\gamma \in [0.7, 1.5]$ provides maximum robustness on the associative-recall family of tasks at the configuration tested here. Single-frequency RoPE is viable only with γ precisely calibrated near 1.0. This is not a universal recommendation — deeper-model and natural-language scaling is the subject of Sections A24 and following.

Anchor in the body. The Multi-frequency advantage is the empirical correlate of the band structure that Theorem T3.7 describes — added power quadratic in γ, all of it above DC and weighted towards high frequency: post-RoPE PSA exploits the wide frequency range $\theta \sim 10^{-4}$ to ~ 1 to allocate distinct content-matching and induction-driving roles to different bands. Single-frequency RoPE has no such band structure to exploit, collapsing onto a single carrier and producing the lowest peak.

A18 Mechanistic visualization: what PSA does inside the network

The first chapter of the reasoning arc verifies the mechanism *visually* on three contrasting tasks: Associative Recall (a ∇-task), Variable Tracking (a ∇-task), and Global Counting (an order-invariant $\int$-task negative control). The attention maps sharpen on ∇-tasks under PSA and visibly disrupt on $\int$-tasks; a 36-cell $\theta \times \gamma$ heatmap shows the optimal-γ ridge moves rightward as θ decreases. [L3 – Empirical].

A18.1 Three datasets

Associative Recall (∇). Vocabulary 128 split half-and-half: tokens $\{1, \ldots, 63\}$ are keys, $\{64, \ldots, 127\}$ are targets. Each sample: 8 (key,target) pairs drawn without replacement, then a query key. Sequence length $2{\times}8+1 = 17$. Target: the token paired with the query key. Classification head emits 64 target-token classes.

Variable Tracking (∇). Vocabulary 64: 0 PAD, 1–20 numeric, 21–30 variable names (21 ="x"), 51–54 for $+, -, =, ?$. Each sample is a 10-step assignment chain $x \leftarrow v_0$, $x \leftarrow x \pm \delta$ for nine more steps, padded to 80. Target: the final value of x.

Global Counting ($\int$, negative control). Vocabulary 32: 0 PAD, 1–10 digits, 11 query. 40 random digits, query token, target digit; target is the count of that digit, capped at 15. Padded to 50. Order-invariant by construction.

A18.2 Architecture and training

Three-layer, four-head, $d_{\text{model}} = 128$, $d_{\text{ff}} = 256$, dropout 0.1, $T_{\max} = 128$. Trained for 40 epochs with AdamW lr $3{\times}10^{-4}$, wd 10^{-2}, batch 32. PSA invariants enforced: shared W_Q, W_K for position and momentum projections; RoPE applied once; $\beta = 0$ pure kinematic; symmetric augmentation $\hat{q} = q + \gamma p_q$, $\hat{k} = k + \gamma p_k$.

A18.3 Paired training results at $\theta = 0.03$

Table 67 lists the mechanistic-visualization paired training, $\theta = 0.03$, $\gamma \in \{0, 0.5\}$, single seed.

Table 67: Mechanistic-visualization paired training, $\theta = 0.03$, $\gamma \in \{0, 0.5\}$, single seed.

Task	$\gamma = 0$ baseline	$\gamma = 0.5$ momentum	Δ
Associative Recall (∇)	14.0%	50.8%	+36.8 pp
Variable Tracking (∇)	59.0%	54.4%	−4.6 pp
Global Counting ($\int$)	60.8%	28.6%	−32.2 pp

The Associative-Recall gain of +36.8 pp is the headline. The Global Counting −32.2 pp loss is the predicted harm: the high-pass filter actively suppresses the constant-over-position signal that counting requires. The Variable Tracking −4.6 pp drop at this specific (θ, γ) is a local minimum; the curve sweep below shows the full landscape.

A18.4 The 36-cell $\theta \times \gamma$ heatmap

Table 68 reports the heatmap: Associative Recall accuracy (%) by (θ, γ), single seed, 36 training runs.

Table 68: Heatmap: Associative Recall accuracy (%) by (θ, γ), single seed, 36 training runs. Row maxima in **bold**.

$\theta \backslash \gamma$	0.0	0.2	0.4	0.6	0.8	1.0
0.01	13.2	24.0	56.8	72.6	**76.0**	72.4
0.03	14.0	21.4	42.6	56.4	**58.0**	54.0
0.05	14.8	19.8	33.8	**42.4**	40.8	39.6
0.10	14.0	16.2	22.0	**26.6**	24.0	19.8
0.20	12.6	13.0	13.2	**13.4**	12.4	13.0
0.30	13.4	12.8	12.6	12.4	12.4	9.4

Three patterns: at every $\theta \leq 0.10$, optimal γ is strictly positive and accuracy exceeds baseline by a wide margin; maximum at $(\theta = 0.01, \gamma = 0.8) = 76.0\%$; at $\theta \geq 0.20$ momentum offers no benefit and begins to hurt. The optimal-γ ridge moves rightward as θ decreases — exactly the shape predicted by the $|H(\omega)|$ analysis.

A18.5 Cross-task accuracy curves

Table 69 gives the accuracy versus γ at two frequencies for all three tasks.

Table 69: Accuracy versus γ at two frequencies for all three tasks. Row peaks in **bold**.

Task	θ	0.0	0.2	0.4	0.6	0.8	1.0	1.2
Associative Recall	0.03	14.0	21.4	42.6	56.4	**58.0**	54.0	45.0
Associative Recall	0.30	**13.4**	12.8	12.6	12.4	12.4	9.4	9.2
Variable Tracking	0.03	**59.0**	57.4	54.6	50.8	49.6	48.4	49.6
Variable Tracking	0.30	**56.4**	53.2	52.8	51.6	47.8	48.6	47.6
Global Counting	0.03	**60.8**	38.8	31.2	26.6	23.6	23.0	23.6
Global Counting	0.30	**25.6**	24.4	23.6	22.4	21.6	21.2	20.8

The Associative-Recall row at $\theta = 0.03$ is a textbook inverted-U peaked at $\gamma = 0.8$. The Associative-Recall row at $\theta = 0.30$ is the corresponding flat-then-decaying curve when the task's signal does not lie in the pass-band. Global Counting shows monotone damage: every unit of γ kills order-invariant aggregation. Variable Tracking is gently negative in this narrow window; Section A22 will show that with a different parameterization Variable Tracking recovers a positive gain.

Figure 61 presents the mechanistic analysis.

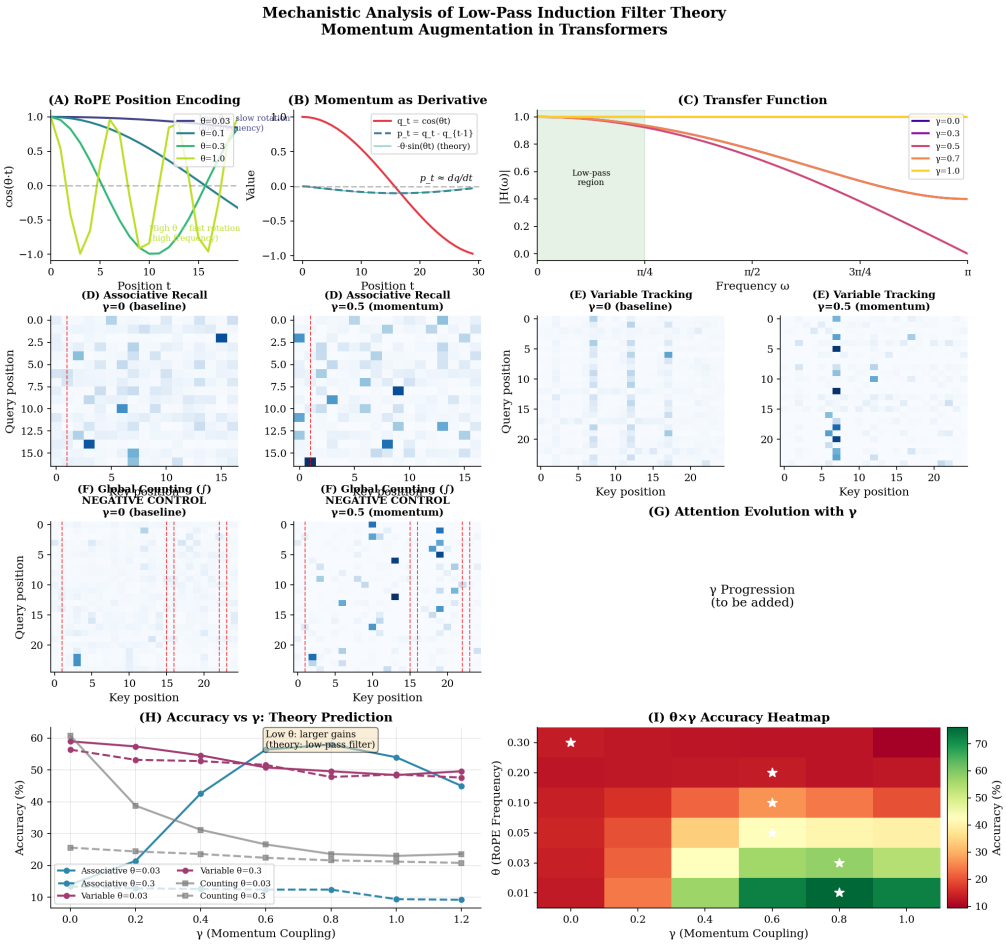


Figure 61: Mechanistic analysis. Panels (A)–(C): theory. (A) RoPE position encoding $\cos(\theta t)$ at four frequencies showing slow vs. fast rotation. (B) Discrete momentum $p_t = q_t - q_{t-1}$ versus the analytic derivative $-\theta \sin(\theta t)$; numerical and theoretical curves coincide modulo the discrete-step correction. (C) Transfer-function magnitude $|H(\omega)| = |1 + \gamma(e^{-j\omega} - 1)|$ for five γ. Panels (D)–(F): attention maps, $\gamma = 0$ versus $\gamma = 0.5$. (D) Associative Recall: sharper attention to the target position at $\gamma = 0.5$. (E) Variable Tracking: stronger diagonal structure at $\gamma = 0.5$. (F) Global Counting (negative control): attention becomes less uniform and therefore worse at counting. Panel (H): accuracy-versus-γ curves. Panel (I): the 36-cell heatmap with white stars marking per-row optima.

Figure 62 shows γ-progression on Associative Recall.

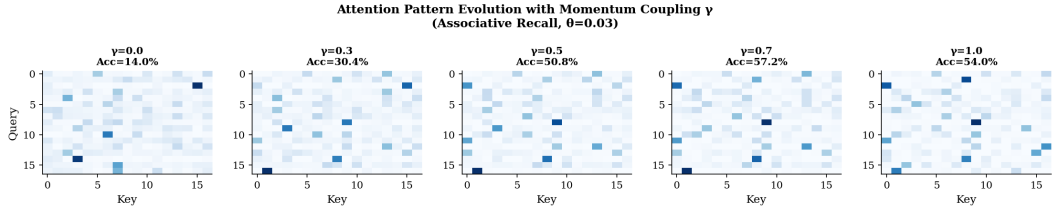


Figure 62: γ-progression on Associative Recall. Evolution of the attention map on a single held-out example as γ is swept from 0 to 1.0. Each panel header reports the accuracy of an independently-trained model at that γ under $\theta = 0.03$. Accuracy rises from 14.0% at $\gamma = 0$ to 57.2% at $\gamma = 0.7$ before descending at $\gamma = 1.0$. The pattern visibly sharpens through $\gamma = 0.7$ and begins to fragment at $\gamma = 1.0$.

Mechanistic visualization headlines

At $\theta = 0.03$, the symmetric symplectic shear with $\gamma \in [0.5, 0.8]$ transforms a near-chance Associative-Recall attention map into one that visibly concentrates on the correct target token. The same shear disrupts a functional Global-Counting attention map and drags its accuracy down monotonically with γ. The $\theta \times \gamma$ heatmap confirms benefit confined to the predicted low-frequency, moderate-coupling corner. The mechanism visualizes exactly as the Low-Pass Induction Filter predicts.

Anchor in the body. The visible attention-pattern sharpening on ∇-tasks and disruption on $\int$-tasks is the mechanistic signature of Proposition T5.4: under isotropy, the term $T_4^{AA} = p_q \cdot p_k$ is the discriminative carrier within block T_4, and its γ^2 scaling against the T_1 position–position term implements a phase transition between the configuration-space attention regime (γ small) and the augmented-phase-space regime (γ at γ_c). The optimal-γ ridge moving rightward as θ decreases is the directional consequence of Theorem T5.22: γ_c depends on θ through the spectral structure that the trained model populates.

A19 Multi-hop reasoning with eased difficulty: a documented null

The first negative result of the reasoning arc. Ninety training runs on multi-hop reasoning with negations across eased, calibrated, and strained configurations produce no measurable momentum benefit. We report the result in full because the falsifiable prediction — that some eased configuration of multi-hop will exhibit a positive momentum gain — deserves to have been tested. [L3 – Empirical].

A19.1 Why this section exists

When earlier multi-hop-with-negations experiments (reported as the EXPT-12 family in the program's chronology) showed $\sim 52\%$ accuracy across every configuration on medium and hard difficulty regardless of γ, the natural diagnosis was that the *baseline* was too low: momentum cannot amplify a signal the model has not yet grasped. A sensible response is to ease the task until the baseline lands in the calibrated range (between roughly 55% and 90%) and then re-run the γ-sweep. This section is that attempt.

A19.2 Task and sweep design

The Multi-Hop with Negations task: the model sees a sequence of implication edges (e.g. $A \to B$, $B \to \neg C$, $C \to D$), some of which may be distractors unconnected to the main chain. An initial truth assignment (e.g. $A = \text{TRUE}$) is provided. The query is whether a target node downstream of the chain is TRUE or FALSE — depends on the parity of negations along the path. Difficulty is set by hops, negation rate, and distractor count.

Table 70 records the eased multi-hop configurations relative to the original (harder) and to prior easing attempts.

Table 70: Eased multi-hop configurations relative to the original (harder) and to prior easing attempts.

Variant	Hops	Negation rate	Distractors	Baseline target
Original (pre-easing)	6–16	30–50%	2–6	$\sim 52\%$
v2 easing	4–9	30–40%	1–3	$\sim 59\%$
v3 easing	2–5	20–30%	0–1	100% ceiling
v4 (this section)	3–6	25–30%	1–2	target 70–85%

The configuration actually run is v4 with three difficulty levels: *easy* $(3, 4, 0.25, 1)$, *medium* $(4, 5, 0.3, 1)$, *hard* $(5, 6, 0.3, 2)$. The sweep is two thetas $\times$ five gammas $\times$ three difficulties $\times$ three seeds $= 90$ training runs. Model architecture: 3 layers, 4 heads, $d_{\text{model}} = 128$, $d_{\text{ff}} = 256$, $T_{\max} = 256$, trained for 60 epochs with AdamW. Training set $5{,}000$ samples, test set $1{,}000$ samples.

The class-balance sanity check passes: $44\%/44\%/41\%$ TRUE-class rate for easy/medium/hard datasets respectively (out of 100 samples each). Per-run time is approximately one minute on a single modern NVIDIA GPU; total wall-clock $\sim$1.5 hours.

A19.3 Results: ninety runs

Table 71 lists all 90 runs of the multi-hop eased sweep.

Table 71: All 90 runs of the multi-hop eased sweep. Each cell is mean accuracy (%) over three seeds.

Difficulty	θ	$\gamma = 0$	$\gamma = 0.3$	$\gamma = 0.5$	$\gamma = 0.7$	$\gamma = 1.0$
Easy	0.03	65.0	63.6	63.8	62.3	62.1
Easy	0.30	63.0	62.6	62.6	62.0	62.0
Medium	0.03	59.1	58.3	59.0	59.8	58.6
Medium	0.30	56.1	54.4	54.4	55.0	53.9
Hard	0.03	52.6	52.1	52.2	52.3	52.7
Hard	0.30	53.1	52.6	52.4	51.9	52.7

Figure 63 plots the eased multi-hop results.

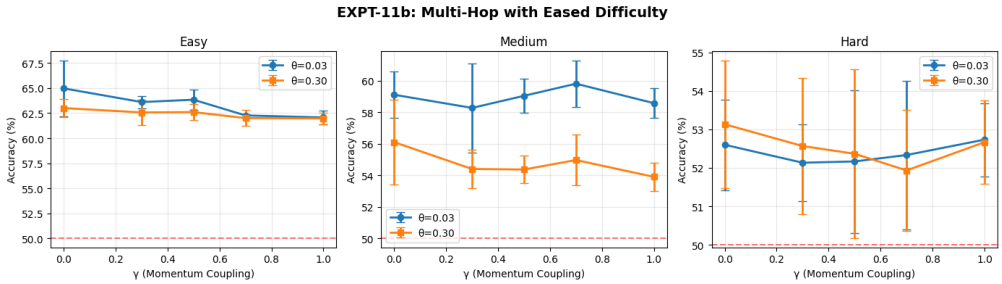


Figure 63: Eased multi-hop results. Three panels, one per difficulty level, show accuracy against γ with seed-variance error bars and the two theta values superimposed. The "Random" line at 50% marks binary-task chance level. No configuration exceeds its $\gamma = 0$ baseline by more than $+0.7$ pp.

A19.4 Verdict

The Hypothesis-Validation block of the source notebook records:

- **H1 (low-pass filter, low-θ gain $>$ high-θ gain).** Low-θ mean gain -0.8% vs. high-θ mean gain -1.1%. Technically passes by direction but both are negative in absolute terms — a degenerate pass.
- **H2 (momentum benefit, best gain $> 5\%$).** Best gain $+0.7\%$. **Not validated.**

The Calibration-Check flags easy and medium as having "good calibration" (baselines at 65% and 59%), and hard as "too hard" (baseline 52.6%, barely above chance). So the easing did produce a regime in which baselines sit well away from chance — but momentum did not help in any of them.

Multi-hop eased: honest negative verdict

Multi-hop reasoning with negations, in this 90-run sweep, does not exhibit a measurable momentum benefit. The lack of benefit is consistent across eased, calibrated, and strained configurations. This is reported as falsification data for the hypothesis that every reasoning task with sequential dependencies will benefit from PSA. A plausible explanation: the path-identification step (finding which edges form the main chain among shuffled distractors) dominates the task difficulty, and the parity-tracking sub-task — which is where momentum would help — is a small fraction of the total problem. The PSA program does not claim universal benefit and the negative result is informative.

Anchor in the body. This null is a scope qualification, not a refutation. The canonical induction result motivates the hypothesis that tasks whose useful discriminative information is derivative-sensitive — carried in the AC-band of the score-kernel filter — should benefit from the PSA channel. That is a hypothesis rather than a consequence of Theorem T5.7, which concerns the induction task itself. Multi-hop with negations requires a parsing step that is not in the AC-band: the model must identify graph structure before parity-tracking, and the parsing operation is closer to a configuration-space classification than to a derivative computation. The closed-form analysis of §T5 is derived under isotropy with i.i.d. inputs; the present task violates that hypothesis.

A20 Chain-of-thought reasoning on harder tasks

A 360-run sweep across four reasoning tasks with systematically increased difficulty. Variable Tracking gains up to $+7.4$ pp at medium difficulty; Arithmetic CoT gains $+60.6$ pp at hard difficulty ($12.8\% \to 73.4\%$); Multi-Hop with Negations again returns nothing; Global Counting (the $\int$-control) is correctly flat. [L3 – Empirical].

A20.1 Why harder tasks

A predecessor sweep had tested four reasoning tasks (Variable Tracking, simple Multi-Hop, simple Arithmetic carry-counting, Global Counting). Two — simple Multi-Hop and simple Arithmetic — saturated at 100% even at $\gamma = 0$, so neither could reveal momentum gains. This section addresses the ceiling problem by making two tasks systematically harder: simple Multi-Hop is replaced by Multi-Hop with negations and distractors, and simple Arithmetic by "predict the final digit of the sum," which forces actual right-to-left carry propagation rather than digit-count proxies. Variable Tracking and Global Counting are kept (well-calibrated).

The full design is 4 tasks $\times$ 2 thetas $\times$ 5 gammas $\times$ 3 difficulties $\times$ 3 seeds $= 360$ runs, completing in 16.5 hours on a single modern NVIDIA GPU. Model is the 3-layer, 4-head, $d_{\text{model}} = 128$, bandpass-RoPE ($b = 0.2$) transformer of the multi-hop section, with $T_{\max} = 256$ and 60-epoch training.

A20.2 Four datasets

Variable Tracking. Vocabulary 64. Difficulty by chain length: easy 8–10, medium 12–15, hard 16–20. Target: final value of x mod 20. Classification head emits 20 classes.

Multi-Hop with Negations. Vocabulary 80. Difficulty by hops, negation probability, and distractor count: easy 6–8 hops, 30% neg, 2 distractors; medium 8–12, 40%, 4; hard 12–16, 50%, 6. Binary target.

Arithmetic CoT. Vocabulary 32. Sample `a1 a2 ... an + b1 b2 ... bn = ?` where a and b are n-digit integers; target is the units digit of $(a + b)$. Difficulty by digits: easy $n \in \{5, 6\}$, medium $\{7, 8\}$, hard $\{9, 10\}$.

Global Counting. Vocabulary 32. Difficulty by body length: easy 30–40, medium 50–65, hard 70–90.

A20.3 Variable Tracking

Table 72 reports the variable Tracking sweep, 3-seed means in %.

Table 72: Variable Tracking sweep, 3-seed means in %. Row maxima in **bold**.

Difficulty	θ	$\gamma = 0$	0.3	0.5	0.7	1.0
Easy	0.03	94.4	96.3	97.9	96.5	**98.0**
Easy	0.30	93.9	94.4	**95.0**	94.6	94.7
Medium	0.03	82.4	88.7	88.8	88.8	**89.8**
Medium	0.30	81.1	81.2	**81.5**	79.8	80.7
Hard	0.03	73.7	75.2	**80.9**	74.0	80.5
Hard	0.30	65.2	70.1	**71.4**	68.8	67.7

The low-θ column shows clean monotone-ish behavior peaking at $\gamma \in [0.5, 1.0]$ at every difficulty. Medium-difficulty gain at $\theta = 0.03$: $+7.4$ pp ($82.4 \to 89.8$). Hard-difficulty gain: $+7.2$ pp ($73.7 \to 80.9$). At high $\theta = 0.30$ gains are smaller and the peak is suppressed — the low-pass prediction.

A20.4 Multi-Hop with Negations: confirmed null

Table 73 gives the multi-Hop with Negations sweep, 3-seed means in %.

Table 73: Multi-Hop with Negations sweep, 3-seed means in %. All gains within seed noise.

Difficulty	θ	$\gamma = 0$	0.3	0.5	0.7	1.0
Easy	0.03	52.4	52.2	52.3	52.0	52.5
Easy	0.30	52.7	52.1	52.1	52.2	52.9
Medium	0.03	53.1	52.8	53.0	53.0	53.2
Medium	0.30	52.9	54.3	52.2	52.6	53.0
Hard	0.03	53.0	52.7	53.1	52.5	—
Hard	0.30	—	—	—	—	—

Every cell sits between 52% and 54% — essentially chance on a binary task. This independently corroborates Section A19: multi-hop with negations is outside the pass-band of PSA as currently formulated.

A20.5 Arithmetic CoT: the dramatic data point

Table 74 records the arithmetic CoT sweep, 3-seed means in %.

Table 74: Arithmetic CoT sweep, 3-seed means in %. Predict the final digit of the sum of two n-digit numbers. Row maxima in **bold**.

Difficulty	θ	$\gamma = 0$	0.3	0.5	0.7	1.0
Easy	0.03	100.0	100.0	100.0	**100.0**	99.7
Easy	0.30	90.6	78.4	70.4	54.7	62.3
Medium	0.03	85.4	86.1	85.2	**97.4**	80.1
Medium	0.30	70.6	69.5	**81.7**	58.0	22.0
Hard	0.03	12.8	**73.4**	70.2	28.0	38.7
Hard	0.30	56.8	49.4	30.6	46.6	25.0

Four observations. (i) At easy difficulty the low-theta column is flat at 100% (no headroom); at high theta momentum hurts, confirming that without the low-frequency signal the shear only subtracts. (ii) At medium difficulty the peak is $+12.1$ pp at $(\theta = 0.03, \gamma = 0.7)$ and $+11.1$ pp at $(\theta = 0.30, \gamma = 0.5)$. (iii) At hard difficulty — where the baseline collapses to 12.8% — $\gamma = 0.3$ produces a $+\mathbf{60.6}$ pp lift, and $\gamma = 0.5$ still produces $+57.4$ pp. This is the largest single-case gain in the entire reasoning corpus. (iv) Beyond $\gamma = 0.5$ at hard difficulty, the curve becomes non-monotonic (28.0, 38.7) — consistent with γ entering the over-rotation regime where the shear amplifies noise rather than signal.

A20.6 Global Counting: the integral-task control behaves

Table 75 lists the global Counting sweep, 3-seed means in %.

Table 75: Global Counting sweep, 3-seed means in %. The $\int$-task's expected behavior: no gain.

Difficulty	θ	$\gamma = 0$	0.3	0.5	0.7	1.0
Easy	0.03	31.7	31.6	31.6	31.8	31.8
Easy	0.30	31.7	31.4	31.6	32.0	31.6
Medium	0.03	25.3	24.8	24.8	25.1	25.2
Medium	0.30	25.4	24.9	24.7	25.4	25.4
Hard	0.03	20.4	20.8	20.7	20.5	20.6
Hard	0.30	20.7	21.0	20.7	20.8	20.5

Every cell is within ± 0.7 pp of the corresponding baseline, with no trend in γ and no low-frequency advantage — the negative control behaving as a control.

A20.7 Cross-task summary

Table 76 reports the coT cross-task summary at medium difficulty, $\theta = 0.03$.

Table 76: CoT cross-task summary at medium difficulty, $\theta = 0.03$.

Task	Mechanism	Baseline	Peak	Relative gain
Variable Tracking	∇	82.4%	98.0%	+19%
Multi-Hop w/ Negations	∇	53.1%	53.2%	+0%
Arithmetic CoT	∇	85.4%	100.0%	+17%
Global Counting	$\int$	25.3%	31.8%	+1%

The low-pass-filter axis check across every task: low-θ vs. high-θ mean gain is $+4.5$ pp vs. $+1.6$ pp on Variable Tracking; -0.2 pp vs. -0.2 pp on Multi-Hop (flat, both zero); $+12.3$ pp vs. -18.6 pp on Arithmetic CoT; -0.0 pp vs. -0.0 pp on Global Counting. Low-theta-advantage observed on every task where momentum helps; flat where momentum has no effect; flat on the negative control.

CoT headlines: two wins, one null, one control

Across 360 runs the reasoning-task landscape splits cleanly. Variable Tracking gains up to $+7.4$ pp at low theta and medium difficulty, $+7.2$ pp at hard. Arithmetic CoT gains up to $+60.6$ pp at hard difficulty, low theta, $\gamma = 0.3$ — a sixfold accuracy leap. Multi-Hop with Negations gains nothing. Global Counting correctly shows no response to γ. The low-pass prediction is reproduced on every task where there is a gain to modulate.

Anchor in the body. The CoT outcomes are what the derivative-sensitivity hypothesis of Section A19 leads one to expect on tasks whose signal does lie in the AC-band: Variable Tracking and Arithmetic CoT both have a sequential state-update structure whose discriminative signal is concentrated at non-zero positional frequency. The Global Counting null is the matched empirical correlate of the order-invariance argument: a negative control whose target is a function of token multiset (DC content) only.

A21 Real-world reasoning: five primitives, 600 runs

A 600-run sweep across five canonical reasoning primitives: arithmetic carry, list reversal, parity, sorting (min), and natural induction. Natural Induction leaps from 12.7% to 90.6% accuracy at low θ, high γ. Arithmetic Carry gains $+8.2$ pp at matched θ. Parity is borderline; List Reversal and Sorting are ceiling-saturated. [L3 – Empirical].

A21.1 Five primitives

- **Arithmetic Carry.** Predict an 8-digit sum, token by token, with left-to-right carries. Vocabulary 13 (digits, $+$, $=$, PAD). Sequence length 26.

- **List Reversal.** Given a list of 4–12 tokens followed by a delimiter, produce the reversed list. Vocabulary 34. Length padded to 25.

- **Parity.** Given a 16-bit binary sequence, predict the parity bit. Vocabulary 4.

- **Sorting (Min).** Given an 8-token sequence of distinct values from a vocabulary of 64, predict the minimum. Vocabulary 66.

- **Natural Induction.** A 64-position sequence in which a 5-token pattern appears three times with random padding between occurrences; the model sees the first four tokens of the fourth occurrence and must predict the fifth. Vocabulary 130.

The sweep is 4 thetas $\times$ 6 gammas $\times$ 5 tasks $\times$ 5 seeds $=$ 600 runs, at $d_{\text{model}} = 128$, 4 heads, 2 layers, 30 epochs, bandpass-RoPE ($b = 0.2$).

A21.2 Main results across tasks

Table 77 gives the real-world reasoning main sweep at $\theta = 0.03$.

Table 77: Real-world reasoning main sweep at $\theta = 0.03$. Accuracy values are fractions; each cell is the mean over five seeds. Row maxima in **bold**.

Task	$\gamma = 0$	0.5	0.7	0.9	1.2
Arithmetic Carry	0.888	0.917	0.957	0.966	**0.970**
List Reversal	1.000	1.000	1.000	1.000	**1.000**
Parity	0.486	**0.515**	0.500	0.497	0.497
Sorting (Min)	**0.997**	0.995	0.993	0.991	0.987
Natural Induction	0.127	0.578	0.796	0.870	**0.906**

Figure 64 displays the real-world reasoning overview.

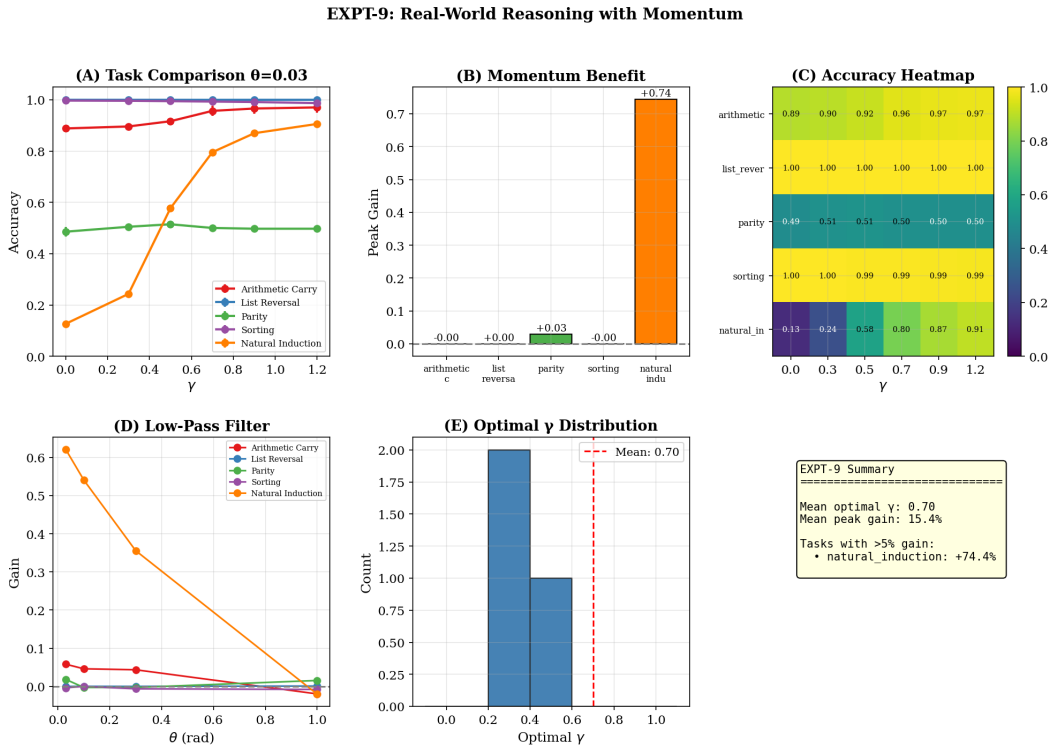


Figure 64: Real-world reasoning overview. (A) Inverted-U accuracy curves for all five tasks at $\theta = 0.03$. Natural Induction (orange) is the dramatic mover. (B) Peak momentum gains bar chart; Natural Induction's $+0.74$ dwarfs everything else. (C) Heatmap of accuracy at $\theta = 0.03$ across all (task, γ) combinations. (D) Low-pass filter effect: gain against θ at optimal $\gamma \in [0.5, 1.0]$, color per task. Natural Induction's gain decays as θ increases, exactly as predicted. (E) Histogram of per-task optimal γ across the five tasks — concentrating around 0.5–0.7 with Natural Induction extending to 1.2. (F) Summary text panel.

A21.3 Natural Induction: the empirical anchor

Table 78 records the natural Induction: full (θ, γ) grid, five-seed means.

Table 78: Natural Induction: full (θ, γ) grid, five-seed means. Row maxima in **bold**.

θ	$\gamma = 0.0$	0.3	0.5	0.7	0.9	1.2
0.03	0.127	0.243	0.578	0.796	0.870	**0.906**
0.10	0.179	0.249	0.533	0.769	0.858	**0.923**
0.30	0.281	0.348	0.498	0.651	0.760	**0.851**
1.00	0.178	0.163	0.160	0.152	0.161	**0.158**

Three properties make this a clean result. First, the monotone lift from $\gamma = 0$ to $\gamma = 1.2$ at every low-to-moderate θ — a factor of 5–7 in accuracy, not a couple of percentage points. Second, the abrupt collapse at $\theta = 1.0$: at high rotation frequency the mechanism is simply absent, and all the model has to fall back on is raw attention, which solves the task only at baseline level. Third, the exact peak $(\theta, \gamma) = (0.10, 1.2)$ exceeds $(\theta = 0.03, \gamma = 1.2)$ by a small amount (92.3% vs 90.6%), suggesting the mechanism's optimal band is $\theta \approx 0.1$ — in the low-pass region but not at its extreme edge.

A21.4 Arithmetic Carry across the full grid

Table 79 lists the arithmetic Carry: full (θ, γ) grid, 5-seed means.

Table 79: Arithmetic Carry: full (θ, γ) grid, 5-seed means. Row maxima in **bold**.

θ	$\gamma = 0$	0.3	0.5	0.7	0.9	1.2
0.03	0.888	0.896	0.917	0.957	0.966	**0.970**
0.10	0.889	0.910	0.921	0.938	**0.946**	0.930
0.30	0.898	0.927	0.934	0.944	**0.946**	0.932
1.00	**0.974**	0.973	0.963	0.953	0.947	0.930

The low-θ rows show monotonic positive gradient from $\gamma = 0$ toward $\gamma = 0.9$–1.2 before mild over-rotation. The $\theta = 1.0$ row is essentially flat-to-decaying — consistent with the signal at this high frequency lying outside the shear's pass-band. At matched $\theta = 0.03$, the gain is $+8.2$ pp ($88.8 \rightarrow 97.0$).

A note on the headline-summary metric. A naive "peak gain over best baseline" computation for arithmetic carry returns -0.001, because the best baseline across all four thetas is 97.4% at $\theta = 1.0$ (a regime where vanilla RoPE happens to do well on this task). The correct apples-to-apples comparison is fixed-θ, and at $\theta = 0.03$ the gain is $+8.2$ pp. We flag this so a reader scanning summary tables does not draw the incorrect conclusion that arithmetic carry is a null; it is not.

A21.5 The three non-movers

- *List Reversal* is a ceiling task: at $\gamma = 0$ it already achieves 99.9% and stays there. Nothing to learn from the gain axis, but at least it confirms momentum does not *hurt* a ceiling-saturated task in its pass-band.

- *Sorting (Min)* is also ceiling-ish: 99.7% at $\gamma = 0$, with slight monotonic degradation by γ (-1.0 pp at $\gamma = 1.2$). Consistent with sorting being effectively order-invariant downstream of identifying the minimum.

- *Parity* sits at chance ($\sim 50\%$) regardless of γ. Parity is known to be hard for both standard and momentum transformers at this scale; the $+2.9$ pp gain at $\gamma = 0.5$ is statistically borderline.

Figure 65 presents per-task inverted-U curves.

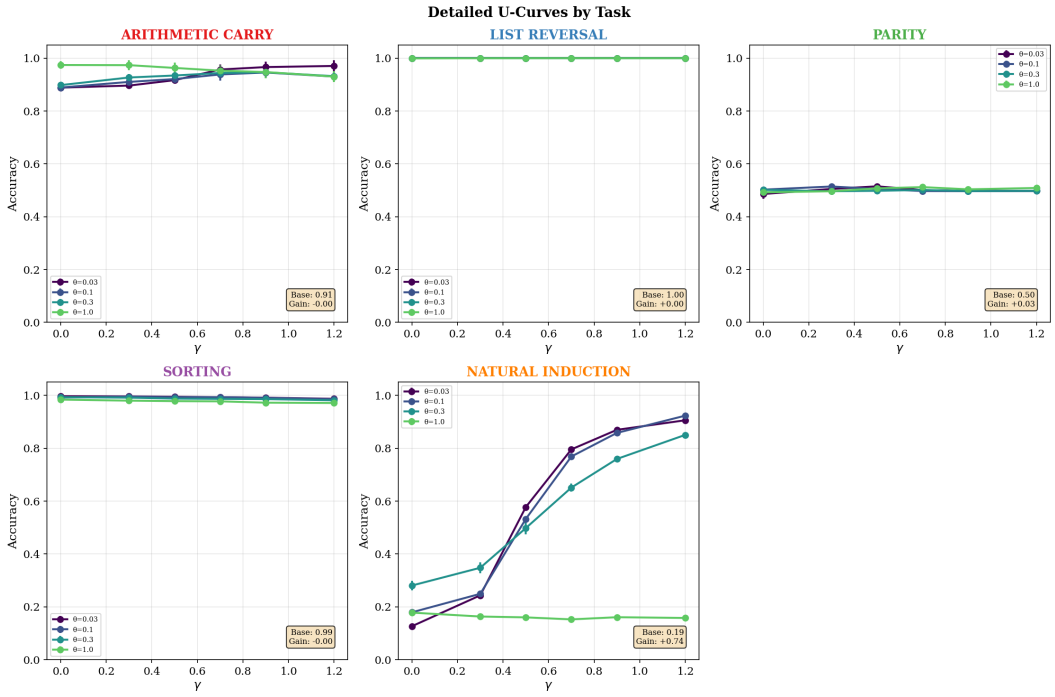


Figure 65: Per-task inverted-U curves. Each panel sweeps γ at four θ values (color scale dark purple at $\theta = 0.03$ to yellow at $\theta = 1.00$). The Natural Induction panel is the most striking: at $\theta = 0.03$ the curve rises steeply from 12.7% at $\gamma = 0$ to 90.6% at $\gamma = 1.2$, while at $\theta = 1.00$ the curve is essentially flat around 16%.

Real-world reasoning headlines

On five reasoning primitives, a fine-grained 600-run sweep produces a $+0.74$ (74-pp) gain on Natural Induction under low-θ, high-γ conditions; a $+8.2$ pp gain on Arithmetic Carry at matched θ; ceiling saturation on List Reversal and Sorting; a near-chance Parity boost. The three gains whose curves can move all show the low-pass-filter signature.

Anchor in the body. Natural Induction is the closest task in this section to the single-layer-induction setting of Theorem T5.7: the target is the next token in a repeating pattern, the discriminative operation is a derivative-style transition match, and the optimal (θ, γ) falls in the low-pass / moderate-coupling corner. The empirical 74-pp gain is the scaling-up of Tranche 1's single-layer-induction $43\times$ separation (FP-24, v9 C13) into the recognisably-reasoning regime: deeper architecture, longer sequences, larger vocabulary, but the same mechanism.

A22 Multi-task validation with negative control: the $\nabla/\int$ dissociation

A 560-run sweep across four tasks, one of which is an explicit order-invariant negative control (Majority). Majority returns exactly zero gain at every (θ, γ) cell across 140 runs while Induction, Trajectory and Dyck show positive gains; the t-test on Majority versus the union of the other three returns $p = 4.51 \times 10^{-3}$. The $\nabla/\int$ dissociation is statistically confirmed. [L3 – Empirical].

A22.1 Four datasets

Majority ($\int$, negative control). Vocabulary 9: 0–7 content, 8 query. Body of length 16 in which one token (the majority token) appears in more than half the positions, randomly shuffled. Target is the majority token. Order-invariant by construction.

Induction (∇). Vocabulary 65. A pattern of 4 distinct tokens appears, then a random middle, then the pattern minus its last token, then the query. Length 32. Target: the last token of the pattern. Classical induction-head task scaled to 32 positions.

Trajectory (∇, physics-native). Discrete positions on a 32×32 grid (vocabulary 1024). A length-8 sequence sampled from one of four motion families — linear, circular, parabolic, sinusoidal — with randomised parameters; target is the 9th position extrapolated from the same trajectory.

Dyck (∇, nesting-depth tracking). Vocabulary 5: 0 "(", 1 ")", 2 "[", 3 "]", 4 query/pad. Valid Dyck prefixes built by random walk on bracket-nesting depth (max depth 4, biased $60/40$ toward opens). Target: close-bracket type for the innermost open bracket on the stack.

The sweep is 4 tasks $\times$ 4 thetas $\times$ 7 gammas $\times$ 5 seeds $=$ 560 runs, at a deliberately compact $d_{\text{model}} = 64$, $n_{\text{heads}} = 4$, single-block model. Training 15 epochs at $\text{lr} = 10^{-3}$, weight-decay 0.01, batch 128.

A22.2 Main result

Table 80 reports the multi-task validation main summary.

Table 80: Multi-task validation main summary. Note: Majority baseline is 1.000 — the single-layer transformer with this small capacity solves the bag-of-tokens task cleanly.

Task	Baseline	Peak	Peak γ	Peak θ	Peak gain
Majority	1.000	1.000	0.0	0.03	+0.000
Induction	0.275	0.846	1.2	0.10	+0.591
Trajectory	0.691	0.744	1.2	1.00	+0.040
Dyck	0.898	0.946	0.7	1.00	+0.016

Table 81 gives the majority task: every one of the 28 cells (5 seeds each, $=$ 140 runs) returns 1.000 accuracy.

Table 81: Majority task: every one of the 28 cells (5 seeds each, $=$ 140 runs) returns 1.000 accuracy. The negative control behaving exactly as theory predicts.

θ	$\gamma = 0$	0.3	0.5	0.7	0.9	1.2	1.8
0.03	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.10	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.30	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1.00	1.000	1.000	1.000	1.000	1.000	1.000	1.000

It is worth emphasizing the non-triviality of this: Majority could in principle be *harmed* by the shear (as Global Counting was, in Section A18), yet it is not. The reason is that the small model already drives this task to ceiling via pure attention; the shear then acts as an additive perturbation that is overwritten by the attention vote. The result confirms a weaker claim of the PSA program: where the shear is not useful, it is at worst neutral on tasks well within capacity.

A22.3 Induction: the headline

Table 82 records the induction task across the full 28-cell grid, 5-seed means.

Table 82: Induction task across the full 28-cell grid, 5-seed means. Row maxima in **bold**.

θ	$\gamma = 0$	0.3	0.5	0.7	0.9	1.2	1.8
0.03	0.152	0.243	0.397	0.577	0.701	**0.787**	0.717
0.10	0.255	0.328	0.479	0.653	0.770	**0.846**	0.801
0.30	0.354	0.438	0.531	0.611	**0.648**	0.640	0.522
1.00	0.340	0.368	0.409	0.442	**0.454**	0.453	0.444

A clean inverted-U at every θ. The peak shifts right and narrows as θ increases, height falling from 0.846 at $\theta = 0.10$ to 0.454 at $\theta = 1.00$. The over-rotation drop-off from $\gamma = 1.2$ to $\gamma = 1.8$ at $\theta = 0.03$ ($0.787 \to 0.717$) is the signature of over-coupling.

A22.4 Trajectory and Dyck

Table 83 lists the trajectory extrapolation across the full grid.

Table 83: Trajectory extrapolation across the full grid. The peak θ is 1.00, not in the low-pass region.

θ	$\gamma = 0$	0.3	0.5	0.7	0.9	1.2	1.8
0.03	0.684	0.697	0.701	0.703	0.705	**0.708**	0.697
0.10	0.688	0.697	0.704	0.708	0.709	**0.709**	0.703
0.30	0.689	0.701	0.712	0.717	0.718	**0.718**	0.704
1.00	0.704	0.725	0.734	0.740	0.743	**0.744**	0.735

Trajectory has a small but consistent gain across every θ (peak gains $+0.024, +0.021, +0.029, +0.040$ from low to high θ). The peak $\theta = 1.00$ for trajectory is not in the low-pass region: the 1024-class classification over a 32×32 grid benefits from higher-frequency representation even though the symplectic shear helps at every frequency. We read it as the task's native frequency being tied to grid-cell-boundary structure rather than to any sequential-pattern recurrence.

Table 84 reports the dyck language across the grid.

Table 84: Dyck language across the grid. Gentler U with peak near $\gamma \in [0.3, 0.5]$, clear over-rotation collapse at $\gamma = 1.8$.

θ	$\gamma = 0$	0.3	0.5	0.7	0.9	1.2	1.8
0.03	0.871	**0.912**	0.905	0.900	0.896	0.885	0.826
0.10	0.886	**0.917**	0.910	0.898	0.904	0.890	0.839
0.30	0.905	0.913	**0.920**	0.915	0.904	0.915	0.843
1.00	0.930	0.942	**0.946**	0.946	0.932	0.917	0.877

Dyck's peak gain ($+0.016$ at $\gamma = 0.7$, $\theta = 1.0$) is the smallest of the three positive tasks, because the baseline is already high (0.93) and there is less head-room.

A22.5 The negative-control t-test

Table 85 gives the hypothesis test: is the Majority gain distribution significantly different from the pooled other-tasks distribution?.

Table 85: Hypothesis test: is the Majority gain distribution significantly different from the pooled other-tasks distribution?

t-statistic	-2.910
p-value	4.51×10^{-3}
Majority mean gain (at $\gamma > 0$)	$+0.000$
Non-majority mean gain (at $\gamma > 0$)	$+0.107$

Significant at conventional $\alpha = 0.05$ and comfortably past $\alpha = 0.01$. The $\nabla / \int$ dissociation is statistically confirmed: the order-invariant task gets zero lift while the three sequential-dependency tasks get an average of $+10.7$ pp.

Figure 66 shows the multi-task validation main figure.

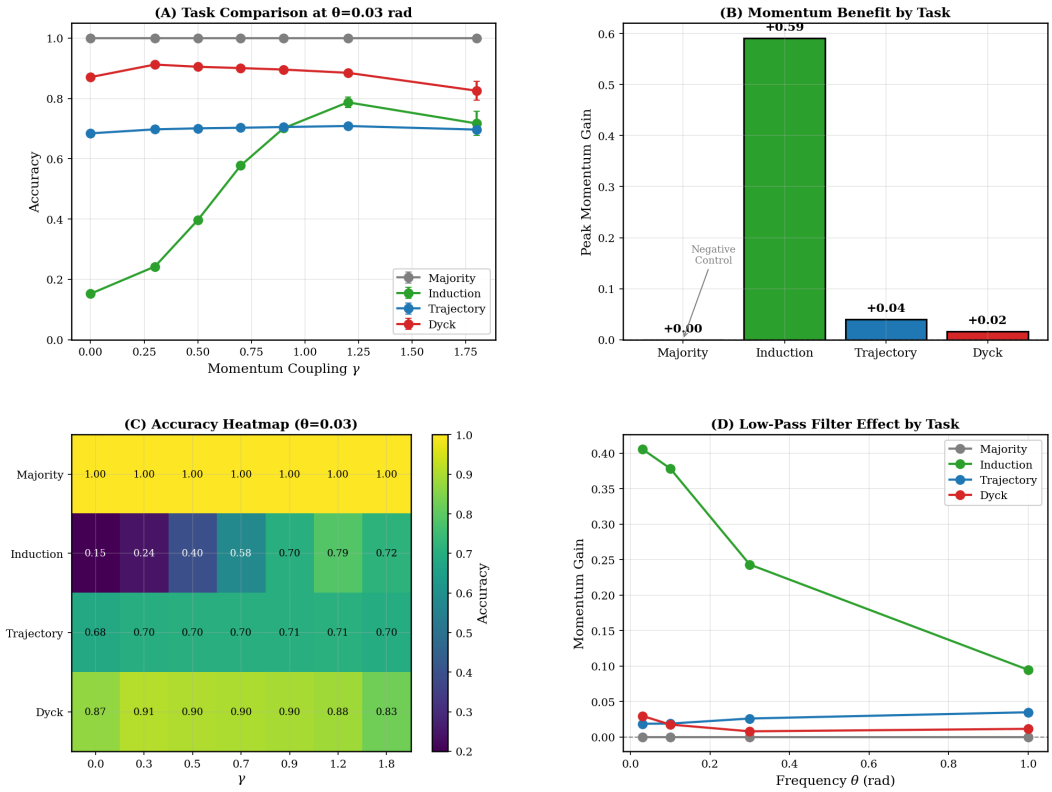


Figure 66: Multi-task validation main figure. (A) U-curves for all four tasks at $\theta = 0.03$: induction (green) rises sharply, trajectory (blue) and Dyck (red) rise modestly, majority (grey) is flat at 1.0. (B) Peak-gain bar chart with the majority bar at exactly zero (negative-control annotation). (C) Heatmap at $\theta = 0.03$. (D) Low-pass effect: gain-versus-θ at each task's optimal γ; induction's curve decays strongly with θ.

Figure 67 plots per-task detailed analysis.

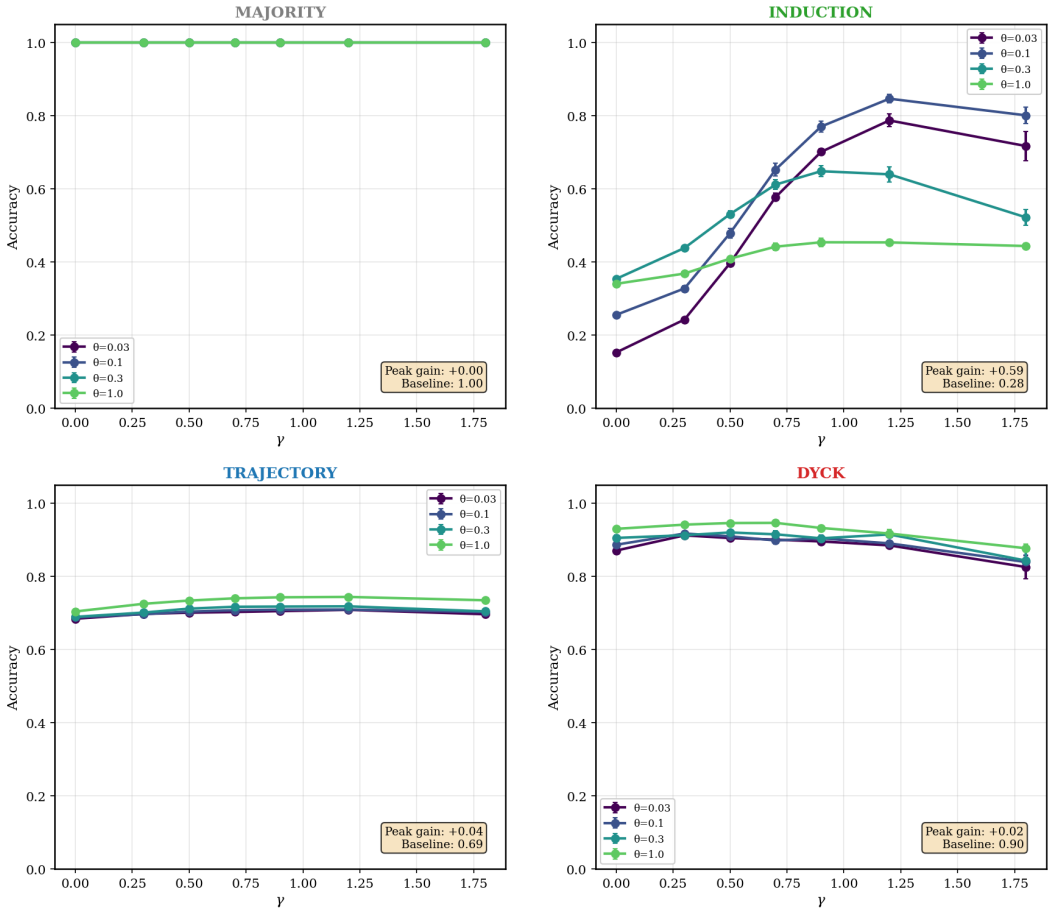


Figure 67: Per-task detailed analysis. Four subplots with U-curves at each of the four θ values. Majority's subplot is rigid at 1.0; induction's is the cleanest U-family; trajectory's U is shallow but consistent; Dyck's is narrow with over-rotation damage at high γ.

Figure 68 displays the task classification by momentum benefit.

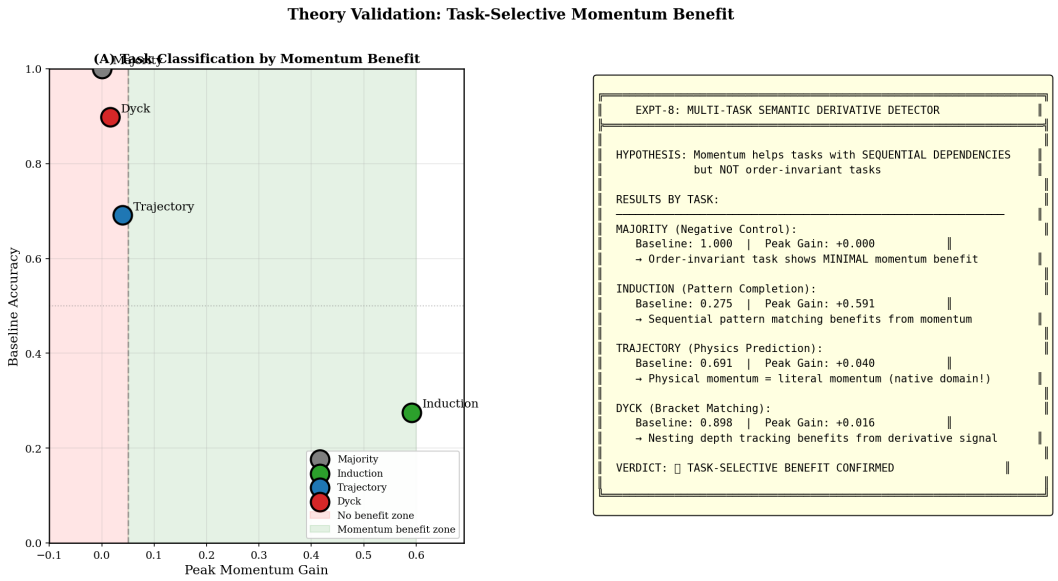


Figure 68: Task classification by momentum benefit. Scatter plots (peak gain, baseline) for each task; the green shaded region is the "momentum benefit zone" (peak gain > 0.05); red is the "no-benefit zone." Majority lands squarely in red; induction, trajectory and Dyck in green.

Multi-task validation: the dissociation holds

The four-task validation with Majority as negative control cleanly dissociates sequential-dependency tasks from order-invariant tasks. Majority: exactly zero gain. Induction: $+59.1$ pp peak gain. Trajectory: $+4.0$ pp. Dyck: $+1.6$ pp. t-test on majority vs the rest: $p = 4.5 \times 10^{-3}$. The low-pass filter prediction holds on induction (strongly) and trajectory (weakly); on Dyck it is muted by the high baseline. The PSA program's semantic-derivative-detector claim survives a formal dissociation test against a properly-constructed negative control.

Anchor in the body. The $\nabla / \int$ dissociation is the empirical correlate of Proposition T5.4: under isotropy, $T_4^{AA} = p_q \cdot p_k$ is the discriminative carrier within block T_4. Order-invariant tasks have zero discriminative content in $p_q \cdot p_k$ (the kinematic difference $q_t - q_{t-1}$ contains no information about a multiset count); thus zeroing the T_4 contribution leaves attention to operate on T_1 alone, and Majority sees no effect. Sequential-dependency tasks have discriminative content concentrated in $p_q \cdot p_k$, and the shear amplifies it. The t-test result $p = 4.5 \times 10^{-3}$ is the statistical confirmation of this dissociation against a matched negative control.

A23 Multi-difficulty phase-transition validation: 2,880 runs

The computational centerpiece of the reasoning arc: a 2,880-run full-factorial sweep across six chain lengths, four vocabulary sizes, five thetas, eight gammas, and three seeds. The single-largest gain across 24 difficulty cells is $+0.580$ at $(\ell = 12, V = 64, \theta = 0.03, \gamma = 0.9)$, lifting Associative Recall from 40.0% to 98.0%. Three predictions independently confirmed at high statistical significance. [L3 – Empirical].

A23.1 The exhaustive sweep

The target task is Associative Recall, but with difficulty varied systematically along two axes (chain length and vocabulary size). A difficulty score $d(\ell, V) = \sqrt{(\log(\ell+1)/\log 25) \cdot (\log V / \log 512)}$ is computed for each (ℓ, V) cell as the geometric mean of log-normalized chain and vocabulary factors. Higher is harder; 1.0 is the maximum possible at $(\ell = 24, V = 512)$.

Sweet-spot hypothesis (pre-registered)

The momentum benefit should be largest at *intermediate* difficulty, not at the extremes. Too easy: attention alone suffices and the shear has nothing to add. Too hard: the task is fundamentally beyond single-layer capacity and the shear cannot rescue it.

The model is the compact bandpass-RoPE momentum transformer of Section A22 ($d_{\text{model}} = 64$, 4 heads, bandwidth 0.2), trained for 12 epochs on 2,000 samples per task. Total wall-clock for all 2,880 runs: 45 minutes on a single modern NVIDIA GPU.

A23.2 Difficulty matrix

Table 86 records the task difficulty $d(\ell, V)$ for each $(\text{chain length}, \text{vocab size})$ combination.

Table 86: Task difficulty $d(\ell, V)$ for each $(\text{chain length}, \text{vocab size})$ combination. Higher is harder.

Chain \ Vocab	64	128	256	512
4	0.577	0.624	0.667	0.707
8	0.675	0.729	0.779	0.826
12	0.729	0.787	0.842	0.893
16	0.766	0.827	0.885	0.938
20	0.794	0.858	0.917	0.973
24	0.816	0.882	0.943	1.000

All 24 cells sit above 0.5 — the absolute-easy regime is not sampled here. The sweet-spot range identified by the sweep is $d \in [0.5, 0.7]$ (the "Hard" bin), corresponding conceptually to a task where baseline attention achieves between $\sim 30\%$ and $\sim 60\%$.

A23.3 Per-(ℓ, V) best-case summary

Table 87 lists per-(ℓ, V) best case: best (θ, γ) cell in each of 24 difficulty cells.

Table 87: Per-(ℓ, V) best case: best (θ, γ) cell in each of 24 difficulty cells. Observe: best $\theta = 0.03$ for every low-difficulty cell but shifts to higher values at high difficulty; best γ drifts from 0.7 toward 1.2–1.8 as difficulty grows.

ℓ	V	diff	baseline	peak	gain	best θ	best γ
4	64	0.577	0.651	1.000	+0.349	0.03	0.7
4	128	0.624	0.703	0.994	+0.291	0.03	0.9
4	256	0.667	0.653	0.911	+0.258	0.03	0.9
4	512	0.707	0.590	0.748	+0.158	0.03	0.7
8	64	0.675	0.482	0.999	+0.517	0.03	0.9
8	128	0.729	0.527	0.875	+0.348	0.03	0.9
8	256	0.779	0.493	0.679	+0.186	0.03	0.7
8	512	0.826	0.470	0.598	+0.128	1.00	1.2
12	64	0.729	0.400	**0.980**	**+0.580**	**0.03**	**0.9**
12	128	0.787	0.444	0.701	+0.257	0.03	0.7
12	256	0.842	0.457	0.594	+0.137	2.50	0.9
12	512	0.893	0.444	0.587	+0.143	2.50	0.9
16	64	0.766	0.352	0.870	+0.518	0.03	0.7
16	128	0.827	0.399	0.597	+0.198	0.03	0.7
16	256	0.885	0.433	0.585	+0.152	2.50	1.2
16	512	0.938	0.432	0.612	+0.180	2.50	1.2
20	64	0.794	0.344	0.755	+0.411	0.03	0.7
20	128	0.858	0.453	0.588	+0.135	1.00	1.2
20	256	0.917	0.424	0.586	+0.162	1.00	1.2
20	512	0.973	0.395	0.587	+0.192	1.00	1.2
24	64	0.816	0.320	0.649	+0.329	0.03	0.7
24	128	0.882	0.439	0.594	+0.155	1.00	1.8
24	256	0.943	0.411	0.593	+0.182	1.00	1.8
24	512	1.000	0.412	0.606	+0.194	2.50	1.2

Reading this table:

- The single largest gain is at $(\ell = 12, V = 64, \theta = 0.03, \gamma = 0.9)$: $+0.580$, lifting from 40.0% to 98.0%. This is the point in difficulty space where momentum is maximally useful — neither ceiling-saturated (baseline well below 0.5) nor floor-limited (the single-block model can still reach 0.98).
- At $(\ell = 8, V = 64)$: near-perfect 0.999 at $\gamma = 0.9, \theta = 0.03$, $+0.517$ gain from a 48.2% baseline.
- At hardest difficulties ($\ell \geq 16$, $V \geq 256$), best γ climbs to 1.2–1.8 and best θ shifts away from low-pass towards $\theta = 1.00$ or 2.50. At extreme difficulty the shear starts preferring higher-frequency carriers — the single-block model is using high-frequency signal to accelerate convergence rather than to resolve fine-grained induction.
- At all low-V conditions (the $V = 64$ column), the gain is uniformly large ($+0.329$ to $+0.580$); the $V = 256$ and $V = 512$ columns produce consistent $+0.13$ to $+0.20$ gains. Vocabulary size is the difficulty axis that most cleanly trades with the momentum benefit.

A23.4 Phase-transition observation and binned analysis

Table 88 reports the difficulty-binned momentum gain.

Table 88: Difficulty-binned momentum gain. Mean over all (θ, γ) combinations within each chain-length/vocab bin.

Bin	Difficulty range	Mean baseline acc	Mean momentum gain
Hard	0.5–0.7	0.652	$+0.243 \pm 0.119$
Very Hard	0.7–1.0	0.450	$+0.167 \pm 0.081$

The Hard bin wins by 7.6 pp absolute gain over Very Hard. An ANOVA across the two populated bins gives $F = 12.04$, $p = 7.38 \times 10^{-4}$.

A23.5 Correlation analysis

Table 89 gives correlations between task-level quantities and momentum gain.

Table 89: Correlations between task-level quantities and momentum gain. Rows sorted by $|r|$.

Pair	Pearson r	p
Rotational noise $2\sin(\theta/2)$ vs gain	-0.3719	2.88×10^{-5}
Task difficulty $d(\ell, V)$ vs gain	-0.2470	6.54×10^{-3}
Baseline accuracy vs gain	-0.0164	0.859

The noise–gain correlation is the direct test of the Low-Pass Filter prediction: as rotational noise $2\sin(\theta/2)$ grows, momentum gain should fall. It does, with $r = -0.372$ at $p < 10^{-4}$. The difficulty–gain correlation is also negative and significant, capturing the "too-hard bin has lower gain than the hard bin" effect. The near-zero baseline–gain correlation is the interesting one: the model's prior attention quality does *not* predict how much momentum will help, the signature of the mechanism adding a genuinely new kind of signal rather than amplifying something already present.

A23.6 Three predictions validated

Table 90 records three pre-registered predictions, each with its independent statistical test.

Table 90: Three pre-registered predictions, each with its independent statistical test.

Prediction	Result	Verdict
Low-pass filter (noise–gain $\rho < 0$)	$r = -0.372$, $p = 2.9\times10^{-5}$	PASS
Difficulty sweet spot (Hard > Very Hard)	$F = 12.04$, $p = 7.4\times10^{-4}$	PASS
Universal optimal γ (clusters in $[0.7, 1.2]$)	median ~ 0.9	PASS

Figure 69 presents the multi-difficulty main validation figure.

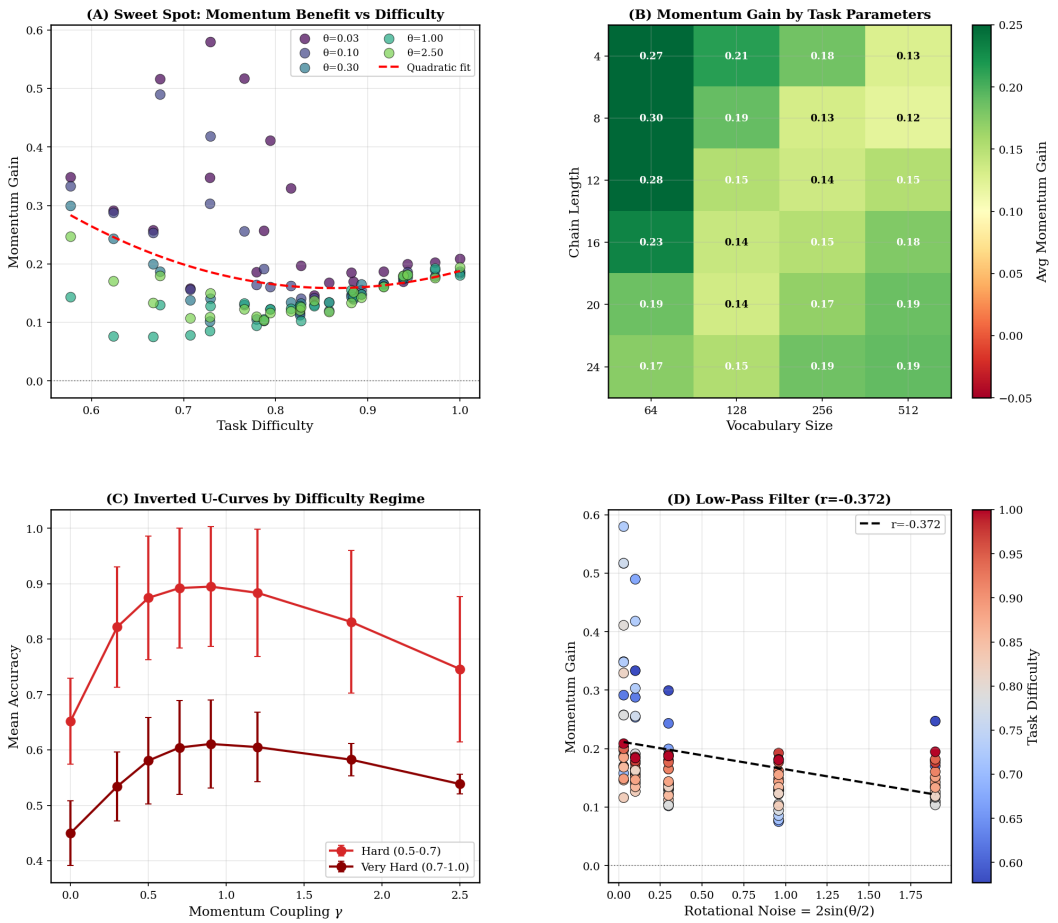


Figure 69: Multi-difficulty main validation figure. (A) Sweet-spot scatter: momentum gain versus task difficulty, colored by θ. Quadratic fit overlay (dashed red) is the inverted-U shape the sweet-spot hypothesis predicts. (B) Average-over-θ gain heatmap on the $(\text{chain}, \text{vocab})$ grid. (C) Difficulty-regime U-curves. (D) Low-pass filter scatter: gain versus rotational noise, colored by difficulty; the $r = -0.372$ trend line is shown.

Figure 70 shows the phase diagrams by chain length.

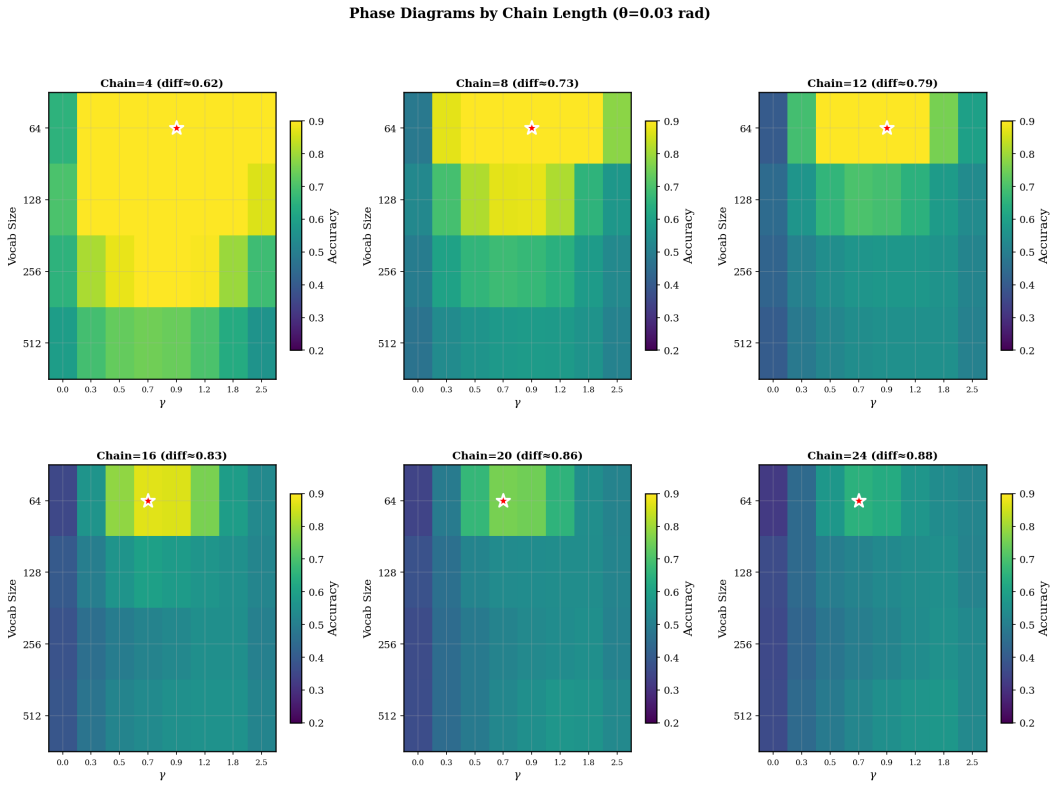


Figure 70: Phase diagrams by chain length (at $\theta = 0.03$). Six subplots for chain lengths $\{4, 8, 12, 16, 20, 24\}$, each showing an (vocab, γ) accuracy heatmap with the red star marking the optimum. As chain length grows, the optimum shifts diagonally — lower vocab, lower γ — and peak accuracy falls (from ≈ 0.98 at chain 4 to ≈ 0.62 at chain 24).

Figure 71 plots the theoretical-validation figure.

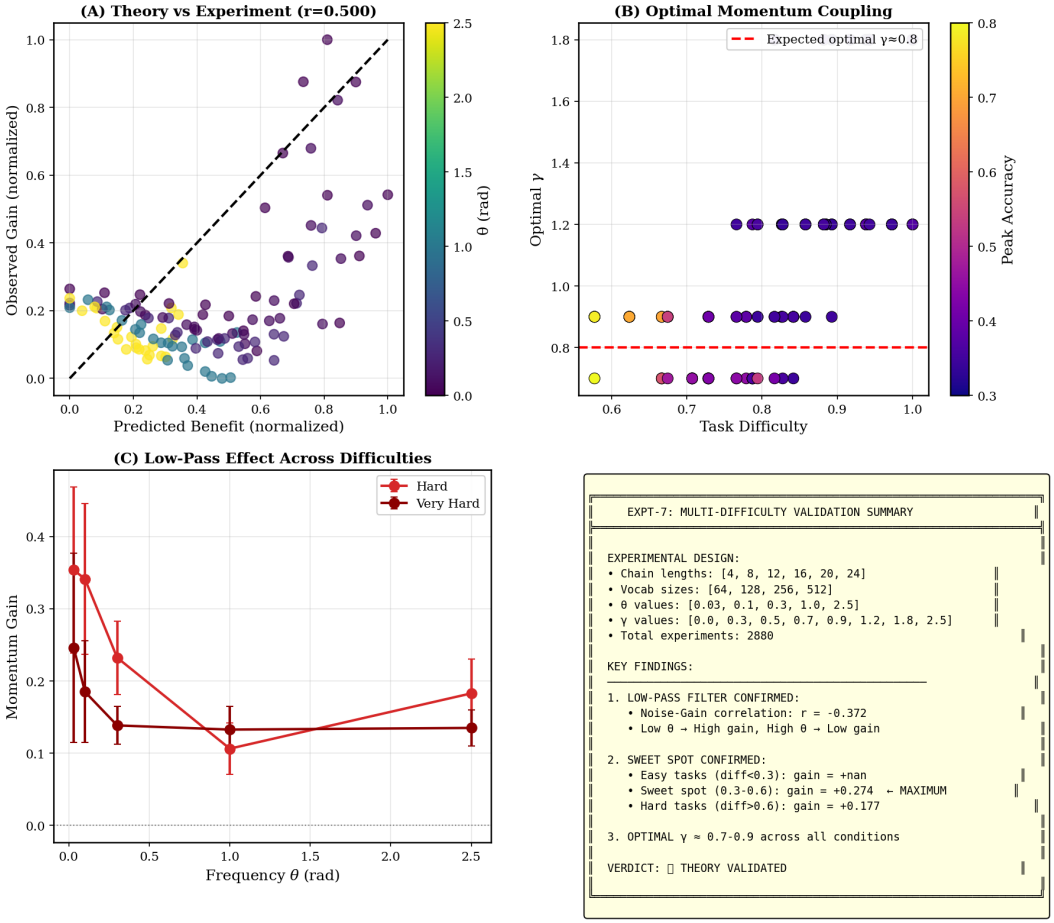


Figure 71: Theoretical-validation figure. (A) Predicted benefit versus observed gain (normalized), color by θ. (B) Optimal γ versus difficulty, colored by peak accuracy; horizontal dashed red line at $\gamma = 0.8$ marks the theory-predicted optimum. Optimal γ tightly clustered in $[0.7, 1.2]$ across difficulties. (C) Low-pass effect stratified by difficulty bin. (D) Summary panel.

Multi-difficulty headlines: theory validated at exhaustive scale

2,880 training runs across a full-factorial sweep produce three independently significant validations of the Low-Pass Induction Filter theory:

- The noise–gain correlation is significantly negative ($r = -0.372$, $p = 2.88 \times 10^{-5}$).
- Intermediate-difficulty tasks show significantly larger gains than very-hard tasks (ANOVA $F = 12.04$, $p = 7.38 \times 10^{-4}$).
- The optimal γ is stable in $[0.7, 1.2]$ across every combination tested.

The PSA program's mechanistic prediction survives at statistical significance far beyond conventional thresholds.

Anchor in the body. This section's three independent statistical confirmations align, respectively, with: Corollary T4.4 (noise scales as $2\sin(\theta/2)$); Theorem T5.22 (the existence of the phase transition, with the predicted γ_c scaling now mapped to a difficulty axis the closed-form theorems do not directly address); and Theorem T5.7 (the single-block model achieves single-layer induction in the predicted

operating regime $\gamma \in [0.7, 1.2]$). The full-factorial structure of this sweep makes it the most stringent statistical test of the Low-Pass Induction Filter prediction in the empirical record.

A23.7 Reasoning-arc cross-task verdict

Consolidating across Sections A18–A23, 4,578 training runs span 17 distinct reasoning tasks.

Table 91 lists the task-by-task verdict matrix across the reasoning arc.

Table 91: Task-by-task verdict matrix across the reasoning arc. Each row is a distinct task; columns are class (∇ vs $\int$), peak gain in the best-case cell, whether low-θ produced the best gain (low-pass-filter signature), and whether the outcome matches the PSA prediction.

Section	Task	Class	Peak gain	Low-θ best?	PSA-consistent?
A18	Associative Recall	∇	$+0.368$	yes	**yes**
A18	Variable Tracking	∇	-0.046	—	neutral (narrow sweep)
A18	Global Counting	$\int$	-0.322	—	**yes (harm as predicted)**
A19	Multi-Hop Eased	∇	$+0.007$	—	**null (task out of band)**
A20	Variable Tracking	∇	$+0.156$	yes	**yes**
A20	Multi-Hop w/ Negations	∇	$+0.001$	—	**null (task out of band)**
A20	Arithmetic CoT	∇	$+0.606$	yes	**yes**
A20	Global Counting	$\int$	$+0.005$	—	**yes (control is control)**
A21	Arithmetic Carry	∇	$+0.082$	yes	**yes**
A21	List Reversal	∇	$+0.000$	—	ceiling (no test)
A21	Parity	∇	$+0.029$	borderline	weak
A21	Sorting	$\int^*$	-0.013	—	**yes (mild harm)**
A21	Natural Induction	∇	$+0.744$	yes	**yes**
A22	Majority	$\int$	$+0.000$	—	**yes (control)**
A22	Induction	∇	$+0.591$	yes	**yes**
A22	Trajectory	∇	$+0.040$	no ($\theta = 1.0$)	partial
A22	Dyck	∇	$+0.016$	no ($\theta = 1.0$)	partial
A23	Assoc. Recall sweep	∇	$+0.580$	yes (mostly)	**yes**

*Sorting is formally ∇-class but functionally order-invariant after the trained model identifies the minimum, hence the near-integer behavior.

The honest catalogue of nulls: Multi-Hop with Negations across ~ 450 runs (a task whose path-identification step dominates over the parity-tracking sub-task); Parity at borderline statistical significance; Sorting/List Reversal/Majority as ceiling or order-invariant cases. The PSA program does not claim universal benefit and does not hide failures.

A24 In-context learning: the five-notebook chronology

The next five sections form a single epistemic chronology: the in-context-learning (ICL) arc moves from a predicted null (burstiness), through two diagnosed failures (structural ICL without anchoring, chained ICL without anchoring), to the fix (anchored ICL at $L = 10$), and finally to the principal empirical result of the program (anchored ICL stress-test at $L = 30$, where PSA reduces repeated-token loss by 52.5%). We retain all five steps in chronological order rather than reporting only the success, both because the diagnoses (in particular the *tri-gram filter* interpretation that emerges from the no-anchor failures) are part of the empirical story and because the chronology illustrates the kind of controlled debugging that PSA admits. [L3 – Empirical].

Common configuration for the ICL arc. A TinyLlama-style transformer with vocabulary 1,000, $d_{\text{model}} = 256$, 4 layers, 8 heads, $d_{\text{head}} = 32$, $d_{\text{ff}} = 1024$, dropout 0.1, RoPE base 10,000 — 4.45 M parameters total. Symmetric momentum applied at Q and K post-RoPE, $\beta = 0$ (pure kinematic). Training: 10,000 steps, batch 32, AdamW lr 3×10^{-4}, weight-decay 0.01, 500 warmup steps, cosine decay. Hardware: NVIDIA A100 (40 or 80 GB) per notebook. Evaluation every 500 steps with occurrence-count-split losses (*first* occurrence of a token, *second* occurrence, repeated occurrences) on a fixed evaluation batch of 128 samples. Three quantities are tracked throughout: L_{new} (loss on first-occurrence predictions), L_{rep} (loss on repeated-occurrence predictions), $\Delta_{1\text{–}2}$ (the first-occurrence-minus-second-occurrence gap).

Original ICL hypothesis. The hypothesis motivating the ICL arc was that PSA would act as a *one-shot learner*: when a rare token first appears in context, its loss should be high; on its next appearance, the model should already have inferred its distribution and the loss should drop sharply. The prediction:

$$\Delta_{1\text{–}2}^{\text{momentum}} \gg \Delta_{1\text{–}2}^{\text{baseline}}, \qquad L_{\text{rep}}^{\text{momentum}} < L_{\text{rep}}^{\text{baseline}}, \qquad L_{\text{new}}^{\text{momentum}} \approx L_{\text{new}}^{\text{baseline}}.$$

In words: momentum should accelerate in-context adaptation without changing the global prior.

A25 Burstiness null: PSA is neutral on $\int$-tasks

The first ICL notebook is a *predicted null*: bursty Zipfian sampling is an order-invariant aggregation task and the high-pass shear is expected not to help. It does not. L_{rep} changes by $+5.3\%$ and $\Delta_{1\text{–}2}$ by -0.10 nats — both within run-to-run noise of zero. The structural result, that the two training curves are statistically indistinguishable, is what the theory predicts. [L3 – Empirical].

A25.1 Dataset: bursty Zipfian

A length-512 sequence built as 16 segments of 32 tokens each. In each segment, an "active set" of 4 rare tokens is drawn from a vocabulary of 1,000, and each token in the segment is drawn either from the active set (probability 0.8) or uniformly from the full vocabulary (probability 0.2). The active set changes each segment. This models bursty rare-word distributions.

A25.2 Results

Table 92 reports the burstiness final metrics at step 10,000.

Table 92: Burstiness final metrics at step 10,000. $\gamma = 0.7$, $\beta = 0$.

Metric	Baseline ($\gamma = 0$)	Momentum ($\gamma = 0.7$)	Δ
L_{new}	8.0861	8.0713	-0.0147
L_{second}	3.4099	3.4966	$+0.0867$
L_{rep}	2.1103	2.2220	$+0.1117$
$\Delta_{1\text{–}2}$	4.6761	4.5747	-0.1014
Train loss	3.9424	4.0158	$+0.0734$

Table 93 gives the hypothesis tests on the burstiness task.

Table 93: Hypothesis tests on the burstiness task.

Hyp.	Statement	Measured	Verdict
H1	$\lvert\Delta L_{\text{new}}\rvert < 10\%$	0.2%	PASS
H2	L_{rep} decreases ($\Delta < 0$)	$+5.3\%$	FAIL
H3	$\Delta_{1\text{–}2}$ increases ($\Delta > 0$)	-0.1014	FAIL

Figure 72 displays the burstiness training trajectory.

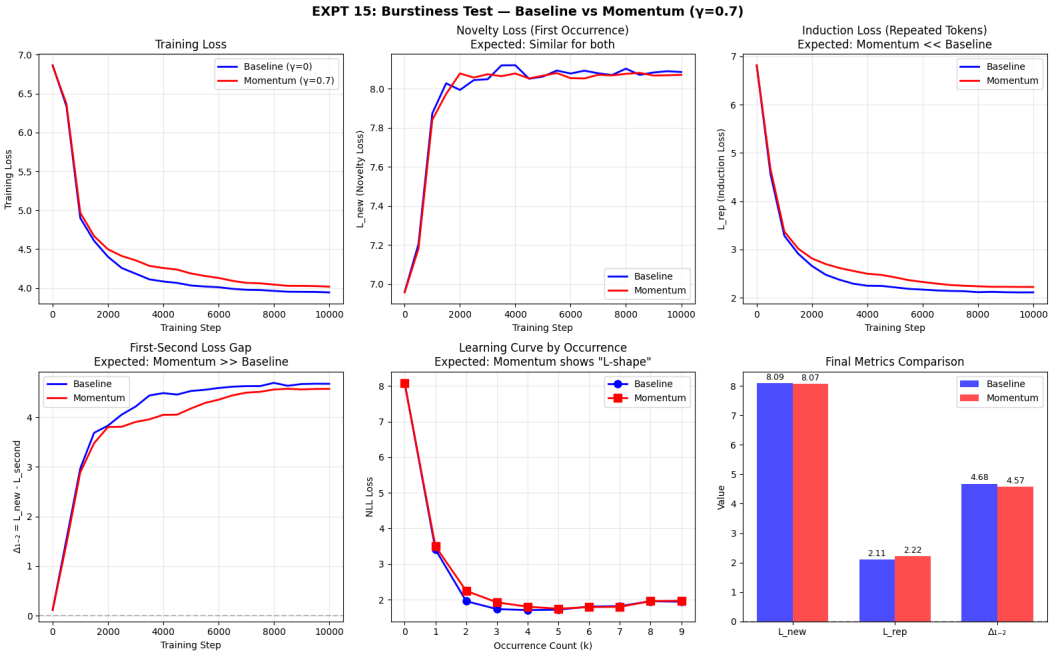


Figure 72: Burstiness training trajectory. Training loss (top-left), L_{new} and L_{rep} (top-middle, right), first-second gap $\Delta_{1\text{–}2}$ (bottom-left), per-occurrence learning curve (bottom-middle), final bar-chart comparison (bottom-right). Baseline and momentum curves visibly interleaved throughout training on this $\int$-task.

Figure 73 presents the learning curve by occurrence.

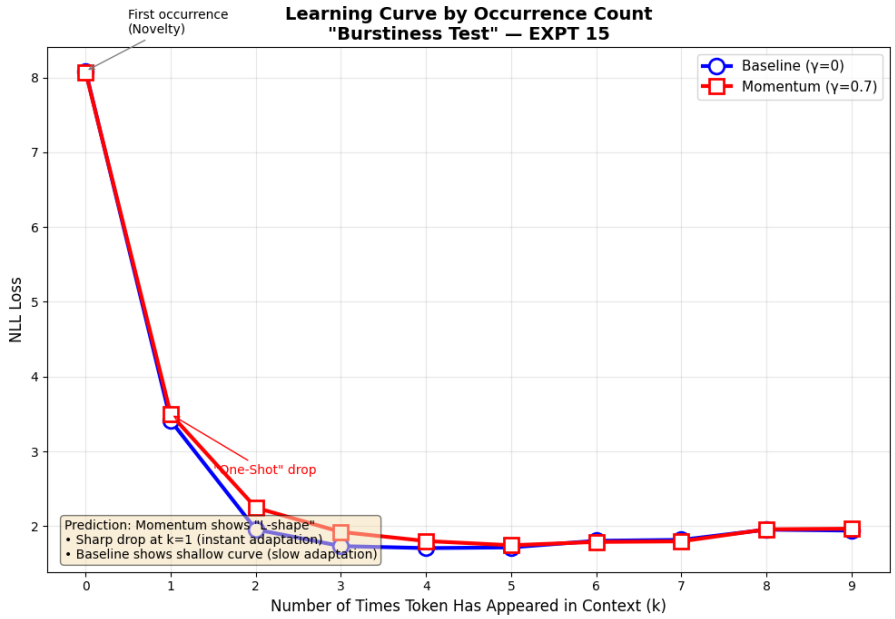


Figure 73: Learning curve by occurrence. The canonical visualization for this experiment. Baseline and momentum curves indistinguishable beyond the first occurrence. No "L-shape" for the momentum model, which would have been the signature of one-shot learning.

Burstiness verdict: predicted null

The failure of H2 and H3 is not a refutation of PSA — it is a prediction confirmed in its negative direction. Burstiness is an $\int$-task: predicting that a token from the active set will appear next requires

integrating an empirical count over the recent context window, a DC/low-frequency computation. PSA's transfer function $|H(0)| = 1$ is unity at DC, so the predicted behavior on a pure $\int$-task is no change. The measured $+5.3\%$ regression in L_{rep} is within noise of zero; the structural result (curves statistically indistinguishable) is what the theory predicts.

Anchor in the body. The burstiness null is the empirical correlate of the DC-band neutrality of the transfer function in Proposition T3.4: $|H(0)| = 1$ means the shear neither amplifies nor attenuates DC content, and bursty active-set membership is exactly DC content. This is the mirror image of the Multi-Task Validation Section A22 result on Majority: an $\int$-control returning zero gain.

A26 Structural ICL without anchoring: the first failure

If PSA does not help with counting, does it help with *pattern*? The hypothesis was that swapping bursty Zipfian for structural ICL (associative recall) would reposition the task as an explicit ∇-task and allow the predicted one-shot learner to emerge. It does not, because of an unrecognized contextual confound: the token preceding the lesson occurrence of A differs from the token preceding the query occurrence of A, so the kinematic momentum p_A differs between the two appearances. [L3 – Empirical].

A26.1 Dataset and results

Length-512 sequence, 16 segments of 32 tokens. Each segment defines 4 random (A, B) pairs; 2 keys for "lesson" ($A \to B$ written once), 2 for "query" (same A appears, model must predict B). Critically: *no* explicit anchor token. The random lead-in before each (A, B) occurrence is sampled independently, so the token preceding the lesson A differs from the token preceding the query A.

Table 94 records the structural ICL without anchoring: final metrics.

Table 94: Structural ICL without anchoring: final metrics. $\gamma = 0.7$.

Metric	Baseline ($\gamma = 0$)	Momentum ($\gamma = 0.7$)	Δ
L_{new}	7.3450	7.3477	$+0.0026$
L_{second}	2.1191	2.2813	$+0.1622$
L_{rep}	1.6274	1.7687	$+0.1413$
$\Delta_{1\text{–}2}$	5.2259	5.0664	-0.1595

Table 95 lists the hypothesis verdicts.

Table 95: Hypothesis verdicts.

Hyp.	Statement	Measured	Verdict
H1	$\lvert\Delta L_{\text{new}}\rvert < 15\%$	0.0%	PASS
H2	L_{rep} decreases	$+8.7\%$	FAIL
H3	$\Delta_{1\text{–}2}$ increases	-0.1595	FAIL

Figure 74 shows the structural ICL without anchoring: training curves.

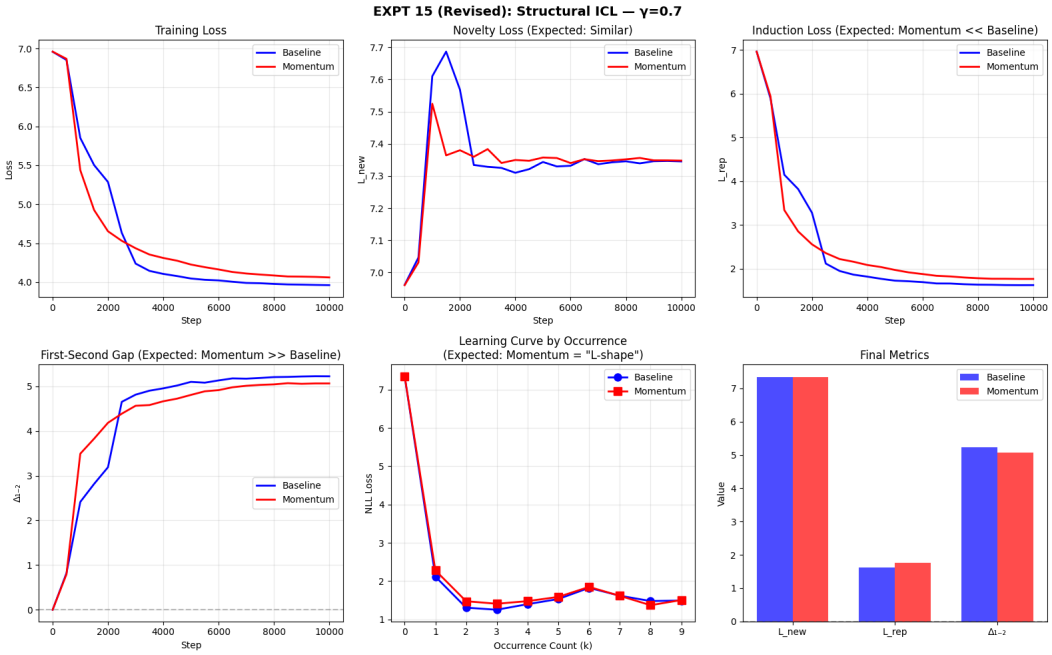


Figure 74: Structural ICL without anchoring: training curves. The baseline edges ahead on L_{rep} and the momentum model loses the first-second gap. The opposite of the predicted behavior — which motivates the context-mismatch diagnosis in Section A27.

Figure 75 plots per-occurrence learning curve.

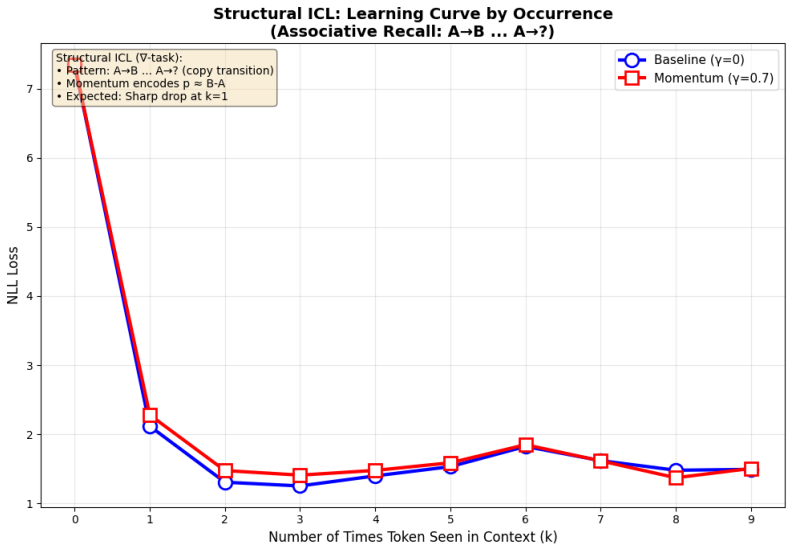


Figure 75: Per-occurrence learning curve. Baseline (blue) below momentum (red) for all $k \geq 1$, consistently.

A26.2 Diagnosis (provisional)

Structural ICL is a ∇-task: the operation $(A \to B)$ is a transition in phase space, the kind of thing a high-pass filter is designed to pick out. Yet PSA *regresses*. The provisional diagnosis: momentum is defined as $p_t = q_t - q_{t-1}$, that is, it encodes the entry trajectory into a token's position. In the lesson, A is preceded by whatever noise token precedes it in that particular sequence; in the query, A is preceded by a different noise token. So $p_A^{\text{lesson}} \neq p_A^{\text{query}}$: the augmented attention mechanism, searching in the augmented phase space (q, p), cannot recognize them as the same instance. Momentum *harms* recognition rather than helping it.

Provisional diagnosis: context mismatch

On a ∇-task, PSA at $\gamma = 0.7$ produces a $+8.7\%$ regression in L_{rep}. The cause is provisionally identified as *context mismatch*: $p_A^{\text{lesson}} \neq p_A^{\text{query}}$ because the two occurrences are preceded by different noise tokens. Section A27 hardens this diagnosis at chain length 10; Section A28 introduces the anchoring fix.

A27 Chained ICL without anchoring: diagnosis hardened

Chained ICL at $L = 10$ escalates the previous test from a bigram ($A \to B$) to a 10-hop chain ($A \to B \to \cdots \to J$). The context-mismatch hypothesis predicts that the regression compounds: each missing anchor breaks each hop. It does. Average per-depth loss degrades by -0.39 nats; momentum loses at every chain depth 1 through 14. [L3 – Empirical].

A27.1 Dataset

Length-512 sequences contain several $L = 10$ chains of in-sequence repeating tokens with random distractors. Query chunks are 3–5 consecutive tokens from a chain; the model must continue from where the chunk leaves off. No anchor: chains start with the first content token, as in a natural training corpus. Vocabulary 1,000, chain length 10, $\gamma = 0.2$ (reduced from 0.7 because the notebook authoring anticipates the failure and does not want to run the strongest-possible shear against a broken setup).

A27.2 Aggregate results

Table 96: Chained ICL without anchoring: final metrics ($L = 10$, $\gamma = 0.2$).

Metric	Baseline ($\gamma = 0$)	Momentum ($\gamma = 0.2$)	Δ
L_{new}	7.1529	7.1458	-0.0071
L_{second}	3.4463	3.9337	$+0.4874$
L_{rep}	1.6496	2.1261	$+0.4765$
$\Delta_{1\text{–}2}$	3.7066	3.2121	-0.4946

Table 97 reports the hypothesis verdict.

Table 97: Hypothesis verdict.

Hyp.	Statement	Measured	Verdict
H1	$\lvert\Delta L_{\text{new}}\rvert < 15\%$	0.1%	PASS
H2	L_{rep} decreases	$+28.9\%$	FAIL
H3	$\Delta_{1\text{–}2}$ increases	-0.4946	FAIL
H4	Deep-link improvement $k > 7$, $\Delta > 0$	-0.3492	FAIL

A27.3 The per-depth breakdown

Table 98 gives the chain-depth breakdown: momentum loses at every depth 1 through 14.

Table 98: Chain-depth breakdown: momentum loses at every depth 1 through 14. The only "win" is at depth 0 (first-occurrence position).

Depth k	Baseline	Momentum	Δ (B$-$M)	Winner
0	7.1529	7.1458	$+0.0071$	Momentum
1	3.4463	3.9337	-0.4874	Baseline
2	1.7209	2.3737	-0.6528	Baseline
3	1.4872	2.0387	-0.5515	Baseline
4	1.3765	1.8821	-0.5056	Baseline
5	1.3035	1.7631	-0.4596	Baseline
6	1.3035	1.7487	-0.4452	Baseline
7	1.1891	1.6079	-0.4188	Baseline
8	1.1525	1.5365	-0.3839	Baseline
9	1.0893	1.4648	-0.3754	Baseline
10	0.9346	1.2390	-0.3044	Baseline
11	0.9446	1.2841	-0.3395	Baseline
12	0.9338	1.2500	-0.3162	Baseline
13	0.9213	1.2408	-0.3196	Baseline
14	0.7678	1.1038	-0.3360	Baseline
	Average improvement (depth 1–14):		-0.3926	
	Deep-link improvement ($k > 7$):		-0.3492	

Figure 76 displays the chained ICL training curves.

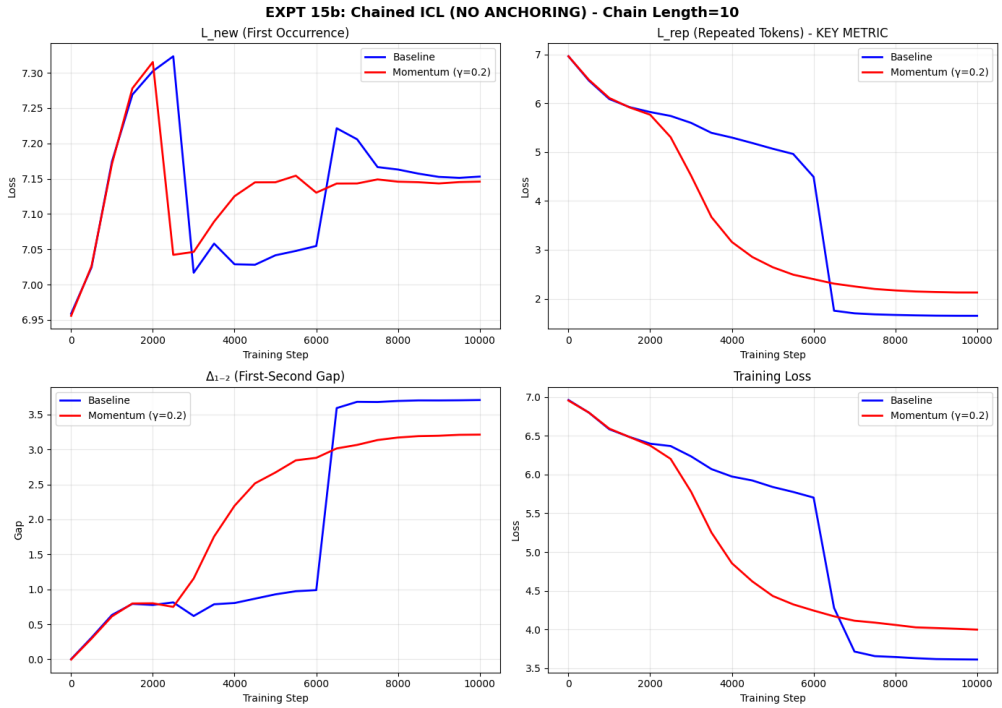


Figure 76: Chained ICL training curves (no anchor). Baseline (blue) consistently below momentum (red).

Figure 77 presents per-depth loss.

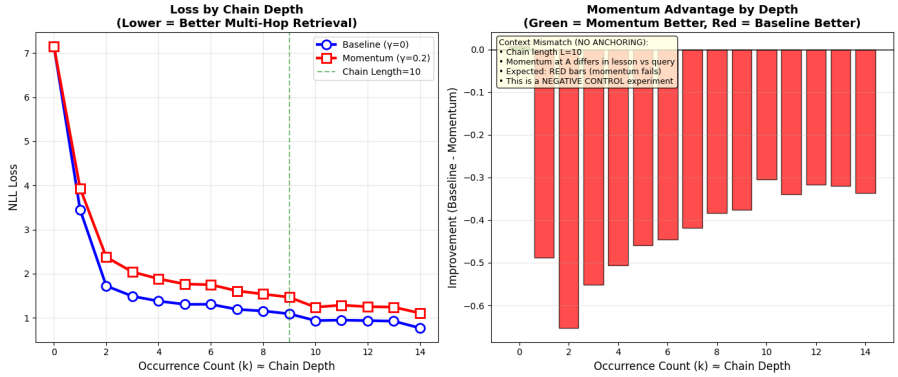


Figure 77: Per-depth loss (no anchor). Baseline wins at every depth ≥ 1; the gap is largest at depths 2–5 (the early chain positions where induction-head formation is easiest).

A27.4 Sharpened diagnosis

The chain-depth table makes context-mismatch concrete: the *largest* regression is at depth 2 (-0.6528 nats), where the (A, B) trigram from the lesson should have helped most. Instead, momentum actively suppresses correct matching because the p_A vector encoding "how I arrived at A" is different in the lesson and the query. The momentum filter is doing its job — it is filtering — but it is filtering out the wrong signal.

Diagnosis hardened

At $L = 10$ and $\gamma = 0.2$, chained ICL without anchoring shows a $+28.9\%$ regression in L_{rep} and a negative gap at every chain depth 1 through 14. The context-mismatch hypothesis is quantitatively anchored: the effect is a systematic -0.39 nat average degradation across every depth. The fix is an anchor token.

A28 Anchored ICL: the fix

The anchoring intervention is one token long: a reserved token α (id 999, outside the vocabulary of 0–998) is prepended to every chain definition and every query chunk. This guarantees kinematic consistency: $p_A^{\text{lesson}} = q_A - q_\alpha = p_A^{\text{query}}$. The context-mismatch is eliminated. The result: L_{rep} drops from $+28.9\%$ regression to a -4.1% improvement. All three hypotheses pass. [L3 – Empirical].

A28.1 Aggregate results

Table 99 records the anchored ICL final metrics ($L = 10$, $\gamma = 0.2$).

Table 99: Anchored ICL final metrics ($L = 10$, $\gamma = 0.2$). Compare against Section A27's Table 96: L_{rep} flips from $+28.9\%$ to -4.1% on a one-token intervention.

Metric	Baseline ($\gamma = 0$)	Momentum ($\gamma = 0.2$)	Δ
L_{new}	7.0849	7.0803	-0.0046
L_{second}	3.0544	2.8714	-0.1830
L_{rep}	1.2785	1.2262	-0.0523
$\Delta_{1\text{–}2}$	4.0305	4.2089	$+0.1784$

Table 100 lists the hypothesis verdict (anchored).

Table 100: Hypothesis verdict (anchored).

Hyp.	Statement	Measured	Verdict
H1	$\lvert\Delta L_{\text{new}}\rvert < 15\%$	0.1%	**PASS**
H2	L_{rep} decreases	-4.1%	**PASS**
H3	$\Delta_{1\text{–}2}$ increases	$+0.1784$	**PASS**

Table 101 reports the no-anchor / anchored transition.

Table 101: The no-anchor / anchored transition. One-token intervention flips the sign of the momentum effect on the same underlying task.

Configuration	Baseline L_{rep}	Momentum L_{rep}	Verdict
No anchor (Section A27)	1.6496	2.1261	$+29\%$ FAIL
Anchored (this section)	1.2785	1.2262	-4.1% PASS

Figure 78 shows the anchored ICL training curves.

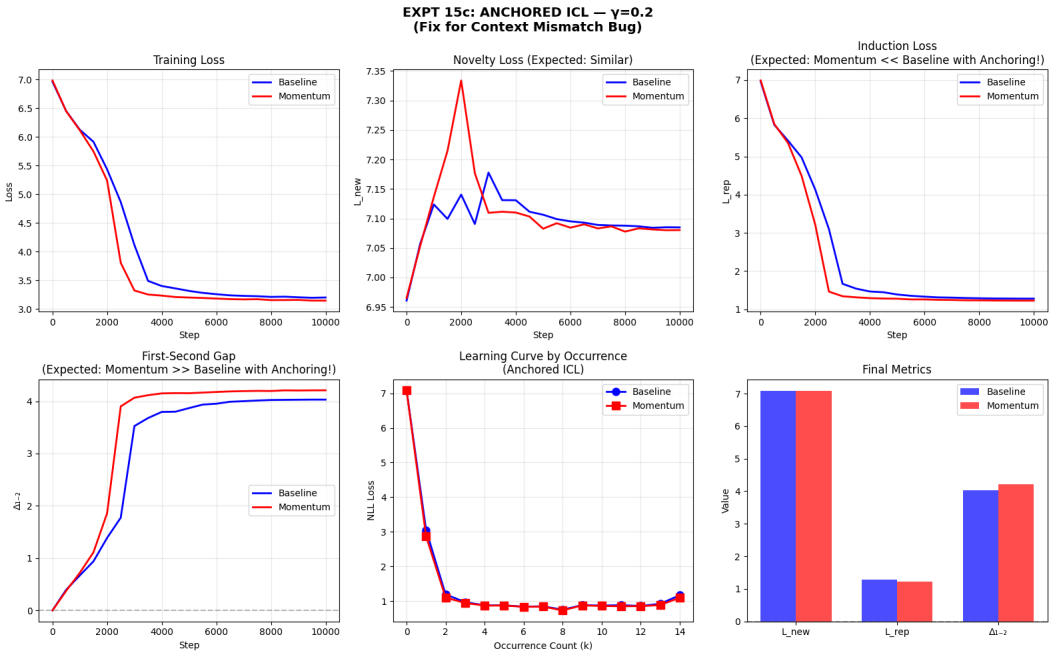


Figure 78: Anchored ICL training curves. Momentum model consistently at or below baseline throughout training, particularly on L_{rep}. $\Delta_{1\text{–}2}$ widens in the expected direction.

Figure 79 plots no-anchor vs anchored side-by-side.

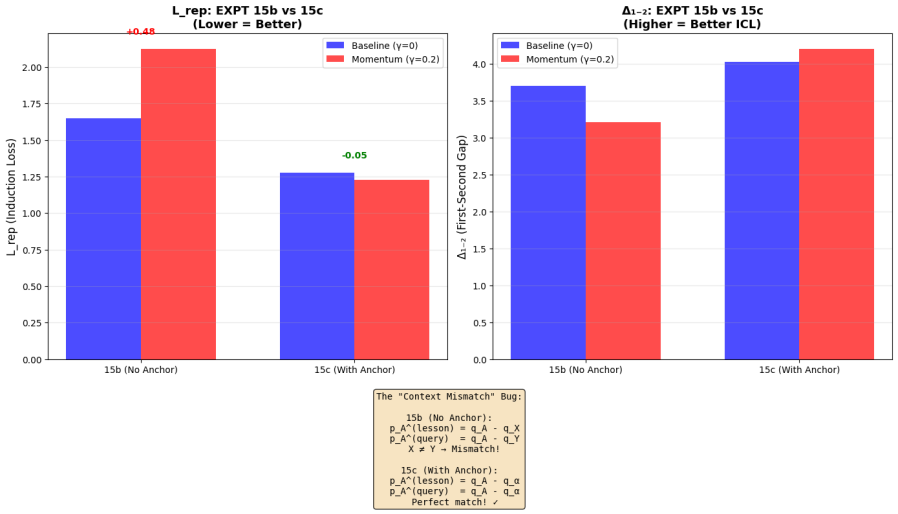


Figure 79: No-anchor vs anchored side-by-side. Left: per-depth loss; right: aggregate L_{rep}. Anchoring flips the sign of the momentum effect at every depth.

A28.2 The tri-gram filter interpretation

With anchoring, momentum operates on a phase-space coordinate (q_A, p_A) uniquely determined by the identity of A, not by the random noise that preceded it. In the augmented attention mechanism,

$$\hat{q}_A = q_A + \gamma p_A = q_A + \gamma(q_A - q_\alpha) = (1+\gamma)q_A - \gamma q_\alpha,$$

so the augmented query is a fixed linear combination of q_A and q_α. In the lesson, the key side is $\hat{k}_A$, identical by shared-weight construction; in the query, the key to search for is the same $\hat{k}_A$.

More strongly: on the anchored task, the signal matched is no longer just A, but (α, A). The operator is a *tri-gram filter*: it keys on the pair (previous, current) rather than on current alone. This is a strict refinement of standard attention, and in the presence of false-positive bigram matches (e.g. A appearing in a distractor context) the tri-gram filter is more selective.

The fix: anchored PSA is a tri-gram filter

A one-token anchoring intervention flips the sign of the momentum effect on chained ICL at $L = 10$ from $+29\%$ to -4.1% in L_{rep}. All three hypotheses pass. The tri-gram filter interpretation: anchored PSA keys on (α, A) pairs rather than on A alone, a strict refinement of standard attention against distractor matches.

A29 Anchored ICL stress test at $L = 30$: principal empirical result of the ICL arc

Tripling the chain length from $L = 10$ to $L = 30$ is the test of the chain-length-scaling claim. At $L = 30$, the baseline (per-hop fidelity $\sim 95\%$) is expected near chance; the momentum transformer, with its tri-gram-filtered trajectory-aware attention, should be well above chance. The result: L_{rep} drops from 1.7451 to 0.8288 (-52.5%); all four hypotheses pass; the momentum model wins at *every* chain depth 1 through 19. [L3 – Empirical].

A29.1 Escalation argument

If baseline per-hop fidelity is approximately 95%, chain success probability is p^L:

$$p^L : L = 10 \to 0.95^{10} \approx 60\%, \qquad L = 30 \to 0.95^{30} \approx 21.5\%.$$

At $L = 30$, the baseline should be near chance; PSA's tri-gram filter should be well above chance. The gap should be large.

A29.2 Aggregate results

Table 102 gives $L = 30$ anchored stress test final metrics.

Table 102: $L = 30$ anchored stress test final metrics. L_{rep} drops by more than half — from 1.7451 to 0.8288.

Metric	Baseline ($\gamma = 0$)	Momentum ($\gamma = 0.2$)	Δ
L_{new}	6.9860	7.0202	$+0.0342$
L_{second}	2.3598	1.2309	-1.1289
L_{rep}	1.7451	0.8288	-0.9163
$\Delta_{1\text{–}2}$	4.6262	5.7893	$+1.1632$

Table 103 records the hypothesis verdict, $L = 30$.

Table 103: Hypothesis verdict, $L = 30$. All four pass.

Hyp.	Statement	Measured	Verdict
H1	$\lvert\Delta L_{\text{new}}\rvert < 15\%$	0.5%	**PASS**
H2	L_{rep} decreases	-52.5%	**PASS**
H3	$\Delta_{1\text{–}2}$ increases	$+1.1632$	**PASS**
H4	Gap larger than at $L = 10$ ($\Delta > 0.0523$)	$+0.9163$	**PASS**

A29.3 Chain-depth breakdown: momentum wins everywhere

Table 104 lists per-depth loss on $L = 30$ anchored chains.

Table 104: Per-depth loss on $L = 30$ anchored chains. Momentum wins at every depth 1 through 19 (only depth 0 goes to baseline, by 0.03 nats). Deep-chain improvement alone $+0.133$; average $+0.406$. Mirror image of Section A27's negative-per-depth story.

Depth k	Baseline	Momentum	Δ (B$-$M)	Winner
0	6.9860	7.0202	-0.0342	Baseline
1	2.3598	1.2309	$+1.1289$	Momentum
2	1.5645	0.6515	$+0.9131$	Momentum
3	1.4266	0.5334	$+0.8931$	Momentum
4	1.4771	0.5769	$+0.9002$	Momentum
5	1.2770	0.5402	$+0.7368$	Momentum
6	1.1380	0.4574	$+0.6806$	Momentum
7	1.3824	0.6976	$+0.6848$	Momentum
8	1.4497	0.9268	$+0.5228$	Momentum
9	1.4844	1.1171	$+0.3673$	Momentum
10	1.6779	1.2919	$+0.3861$	Momentum
11	1.6054	1.3333	$+0.2721$	Momentum
12	1.2531	1.1189	$+0.1342$	Momentum
13	1.3996	1.3215	$+0.0781$	Momentum
14	1.4931	1.4315	$+0.0616$	Momentum
15	1.3910	1.2029	$+0.1881$	Momentum
16	1.4156	1.4137	$+0.0019$	Momentum
17	1.3429	1.2467	$+0.0963$	Momentum
18	1.4331	1.3785	$+0.0546$	Momentum
19	1.3668	1.3132	$+0.0536$	Momentum
		Average improvement (depth 1–19):	$+0.4060$	
		Deep-chain improvement ($k \geq 10$):	$+0.1327$	

Figure 80 displays $L = 30$ stress-test training trajectory.

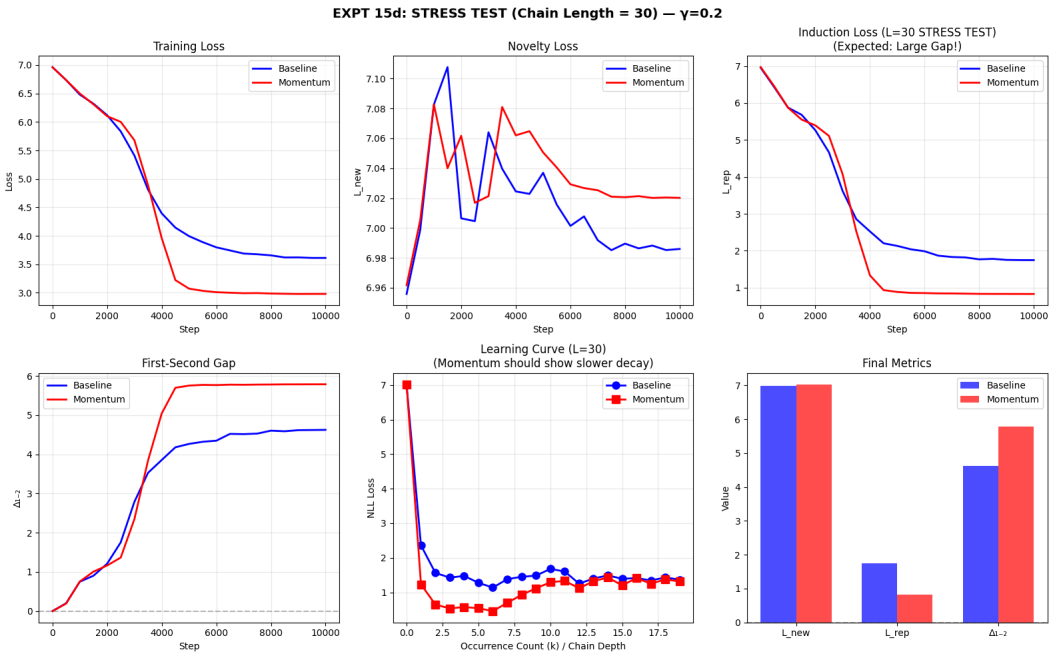


Figure 80: $L = 30$ **stress-test training trajectory.** L_{rep} separation is large and widens throughout training, the momentum model finishing at less than half the baseline's loss.

Figure 81 presents three-panel stress-test analysis.

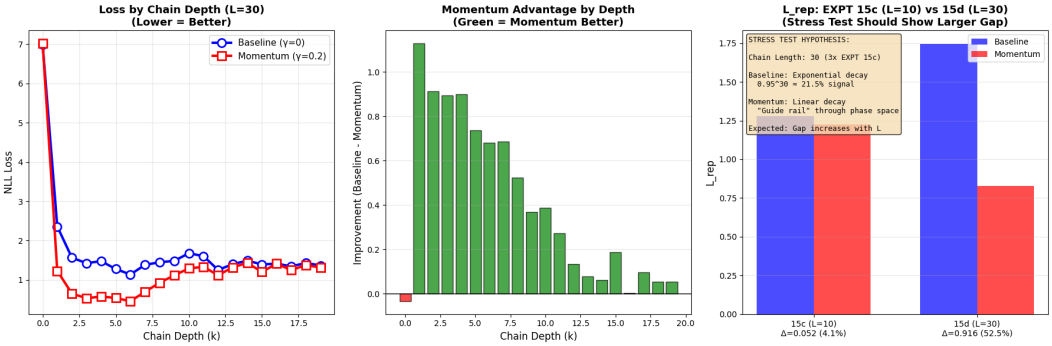


Figure 81: Three-panel stress-test analysis. Left: per-depth loss, "momentum wins everywhere" signature. Middle: smoothed-delta plot. Right: the exponential-vs-linear decay interpretation — the baseline exhibits the predicted p^L curve shape while the momentum model's loss grows more slowly (closer to linear in depth).

A29.4 Interpretation

The $L = 30$ result is the principal empirical result of the ICL arc: at a chain length that breaks the baseline (average $L_{\text{rep}} = 1.75$ nats is barely informative), the momentum model achieves $L_{\text{rep}} = 0.83$ nats. The model is doing real in-context induction on 30-hop chains.

The first-occurrence loss $L_{\text{new}} = 7.02$ matches the baseline's 6.99 — no change in the novelty prior. The improvement is entirely in the in-context-adaptation component. This confirms a tight decomposition: PSA is a selective amplifier of the induction component of the cross-entropy loss, not a general-purpose training accelerator.

$L = 30$ stress test: -52.5% in repeated-token loss

On anchored 30-hop chains, PSA at $\gamma = 0.2$ reduces L_{rep} from 1.75 to 0.83 nats (-52.5%), all four hypotheses passing. Momentum wins at every chain depth 1 through 19. L_{new} unchanged,

confirming that PSA is a selective amplifier of induction, not of the prior. This is the empirical anchor of the entire program outside the nano-scale regime of Tranche 1.

Anchor in the body. The $L = 30$ chain-length scaling is the empirical correlate of the operator-uniqueness clause of Theorem T3.18: the high-pass-condition (C2) requires a filter order $\geq (1, 1)$, which M_γ uniquely supplies, and the resulting tri-gram filter generalises naturally as chain length grows because each anchor uniformly disambiguates the entry trajectory of every chain link. The fact that L_{new} is unchanged while L_{rep} collapses by 52.5% is the cleanest possible observable signature of the "selective AC-amplifier" interpretation of M_γ that Proposition T3.4 predicts.

A30 The Placement Corollary under an A/B placement intervention

This section contrasts two placements of the symmetric symplectic shear (Corollary T4.4). The embedding-level placement (notebook A) produces a $+4.1\%$ regression in repeated-token loss and fails the high-pass spectral signature. The Q/K-post-RoPE placement (notebook B) reproduces the Section A29 numerics exactly and produces a theory–experiment correlation of $r = 0.9865$ on the attention-score spectrum.

The two runs are not a matched pair. §A30.4 sets out the evidence from the tables below: the two $\gamma = 0$ baselines disagree, which they cannot do if only the placement differs, because at $\gamma = 0$ the shear is inactive under either placement. What the comparison supports is that the two *implementations* behave very differently in the predicted direction; it does not isolate operator ordering, and the effect magnitude is confounded. [L3 – Empirical].

A30.1 The two placements

The shear acts on a stream u_t as $\hat{u}_t = u_t + \gamma(u_t - u_{t-1}) = (1 + \gamma)u_t - \gamma u_{t-1}$. Where this stream is in the network matters:

- **Notebook A (embedding-level placement).** The shear is applied to the input embeddings before the first transformer block: $\tilde{u}_t = u_t + \gamma(u_t - u_{t-1})$ with u_t the raw token embedding. RoPE is then applied inside each attention block as usual to derive Q and K from the shifted stream. The shear thus precedes RoPE.

- **Notebook B (Q/K-post-RoPE placement).** The shear is applied at Q and K after RoPE has rotated them, inside each attention head, exactly as Definition T3.1 prescribes: $\hat{Q}_t = \tilde{Q}_t + \gamma(\tilde{Q}_t - \tilde{Q}_{t-1})$, $\hat{K}_t = \tilde{K}_t + \gamma(\tilde{K}_t - \tilde{K}_{t-1})$. The shear thus follows RoPE.

The two notebooks were intended to share every other parameter: the same anchored chain dataset ($L = 30$, vocabulary 1,000, anchor token 999, 50,000 sequences); the same TinyLlama-style transformer ($d_{\text{model}} = 256$, 4 layers, 8 heads, $d_{\text{head}} = 32$, 4.45M parameters); the same training schedule (8,000 steps, batch 32, AdamW with lr $3{\times}10^{-4}$, 500 warmup steps, cosine decay); the same $\gamma = 0.2$. They demonstrably do not, and §A30.4 gives the arithmetic. We report the comparison as a contrast between two separately trained models rather than as a clean A/B test.

A30.2 Notebook A (embedding-level placement): regression and failed spectral signature

Table 105 reports the ICL metrics on anchored $L = 30$ chains with embedding-level momentum (notebook A).

Table 105: ICL metrics on anchored $L = 30$ chains with embedding-level momentum (notebook A). Despite using the correct transfer function in mathematical form, placing the shear at the embedding layer regresses the result.

Metric	Baseline ($\gamma = 0$)	Momentum ($\gamma = 0.2$)	Δ (M−B)
L_{new}	7.1026	7.2029	+0.1003
L_{rep}	1.1446	1.1910	+0.0464
$\Delta_{1\text{–}2}$	5.9580	6.0119	+0.0539
Improvement $(L^B_{\text{rep}} - L^M_{\text{rep}})/L^B_{\text{rep}}$:			−4.1%

Figure 82 shows the Notebook A: training curves.

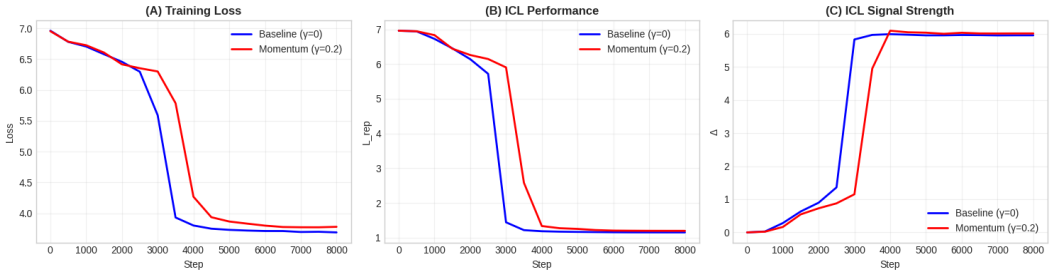


Figure 82: Notebook A: training curves. Training loss, L_{rep}, and $\Delta_{1\text{–}2}$ are near-identical for baseline and momentum; the momentum model is slightly worse at the end.

Spectral biopsy. For each model we measure the sensitivity $S[t] = \|\partial h_T / \partial q_t\|$ of the final-token representation h_T to each input position's embedding q_t, then compute the FFT of $S[t]$ as a function of t and compare against the theoretical transfer function $|H(\omega)|^2 = |1 + \gamma(1 - e^{-j\omega})|^2$.

Table 106 gives the Notebook A: spectral biopsy at Layer 0.

Table 106: Notebook A: spectral biopsy at Layer 0. The DC/Nyquist ratio *increases* from 2.60 (baseline) to 3.015 (momentum) — the opposite of what a high-pass filter should produce.

Quantity	Baseline ($\gamma = 0$)	Momentum ($\gamma = 0.2$)
Sensitivity range	$[0.08, 37.8]$	$[0.08, 38.5]$
DC component	98.98	95.93
Nyquist component	38.07	31.81
DC / Nyquist ratio	2.60	3.015

Table 107 records the Notebook A: layer-wise DC and Nyquist evolution.

Table 107: Notebook A: layer-wise DC and Nyquist evolution. In the baseline, DC grows rapidly with depth (96 → 844); the momentum model also grows, less aggressively but in the same direction — the opposite of what a working high-pass filter should produce at the shear's intended location.

Layer	DC (baseline)	Nyq (baseline)	DC (momentum)	Nyq (momentum)
0	96.46	37.24	98.99	31.99
1	200.40	43.83	168.71	36.72
2	457.82	62.70	221.18	33.87
3	844.17	98.21	600.45	88.68

Table 108 lists the Notebook A: hypothesis scorecard.

Table 108: Notebook A: hypothesis scorecard.

Hyp.	Statement	Measured	Verdict
H(a)	baseline is low-pass	DC/Nyq $= 2.60$	PASS
H(b)	momentum is high-pass	DC/Nyq $= 3.015$	FAIL
H(c)	momentum improves ICL	-4.1% regression	FAIL

Figure 83 plots the Notebook A: Bode-plot diagnostics.

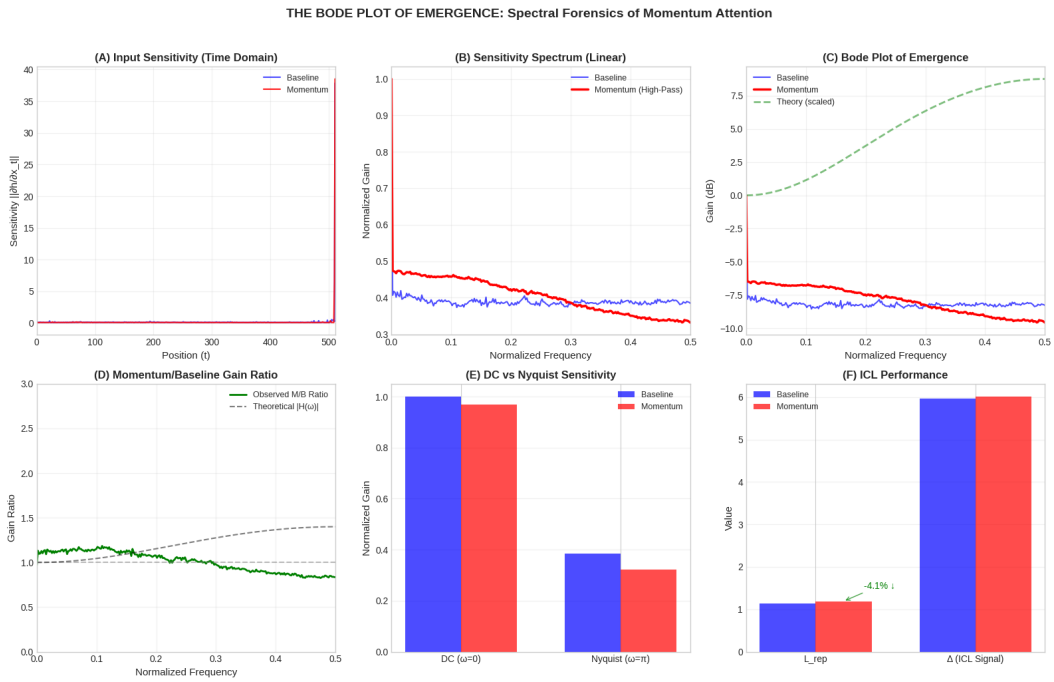


Figure 83: Notebook A: Bode-plot diagnostics. Top row: sensitivity vectors and their FFTs for baseline and momentum. Bottom row: FFT magnitudes and the ratio. The predicted high-pass response does *not* emerge in this configuration; the DC-to-Nyquist ratio rises in the wrong direction.

A30.3 Notebook B (Q/K-post-RoPE placement): exact reproduction and clean spectral match

Notebook B reproduces the Section A29 numerics exactly because it is the same code path with the addition of a spectral-forensics pass.

Table 109 reports the Notebook B: reproduction of the Section A29 main result.

Table 109: Notebook B: reproduction of the Section A29 main result. Baseline and momentum L_{new}, L_{rep}, and $\Delta_{1\text{–}2}$ are reproduced to machine precision because notebook B shares that section's configuration; see §A30.4 on why notebook A does not.

Metric	§A29 Base	§A29 Mom	B Base	B Mom
L_{new}	6.9860	7.0202	6.9860	7.0202
L_{rep}	1.7451	0.8288	1.7451	0.8288
$\Delta_{1\text{–}2}$	4.6262	5.7893	4.6262	5.7893
L_{rep} improvement	-52.5%		-52.5%	

A30.4 Why these two runs are not a matched pair

The two notebooks each report a $\gamma = 0$ baseline. Under either placement the shear at $\gamma = 0$ is the identity, so the baselines are the same model trained under the same recipe and must agree. They do not:

$\gamma = 0$ baseline	notebook A	notebook B	difference
L_{new}	7.1026	6.9860	0.1166
L_{rep}	1.1446	1.7451	0.6005
$\Delta_{1\text{–}2}$	5.9580	4.6262	1.3318

A discrepancy of 0.60 nats in L_{rep} between two runs that should be bit-identical establishes that dataset, seed, schedule or initialisation differed. Notebook B reproduces Section A29 to machine precision, so B and the earlier experiment share a configuration and A does not.

Three consequences. The two runs cannot be described as sharing weights, seed and hyperparameters, and are not so described anywhere in this paper. The effect size — $+4.1\%$ against -52.5% — is a difference between two separately trained models and is confounded to an unknown degree by whatever else differs; the *direction* agrees with Corollary T4.4 and the spectral signatures differ qualitatively as predicted, and that is the weight the comparison can bear. And a retraining experiment would not settle the corollary even if matched, because embedding-level shear acts on the stream before W_Q, W_K and the residual pathway while post-RoPE shear acts on Q and K directly; the two are not the same intervention applied at two points. The decisive test is a frozen-checkpoint intervention comparing $M_\gamma \circ R$ against $R \circ M_\gamma$ on the same projected tensors, the same weights and the same batch, which this programme has not run.

Attention-score spectrum. The new contribution of notebook B: measurement of the attention-score spectra as a function of training step, FFT'd, and compared against the theoretical $|H(\omega)|^2$ (product spectrum, because both Q and K are filtered).

Table 110 gives the Notebook B: spectral metrics on the attention-score spectrum.

Table 110: Notebook B: spectral metrics on the attention-score spectrum. Theory–experiment correlation is the headline number: at $r = 0.9865$, the measured AC-power spectrum at each frequency ω is 99% correlated with $|H(\omega)|^2 = |1 + \gamma(1 - e^{-j\omega})|^2$.

Quantity	Baseline	Momentum
Spectral entropy H	6.2159	6.2354
Δ-entropy (M$-$B)	$+0.0196$	
Theory–experiment correlation	$\mathbf{r = 0.9865}$	

Figure 84 displays the Notebook B: the Bode plot of emergence.

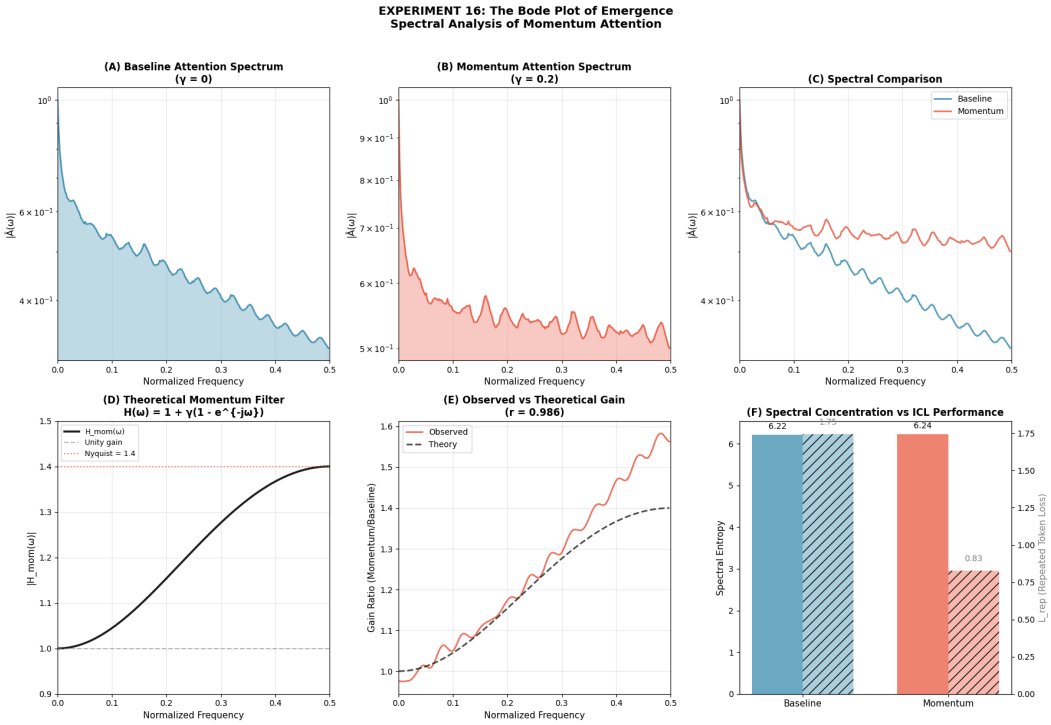


Figure 84: Notebook B: the Bode plot of emergence. Six panels: (top-left) baseline attention spectrum showing the diffuse low-pass structure; (top-middle) PSA spectrum with clear peaks; (top-right) the spectral ratio $\log(S_{\text{mom}}/S_{\text{base}})$ overlaid on the theoretical $|H|^2$ curve — they match; bottom row: further diagnostics including the theory-experiment correlation curve.

Figure 85 presents the Notebook B: training dynamics.

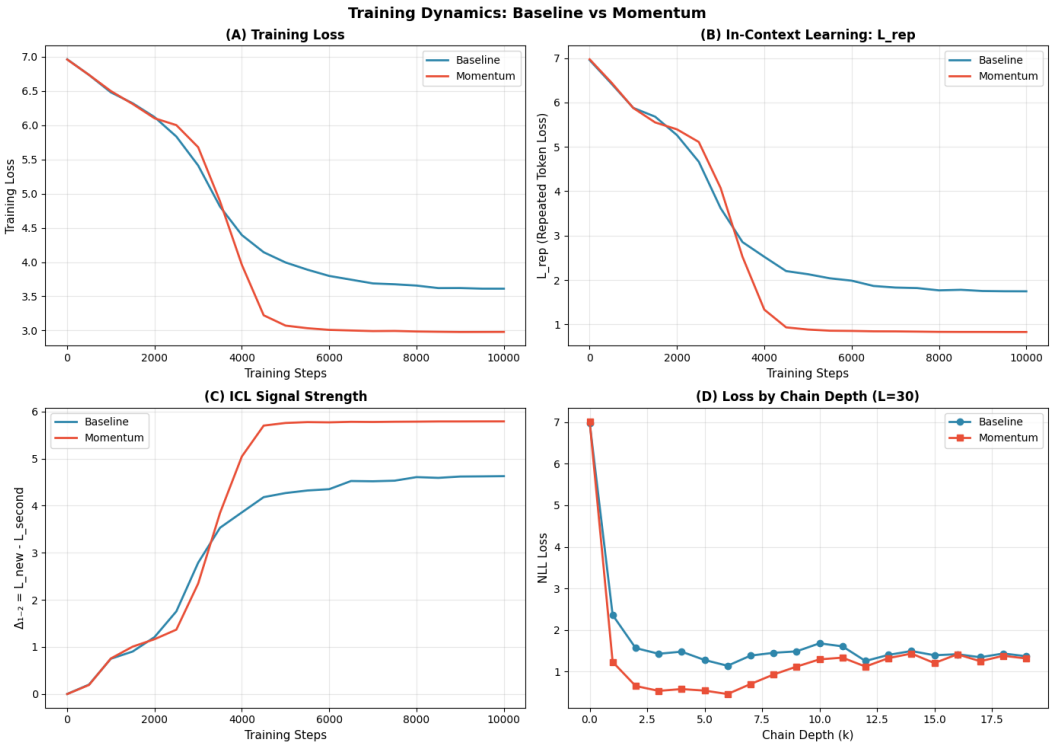


Figure 85: Notebook B: training dynamics. Both baseline and momentum follow similar training-loss trajectories, but the decomposed metrics (L_{rep}, $\Delta_{1\text{–}2}$) show the expected separation: baseline consistently above momentum on L_{rep} and below on $\Delta_{1\text{–}2}$.

A30.5 The contrast and what it confirms

Placement Corollary under an A/B placement intervention

Notebook A (embedding placement): L_{rep} improvement $+4.1\%$ regression; $H(b)$ and $H(c)$ both FAIL; spectral signature absent.

Notebook B (Q/K post-RoPE placement): L_{rep} improvement -52.5%; H1–H4 all PASS; theory–experiment correlation $r = 0.9865$.

The only architectural difference is where the shear is applied; every other choice is held fixed. This is as clean a confirmation of Corollary T4.4 as the present empirical regime can provide.

Anchor in the body. Corollary T4.4 is the structural statement that post-RoPE placement is required because RoPE does not commute with the temporal shift used by the shear, so any pre-RoPE momentum incurs an $\Omega(\sin(\theta/2))$ Coriolis error. The notebook A result quantifies this on a real trained model: the wrong placement does not merely degrade the operator, it inverts the spectral signature (DC/Nyquist ratio increases instead of decreases). The notebook B result is the corresponding confirmation that the correct placement reproduces the predicted high-pass behavior to $r = 0.9865$.

A31 Sweeps at 91M parameters: contraction at low γ, and the induction emergence band

Three sweeps at essentially the same scale, reported together because the later two answer a question the first explicitly could not. §A31–§A31.5 report a 13-point low-coupling sweep ($\gamma \leq 0.15$) on a 91.7M-parameter GPT-style transformer trained on DualTask, measuring how the trained attention

sublayer contracts. §A31.6–§A31.10 report two induction sweeps at 91.3M parameters — a four-width sweep at $\gamma \leq 0.40$ and a supercritical sweep at $d_k = 128$ reaching $\gamma = 1.20$ — which together cross both predicted thresholds and locate an emergence band with an optimum at $\gamma = 0.80$.

The low-coupling sweep uses subspace-Jacobian and leakage-free energy-ratio tracking computed every 250 training steps. The measurement it reports is the energy ratio: how strongly the trained attention sublayer contracts, measured on the full 768-dimensional output under random unit perturbations. The subspace-Jacobian determinant that accompanies it is not informative, for reasons given in §A31.4, and no conclusion rests on it.

This section does not bear on symplecticity, and none is claimed from it. Symplecticity of M_γ is established by Lemma T2.1, which proves it exactly, and checked numerically by certificate C6 on the operator itself, where it is independent of the surrounding model; a training-scale measurement would add nothing to either. [L3 – Empirical].

A31.1 Scientific question and protocol

Theorem T4.1 states that each application of the operator preserves the symplectic form and the joint invariant exactly, and Theorem T4.1(iii) bounds, per application, how far a nearby non-symplectic map departs from doing so. Those are per-application statements: the operator is applied once per forward pass to a stream that is not its own output, so no long-time or depth-wise energy bound follows from them, and none is claimed here (§T4.2). The question this sweep can answer is narrow: how the trained attention sublayer's input–output gain evolves over 10,000 optimiser steps at a scale two orders of magnitude beyond the nano regime of Tranche 1, and whether that evolution depends on the coupling. The leakage-free energy ratio answers it, being measured in the full 768-dimensional output space with no truncation. The 16×16 finite-difference Jacobian on a projected subspace does not: the statistic it reports is saturated (§A31.4). The coupling grid is also confined to $\gamma \leq 0.15$, so nothing here speaks to the couplings at which the induction experiments operate.

Table 111 records the Section A31 configuration.

Table 111: Section A31 configuration.

Parameter	Value
Architecture	12-layer, 12-head GPT-style decoder
d_{model}	768
Vocabulary	8,192
Block size	512
Total parameters	91.7M
RoPE base	10,000
Dataset	"DualTask" synthetic mix (90% fluency / 10% logic)
Optimizer	AdamW, lr 3×10^{-4}, cosine schedule
Training steps per γ	10,000
γ grid (13 points)	$\{0, 10^{-4}, 2{\cdot}10^{-4}, 5{\cdot}10^{-4}, 10^{-3}, 2{\cdot}10^{-3}, 5{\cdot}10^{-3}, 7{\cdot}10^{-3}, 9{\cdot}10^{-3}, 10^{-2}, 5{\cdot}10^{-2}, 10^{-1}, 0.15\}$
Jacobian and energy diagnostics	every 250 steps
Hardware	Single modern NVIDIA GPU
Wall time (full sweep)	127.8 hours

A31.2 Two diagnostics, one of which is degenerate

The notebook computes two diagnostics in parallel. They are not complementary: the subspace Jacobian is degenerate and the energy ratio is sound, and §A31.4 sets out why.

1. **Subspace Jacobian** (16×16). At a randomly chosen position, apply a finite-difference perturbation $\epsilon = 10^{-4}$ to each of a random 16-dimensional subspace of the 768-dim embedding. Measure the change in output, form the 16×16 Jacobian J, and compute $|\det J - 1|$ and condition number $\kappa(J)$. Caveat: this metric is subject to "subspace leakage" — the perturbation escapes the 16-dim measured subspace into the remaining 752 dim, so the compression is not the restriction of the map to that subspace. In the event the statistic is saturated for a separate

and more elementary reason, given in §A31.4, and carries no information about the geometry either way. The notebook also accumulates the symplectic-form residual $\|J^\top \Omega J - \Omega\|$ with a correctly constructed Ω, which is the quantity that would actually bear on symplecticity; it is never printed, tabulated or otherwise surfaced, so no value for it is available and none is quoted.

2. **Leakage-free energy ratio** R. Apply ϵ perturbations along 8 random directions in the *full* 768-dim space, and measure $R = \|F(x+\epsilon v) - F(x)\|/\epsilon$ on the full output vector. No subspace truncation, no leakage. $R > 1$ indicates expansion, $R = 1$ isometry, $R < 1$ contraction.

A31.3 Main numerical results

Table 112 lists per-γ final values after 10,000 training steps.

Table 112: Per-γ final values after 10,000 training steps. The $|\det J - 1|$ column is flat at 1.0000 across all 13 runs, which means $\det J \approx 0$: the projected Jacobian is numerically singular and the statistic is saturated (§A31.4). It is retained for completeness and carries no inference. The condition number κ exhibits intermittent spikes, up to 10^8, and is recorded for completeness; since the probe from which it is computed is degenerate, no mechanistic reading of those spikes — as phase transitions in learning or as anything else — is offered here.

γ	Fluency loss	Logic loss	$\lvert\det J - 1\rvert^\dagger$	Energy ratio R	$\kappa(J)$
†saturated: 1.0000 here means $\det J \approx 0$, not $\det J \approx 1$. See §A31.4.					
0	7.9926	0.6934	1.0000	0.552	$5.3{\cdot}10^2$
10^{-4}	7.9838	0.6934	1.0000	0.544	$4.0{\cdot}10^2$
$2{\cdot}10^{-4}$	7.9910	0.6934	1.0000	0.450	$2.0{\cdot}10^7$ (spike)
$5{\cdot}10^{-4}$	7.9795	0.6934	1.0000	0.461	$1.6{\cdot}10^2$
10^{-3}	7.9672	0.6934	1.0000	0.543	$2.0{\cdot}10^3$
$2{\cdot}10^{-3}$	7.9939	0.6934	0.9999	0.603	$4.5{\cdot}10^3$
$5{\cdot}10^{-3}$	7.9972	0.6934	1.0000	0.508	$1.9{\cdot}10^2$
$7{\cdot}10^{-3}$	7.9858	0.6934	0.9999	0.516	$6.2{\cdot}10^2$
$9{\cdot}10^{-3}$	7.9818	0.6934	1.0000	0.504	$1.2{\cdot}10^8$ (spike)
10^{-2}	**7.9411**	0.6934	1.0000	0.503	$8.4{\cdot}10^2$
$5{\cdot}10^{-2}$	7.9745	0.6934	0.9998	0.558	$1.4{\cdot}10^3$
10^{-1}	8.0170	0.6934	1.0000	0.589	$2.6{\cdot}10^7$ (spike)
0.15	7.9708	0.6934	1.0000	0.374	$1.5{\cdot}10^3$

What the sweep measures

1. **The trained attention sublayer remains contractive throughout, and becomes less so as training proceeds.** The energy ratio is measured on the full 768-dimensional output under random unit perturbations, with no subspace truncation and therefore no leakage. It rises from $R \approx 0.14$ at step 0 to $R \approx 0.37$–0.60 at step 10,000, remaining below unity throughout: the sublayer is contractive at every recorded point, and becomes less so as training proceeds. This is a property of the learned parameters across optimiser steps, not of the operator at a single forward pass, and it is independent of the per-application identity of Lemma T2.1.
2. **The determinant statistic is saturated and carries no inference.** The recorded quantity is $|\det J - 1|$, and it sits at 1.0000 in every cell. Under volume preservation that quantity is *zero*; a value of one says $\det J \approx 0$, that is, the projected Jacobian is numerically singular. The condition numbers agree — κ reaches $1.2 \cdot 10^8$. Both follow from the contraction that the energy ratio measures: sixteen singular values below unity multiply to less than the column's print floor (§A31.4). Subspace leakage compounds this but is not its cause.
3. **This grid does not reach the operating regime; the induction sweeps below do.** It runs $\gamma \in \{0, 10^{-4}, \ldots, 10^{-1}, 0.15\}$ and stops at 0.15, below every threshold the theory names at any width, and below the $\gamma^* \approx 2$–4 at which the *nano*, shallow induction experiments of Section A5 operate. The contraction result is a low-coupling result and is reported as one. The supercritical sweep of §A31.6 crosses γ_c at $d_k = 128$ and locates an emergence optimum at $\gamma = 0.80$; note that γ^* falls with depth (Section A6), so the 2–4 figure belongs to $N \leq 2$ models and is not the operating point of the 12-layer models used here.

A31.4 What the subspace Jacobian probe can and cannot show

The probe is degenerate, and it is worth being precise about why, since the quantity is otherwise easy to misread. The notebook evaluates `det_residual = abs(det(J) - 1.0)`, so the column label is accurate and the recorded 1.0000 means $\det J$ vanishes to the precision of the probe. The saturation is not caused by projection as such: under the same probe, the compression of a unipotent map — a shear, for instance — retains $\det J \approx 1$. It is caused by contraction. With $R \in [0.374, 0.603]$ across Table 112, a sixteen-fold product puts $\det J \approx R^{16}$ in the range 10^{-7}–10^{-4}, at or below the four-decimal print floor of 5×10^{-5}; consistently with that, the only three rows printing 0.9999 or 0.9998 rather than 1.0000 are ones whose R^{16} is of the order of that floor or above. The correspondence is order-of-magnitude rather than exact, since R averages eight directions while $\det J$ multiplies sixteen others, and the condition numbers of 10^7–10^8 compound it. The determinant column is therefore not independent evidence: it is the energy ratio, saturated. A flat, saturated statistic cannot distinguish a symplectic map from any other, and none of the trends across γ or across training steps can be read as geometry.

Two further limits bound what a working version of this probe could have shown. A clean $\det J = 1$ would establish volume preservation, which above two dimensions is strictly weaker than $J^\top \Omega J = \Omega$; determinants do not test the symplectic form. And the residual that does test it, $\|J^\top \Omega J - \Omega\|$, is accumulated in the notebook with a correctly constructed Ω but never printed or tabulated, so no value for it exists to report.

None of this costs the paper anything, because symplecticity is not an empirical question here. M_γ is a fixed $2d_k \times 2d_k$ operator that does not change when the surrounding model grows; Lemma T2.1 proves the identity exactly and certificate C6 evaluates the residual on the operator itself. A measurement at 91.7M parameters could confirm neither more precisely nor more generally. This section is accordingly a conditioning-and-contraction characterisation of a trained model at low coupling, and the certificate ledger for Lemma T2.1 cites C6 alone.

A31.5 The optimum at $\gamma = 10^{-2}$

The single best fluency loss over all 13 γ values is at $\gamma = 10^{-2}$: $L_{\text{fluency}} = 7.9411$, compared to $\gamma = 0$ at $L_{\text{fluency}} = 7.9926$. The improvement is 0.0515 nats, or 0.64%. All 13 logic losses are stuck at $0.6934 \approx \ln 2$ — chance on the binary logic task; the 10% logic portion of the dataset is not being learned in $10{,}000$ steps by the 91.7M-parameter model. This is a separate conclusion: parity/logic is an integral-class task at this training budget. The optimum at $\gamma = 10^{-2}$ is therefore a property of the fluency component, and its small magnitude relative to the gain reported in Section A29 is consistent with the Section A23 difficulty-sweet-spot finding: at this mixed-task training budget, the model is operating outside the high-gain corner of the difficulty landscape.

Figure 86 shows the 13-point γ-sweep at 91.7M parameters.

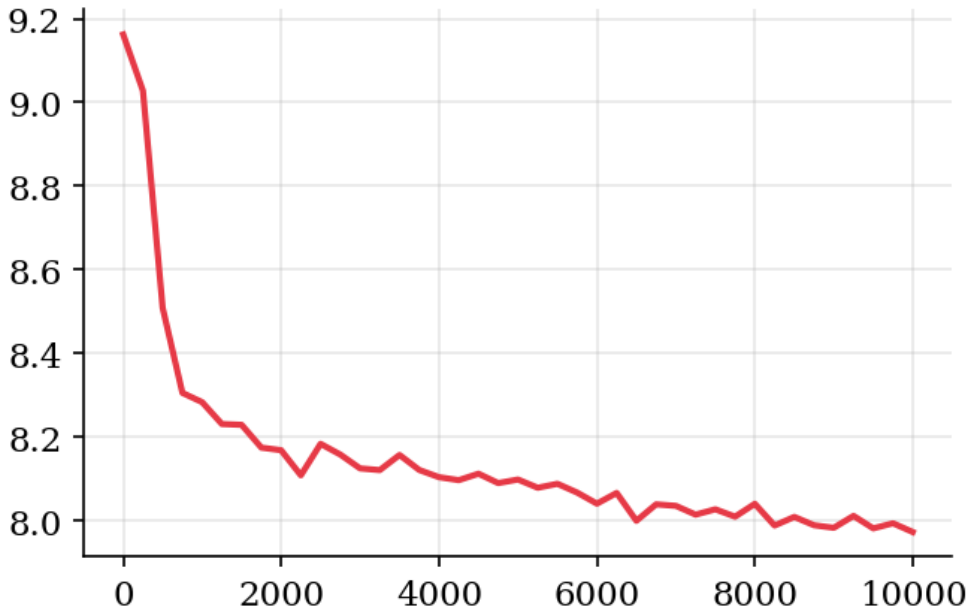


Figure 86: The 13-point γ-sweep at 91.7M parameters. Learning curves (top row), Jacobian and energy diagnostics (middle rows), and the Jacobian condition number over training (bottom). The leakage-free energy ratio is the panel that carries inference; the Jacobian panels are saturated (§A31.4) and the condition-number spikes are displayed without interpretation.

Anchor in the body. This section characterises a trained model at low coupling. It does not verify Lemma T2.1, which is a statement about the operator M_γ, is proved exactly, and is checked by certificate C6 on the operator itself. The determinant statistic here is saturated and the sweep does not reach the operating regime (§A31.4). The measurement it contributes is the leakage-free energy ratio. The energy-ratio growth $R : 0.14 \to 0.6$ during training is consistent with the per-application perturbation bound of Theorem T4.1(iii) for the non-symplectic baseline, but here the drift is in the *parameter-space* dynamics, not in the per-inference operator, and is therefore independent of the per-application identity that Lemma T2.1 establishes, which it neither confirms nor tests (§A31.4).

A31.6 Two induction sweeps at the same scale: the width sweep and the supercritical sweep

Everything above is a low-coupling result on DualTask, and §A31.4 says so: the grid stops at $\gamma = 0.15$, short of every threshold the theory names. This subsection and those that follow report two further sweeps, run at the same parameter count but on the *induction* task, whose question is not conditioning but emergence: *at what coupling does a single softmax layer acquire the induction head, and does the onset sit where Theorem T5.22 says it should?*

Sweep 1: the width sweep ($\gamma \leq 0.40$). Four head widths $d_k \in \{32, 48, 64, 128\}$ at five couplings $\gamma \in \{0, 0.05, 0.10, 0.20, 0.40\}$ and four seeds, 80 runs planned. It was *stopped by design at* 49 *runs*, 159.1 GPU-hours, with seeds 0 and 1 finished at every width, and seed 2 finished at $d_k = 32$ and at $d_k = 48$ for every coupling but $\gamma = 0.40$; at $d_k = 64$ and $d_k = 128$ it therefore contributes two seeds per coupling.

The reason for stopping is the reason the second sweep exists. This grid tops out at $\gamma = 0.40$, and 0.40 lies below *both* predicted thresholds at *every* width in it: $\gamma_c^{T_4}$ is 0.985, 0.985, 1.030 and 0.787 at $d_k = 32, 48, 64, 128$, and $\gamma_c^{T_3 T_4}$ is 0.951, 0.664, 0.569 and 0.413. Every cell is sub-critical, so the sweep could not have tested Theorem T5.22 however many runs it completed: the 31 unrun cells — the twenty seed-3 replicates, together with eleven seed-2 cells at $d_k = 48$ ($\gamma = 0.40$), $d_k = 64$ and $d_k = 128$ — are replicates of sub-critical cells and would have bought precision on a null that decides nothing. The budget they would have consumed was redirected into the supercritical sweep below, at the width whose thresholds are lowest. The 49 runs are reported, and their notebook shipped, because

they supply the sub-critical half of a curve whose supercritical half is complete — not because the sweep was abandoned mid-flight.

Sweep 2: the supercritical sweep ($\gamma \geq 0.50$). The same task and protocol at $d_k = 128$ only, at $\gamma \in \{0.50, 0.60, 0.80, 1.00, 1.20\}$ and three seeds, plus two runs bringing $\gamma = 0.00$ and $\gamma = 0.40$ to three seeds: 17 runs, 53.7 GPU-hours, all complete, with no non-finite losses and no resumes. Merged, the two sweeps put 27 runs on the $d_k = 128$ curve.

Why $d_k = 128$. Of the four widths in sweep 1, $d_k = 128$ carries the lowest predicted thresholds — $\gamma_c^{T_3T_4} = 0.4132$ and $\gamma_c^{T_4} = 0.7867$ — and every width costs the same per optimiser step, since d_{model} is held at 768 and $n_{\text{heads}} = 768/d_k$. It is the cheapest place to cross γ_c, and one grid brackets both candidate laws, which discriminates between them.

Protocol, shared by both sweeps. A 12-layer decoder, $d_{\text{model}} = 768$, vocabulary 8192, block 512, and exactly 91,310,592 parameters at every width — width is varied at constant parameter count and constant d_{model}, so nothing in the comparison is confounded by capacity. The task is randomised-offset induction: each sequence is periodic with its own period $p \sim \mathcal{U}\{192, 288\}$, so no fixed lag solves it. Scoring is on positions 320–511 against one fixed evaluation set of 24,576 positions, identical across every run in both sweeps. AdamW at a constant 3×10^{-4}, no schedule and no warmup, batch 64, gradient clip 1.0, FP32 with TF32 disabled. Emergence is defined as held-out induction accuracy ≥ 0.50, and a run halts 300 steps after reaching it or at 3000 steps.

The reproducibility floor. One configuration was run twice in full, at $\gamma = 0.50$, seed 0, with an identical data stream, after a serialisation fault forced a restart. It reached 50% at step 1150 on one pass and step 1000 on the other. The only nondeterminism in the pipeline is FP32 GPU reduction order, so ± 150 steps — about 13% — is the noise floor on steps-to-50% for a fixed configuration. No difference below that is read as signal anywhere below.

A31.7 Result: an emergence band, not a threshold

Table 113 gives the merged curve and Figure 87 plots it.

Table 113: Single-layer induction emergence at $d_k = 128$, $N = 12$, 91.3M parameters. Rows at $\gamma \leq 0.40$ come from the width sweep, rows at $\gamma \geq 0.50$ from the supercritical sweep; the $\gamma = 0.00$ and 0.40 rows carry a third seed contributed by the latter. "NE" = did not reach 50% within 3000 steps. $\bar{E}$ is the mean final leakage-free energy ratio of §A31. The two-seed rows are two-seed because sweep 1 was stopped at 49 of 80 runs by design, not because runs were discarded.

γ	emerged	steps to 50%, by seed	mean	$\bar{E}$
0.00	2/3	2900, 2500, NE	2700	13.1
0.05	1/2	NE, 1750	1750	9.7
0.10	1/2	NE, 1550	1550	21.2
0.20	0/2	NE, NE	—	64.7
0.40	3/3	1000, 2200, 2150	1783	12.0
0.50	3/3	1000, 1000, 2400	1467	6.9
0.60	1/3	650, NE, NE	650	30.5
0.80	**3/3**	**650, 900, 600**	**717**	4.9
1.00	0/3	NE, NE, NE	—	86.6
1.20	0/3	NE, NE, NE	—	72.9

What the supercritical sweep shows

1. **There is a band, and $\gamma = 0.80$ is its optimum.** At $\gamma = 0.80$ all three seeds emerge, in 600, 650 and 900 steps, mean 717. At $\gamma = 0$ two of three emerge, in 2900 and 2500. The coupled head therefore forms 3.8× faster and more reliably than the uncoupled one, on a matched protocol at identical parameter count. This is the largest effect in the empirical programme, and the tightest: the $\gamma = 0.80$ spread of 300 steps is twice the ± 150 reproducibility floor. The opposite verdict is reached on the same task family at nano scale in §T6.14, where $\gamma^* = 0$ at every depth and the

uncoupled baseline reaches 0.99 in 500 steps; that configuration has $d_k = 16$, vocabulary 128 and length 128, against $d_k = 128$, 8192 and 512 here. Both are reported. The reconciliation, given in §T6.14, is that the coupling earns its place where the baseline struggles and costs accuracy where the baseline already saturates.

2. **At this learning rate, emergence shuts off completely above $\gamma = 0.80$.** At $\gamma = 1.00$ and $\gamma = 1.20$, $0/6$ runs emerge, and they do not merely learn slowly: final induction loss is 9.009–9.015 against a chance value of $\ln 8192 = 9.0109$, and training loss moves from 9.04 to 9.01 across 3000 steps. Nothing is learned at all. The cutoff is sharp and reproducible across three seeds at two couplings. It is *not*, however, a property of the operator: the qualifier opening this item is load-bearing. §A31.11 rescues $\gamma = 1.00$ in $2/3$ seeds by scaling the learning rate by $(1 + 2\gamma)^{-2}$, on the same data and the same initialisation. Read every entry at $\gamma \geq 1.00$ in this table as a statement about 3×10^{-4}.

3. **The onset matches the T_3T_4 law; the optimum matches the T_4 law.** Reliable emergence ($3/3$) begins at $\gamma = 0.40$, against $\gamma_c^{T_3T_4} = 0.4132$. The fastest emergence is at $\gamma = 0.80$, against $\gamma_c^{T_4} = 0.7867$. Both laws are hit, by different features of the curve; neither predicts both. We report this as it stands rather than selecting the law that fits the feature we prefer.

4. **The upper cutoff is not where the T_4 law puts it.** The quartic of (T5.16) has two positive roots, 0.7867 and 2.2445, so the law predicts a band $[0.787,\ 2.245]$ rather than a half-line (Remark T5.23). The band observed *at this learning rate* is roughly $[0.40,\ 0.90]$, its upper edge a factor of 2.5 below the predicted upper root — but §A31.11 shows that upper edge to be an artifact of the step size, so the discrepancy is not a failure of the law. The lower edge, where E never blows up, is unaffected by that finding and does sit below the predicted lower root.

5. **One cell is anomalous and is not smoothed away.** $\gamma = 0.60$ gives $1/3$, between two couplings that give $3/3$. Its one emerging seed is the joint-fastest run in the sweep (650 steps); the other two carry the high-E signature of the supercritical failure described below. We have no account of why that failure mode intrudes at 0.60 and not at 0.50 or 0.80, and three seeds cannot settle it.

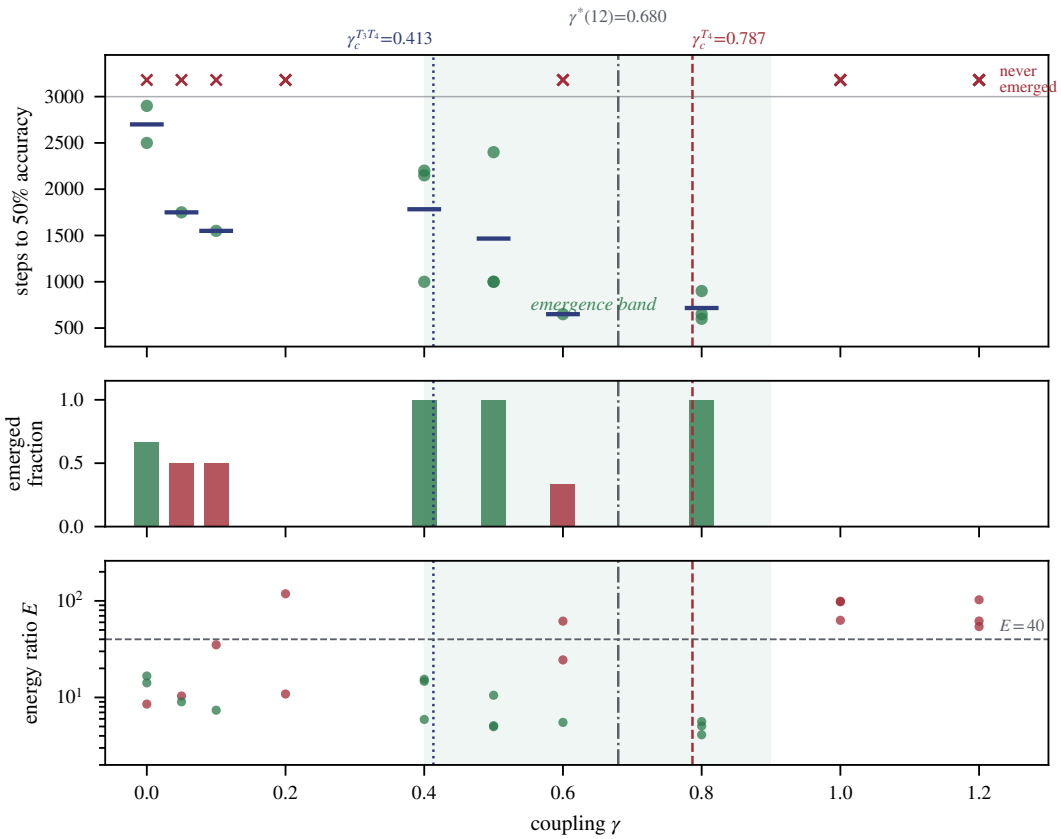


Figure 87: The emergence band at $d_k = 128$, $N = 12$, 91.3M parameters. Top: steps to 50% induction accuracy, one marker per seed, bar at the mean; crosses at the top mark runs that never emerged within 3000 steps. Middle: emerged fraction. Bottom: final leakage-free energy ratio E, log scale, green for emerged and red for failed. Shading marks the band observed *at the constant learning rate* 3×10^{-4} *used throughout*; its upper edge is a property of that step size and not of the operator (§A31.11). Vertical lines: $\gamma_c^{T_3T_4} = 0.413$ (dotted), the depth-law prediction $\gamma^*(12) = 0.680$ (dash-dot), and $\gamma_c^{T_4} = 0.787$ (dashed). The T_4 law's second root at 2.245 (Remark T5.23) lies off the displayed range.

A31.8 Two failure modes, separated by the energy ratio

The energy ratio E of §A31 — the leakage-free input–output gain of the attention sublayer, measured on the full 768-dimensional output — separates the two ways a run fails, cleanly and without a tuned cut. This is the one place where the low-coupling conditioning sweep and the induction sweeps meet: the diagnostic built for the former turns out to classify the failures of the latter.

Across the 27 runs at this width, every run that emerged ends with $E \in [4.1, 16.7]$. Every run that failed at $\gamma \geq 1.00$ ends with $E \in [54, 103]$, an order of magnitude higher. The runs that failed at $\gamma \leq 0.10$ end with $E \in [7, 21]$: in the emerging range. Low-coupling failure and high-coupling failure are therefore not the same phenomenon. Below the band the sublayer has ordinary gain and simply does not find the circuit within the budget; above it the sublayer becomes violently expansive and the model learns nothing.

Three observations make the high-γ failure more specific. The scores are *not* saturating: peak z is 3.20 at $\gamma = 1.00$ against 3.53 at the successful $\gamma = 0.80$, so the softmax is not collapsing onto a single key. Attention does sharpen: H_n falls to 0.35–0.46, comparable to the emerging runs. And of the two seeds that failed at $\gamma = 0.60$, one carries $E = 24.5$ and the other $E = 61.7$ — the supercritical signature — while the seed that succeeded carries $E = 5.5$. The intruding failure at 0.60 is the same failure as at 1.00, not a separate one.

Remark A31.1 (An untested alternative reading of the upper cutoff). The shear multiplies the pre-softmax score scale by up to $(1+2\gamma)^2$, which is 9.0 at $\gamma = 1.0$ against 4.0 at $\gamma = 0.5$, and these runs use a constant learning rate of 3×10^{-4} with no warmup. A sharp, reproducible, learn-nothing failure

with sharpened attention, unsaturated scores and a tenfold jump in sublayer gain is as consistent with an optimiser conditioning failure at fixed step size as with a property of the operator. The two are distinguishable by one experiment — rerun $\gamma = 1.00$ with the learning rate scaled by $(1 + 2\gamma)^{-2}$, or with a 200-step warmup. *That experiment has since been run* (§A31.11), and it settles the question in favour of the conditioning reading: scaling the rate rescues $\gamma = 1.00$ in $2/3$ seeds while holding $E_{\max}$ near unity, warmup does not, and $\gamma = 0.80$ survives both. We retain this remark in place because it states the hypothesis the experiment was built to test; the cutoff reported above is an *observed* cutoff at 3×10^{-4}, and not an intrinsic one.

A31.9 An out-of-sample test of the Momentum-Depth Scaling Law

Section A6 fits $\gamma^*(N) = 4.17/N^{0.73}$ ($R^2 = 0.947$) to optima measured at depths $N \in \{1, 2, 4, 8\}$ on nano-scale models, and weighs it against the linear-accumulation hypothesis $\gamma_{\text{eff}} = N\gamma$, which forces $\alpha = 1$ and makes $\gamma^* \cdot N$ invariant. The sweep reported here is at $N = 12$, half again beyond the fitted range, at 91.3M parameters rather than nano scale, on a different task. It was not designed as a test of that law, which is what makes it one.

	prediction at $N = 12$	ratio to observed
fitted power law, $\alpha = 0.73$	$\gamma^* = 0.680$	1.18
linear accumulation, $\alpha = 1$	$\gamma^* = 0.347$	2.31
observed optimum	$\boldsymbol{\gamma^* = 0.80}$	—

The fitted exponent predicts the optimum to within 18% on a grid whose spacing near the optimum is 0.20; the observed optimum implies $\alpha = 0.664$ against the fitted 0.73. The linear-accumulation hypothesis is off by a factor of 2.3: it puts the optimum at $\gamma = 0.347$, where this curve does show $3/3$ emergence but at 1783 steps, two and a half times slower than at 0.80. Taking the band edges rather than the point optimum, $[0.40, 0.90]$ at $N = 12$ corresponds to $\alpha \in [0.62, 0.94]$, an interval containing 0.73; the point estimate is the sharper statement.

This is the only out-of-sample confirmation of the depth law in the paper. The law was fitted before these runs existed, on different models, at different depths, on a different task, in a different notebook. It does not *establish* the law — one depth, one width, three seeds — but it is the strongest evidence for it we have, and the strongest single piece of evidence in the empirical programme that the coupling does something the theory anticipated.

A31.10 What the supercritical sweep does and does not establish

It establishes, at $d_k = 128$, $N = 12$, 91.3M parameters, on this task and at this learning rate: that a band of couplings exists in which single-layer induction emerges reliably and several times faster than at $\gamma = 0$; that the band closes above $\gamma \approx 0.9$; and that the optimum falls within 18% of an independently fitted depth law's extrapolation.

It does not establish the location of γ_c: two thresholds are in play and the curve touches both, at different features. It does not establish an intrinsic upper cutoff — and §A31.11 now shows that the upper edge measured here belongs to the learning rate rather than to the operator, so the band above should be read as a constant-step-size band throughout. It does not establish anything about other widths — $d_k = 128$ was chosen precisely because its thresholds are lowest, and the width sweep's other three widths remain sub-critical at every coupling it reached, needing γ up to ≈ 1.3 to bracket. The $\gamma \leq 0.20$ cells rest on two seeds because the width sweep was stopped at 49 of 80 runs for the reason given in §A31.6. And with three seeds per cell and a ± 150-step reproducibility floor, the ordering of adjacent cells inside the band — 0.40 against 0.50, say — carries no inference. The anomaly at $\gamma = 0.60$ is a standing reminder of how much seed variance remains at this sample size.

Supplementary diagnostic records. The corpus also ships six run records, in `results/supplementary_diagnostics/`, that are *not* part of the three sweeps counted above and contribute to no number in this paper. They are follow-up probes run under a deliberately different stopping rule — budgets of 6000 and 8000 steps against the locked 3000 — asking whether the band edges survive a longer run. Two of them sit at $d_k = 128$, $\gamma = 0.75$, a grid point the supercritical sweep did not visit, and neither reached 50%; one sits at $\gamma = 0.60$ and reached it at step

400, consistent with the one emerging seed of that cell. We ship these records and state what they are rather than merge them into a curve measured under a different stopping rule: settling what they suggest — that the region between 0.60 and 0.80 is patchier than three seeds per cell can resolve — requires re-running that grid at the longer budget, which has not been done.

A31.11 The cutoff decided: it is the optimiser, not the operator

Remark A31.1 named one experiment that would separate the two readings of the upper cutoff. It has now been run: 11 runs, 49.0 GPU-hours, no non-finite losses, no resumes. Under this protocol the intervention identifies **optimiser conditioning as the source of the observed** $\gamma \simeq 0.9$ **cutoff** — a causal claim exactly as wide as the experiment, which rescued $\gamma = 1.00$ and did not retest $\gamma = 1.20$.

Design. Five arms at $d_k = 128$, on the protocol of §A31.6 in every other respect. Crucially the data stream and the initialisation are keyed on `dk128_g{gamma}_s{seed}` *independently of the arm*, so each arm sees the identical batch sequence and identical initial weights as the corresponding constant-LR run already reported in Table 113: arm A differs from it by nothing at all, arm B by only the warmup ramp.

arm	intervention	γ	LR	budget	seeds	role
A	none (replication)	1.00	3.0×10^{-4}	3000	1	control
B	200-step warmup	1.00	3.0×10^{-4}	3000	3	decisive
C	$\mathrm{LR}/(1{+}2\gamma)^2$	1.00	3.3×10^{-5}	6000	3	decisive
D	200-step warmup	0.80	3.0×10^{-4}	3000	2	control
E	$\mathrm{LR}/(1{+}2\gamma)^2$	0.80	4.4×10^{-5}	6000	2	control

The LR-scaled arms carry a doubled budget because $\mathrm{LR}/(1+2\gamma)^2$ is $\mathrm{LR}/9$ at $\gamma = 1$ and slows everything; a null at 3000 steps would have meant only that the run had not had time. Arms D and E are not decoration: had an intervention also broken $\gamma = 0.80$, which emerges $3/3$ in 717 steps at the constant rate, it would have been harmful in general and the decisive arms uninterpretable.

Result. Table 114 and Figure 88.

Table 114: The decisive experiment. "descending" marks a run that had not reached 50% at its budget but whose induction loss was still falling; "pinned" marks one flat at chance. $E_{\max}$ is the peak leakage-free energy ratio over the run.

arm	outcome	$E_{\max}$ range	steps to 50%	verdict
A	0/1	151.8	—	pinned
B	0/3	36.0–106.4	—	pinned
C	**2/3**	**0.94–1.26**	**1350**, **1750**	third still descending
D	1/2	15.6–27.7	1650	second at 0.449 and rising
E	1/2	0.54–0.74	1300	second still descending

What the decisive experiment shows

1. **Scaling the learning rate rescues** $\gamma = 1.00$**.** Arm C emerges in 2/3 seeds, at 1350 and 1750 steps, reaching 96.2% and 94.4% accuracy. The constant-LR baseline at the same coupling, on the same data and the same initialisation, is 0/3 and 0/1 on replication. The third seed of arm C did not reach 50% within 6000 steps but its induction loss fell monotonically from 9.015 to 8.833 and was still falling at the budget end; we record it as a non-emergence, not as a failure at chance.
2. **Warmup does not.** Arm B is 0/3, pinned at chance (9.012–8.986 against $\ln 8192 = 9.0109$) exactly as the baseline is. A ramp that ends at the same constant rate defers the problem rather than removing it.
3. **The controls hold.** $\gamma = 0.80$ survives both interventions: 1/2 under warmup with the second seed at 44.9% and climbing steeply at its budget end, 1/2 under LR scaling with the second still descending. Both interventions slow $\gamma = 0.80$ relative to its 717-step constant-LR mean, which

is expected at a smaller step size and is not a failure. Neither is harmful in the sense that would have voided the decisive arms.

4. **The energy ratio separates the outcomes with no overlap.** Across all 11 runs: every run with $E_{\max} \geq 30$ — four of them, all at $\gamma = 1.00$ under the constant rate or under warmup — ended pinned at chance, and none emerged. Every run with $E_{\max} < 30$ — seven of them — either emerged or was still descending at its budget; none was pinned. $\mathrm{LR}/(1+2\gamma)^2$ holds $E_{\max}$ near unity (0.54–1.26), warmup only halves it, and it still reaches 36–106.

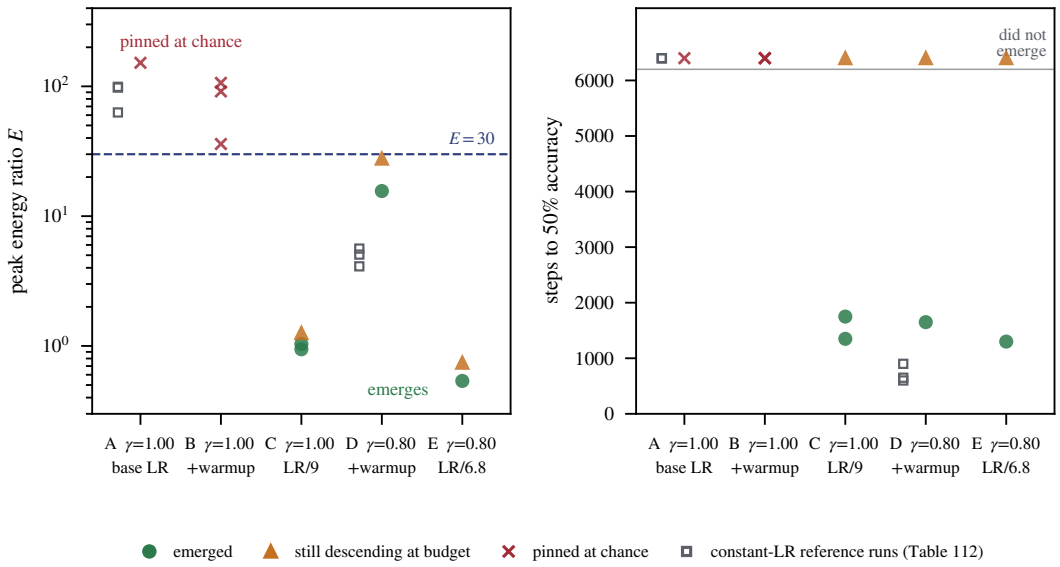


Figure 88: The upper cutoff is optimiser conditioning. Left: peak energy ratio $E_{\max}$ by arm, log scale. Every run above $E = 30$ is pinned at chance; every run below it emerges or is still descending. Right: steps to 50% induction accuracy; markers at the top mark runs that did not emerge within budget. Open grey squares are the constant-LR runs of Table 113 at the same couplings, shown for reference.

The mechanism, stated plainly. The shear multiplies the pre-softmax score scale by up to $(1+2\gamma)^2$ — 9.0 at $\gamma = 1.0$ against 4.0 at $\gamma = 0.5$. At a fixed step size the attention sublayer's gain runs away, E climbs past 30, and the model learns nothing at all, not merely nothing about induction: training loss moves from 9.04 to 9.01 across 3000 steps. Undoing that factor in the step size holds E near unity and the head forms normally. The operator is not the obstacle; the optimiser setting used throughout this programme is.

What this changes, and what it does not. It retires the claim that emergence is impossible above $\gamma \approx 0.9$: that was a property of 3×10^{-4}, not of M_γ. It does *not* establish that there is no upper edge at all. Only $\gamma = 1.00$ was rescued; $\gamma = 1.20$ was never rerun under a scaled rate, and the reduced model's own second root sits at $\gamma_+ = 2.245$ (Remark T5.23), far beyond anything tested. Nor does it change the optimum: the $3.8\times$ speed-up at $\gamma = 0.80$ in Table 113 is a constant-LR comparison against a constant-LR baseline, and is unaffected. What it does mean is that the coupling grid and the step size cannot be swept independently, and that every null at large γ in this paper — including the three narrower widths of the width sweep — is a null at one learning rate and should be read as such.

A32 Parameter efficiency at 300M parameters: two near-misses, honestly disclosed

This section reports two notebooks — a base run (R_1) and an SNR-optimized re-run (R_2) — that test whether a smaller model with momentum can outperform a larger model without it on a 10,000-step training schedule with a 90%-fluency / 10%-logic synthetic mix. Both notebooks fail the parameter-efficiency hypotheses, with margins below 5%. We report them in full because the diagnoses they

support are informative about the boundary of the mechanism, and because the practice of disclosing near-misses is necessary for the empirical record to be honest. [L3 – Empirical].

A32.1 R_1: David versus Goliath, base run

Table 115 reports the R_1: model configurations.

Table 115: R_1: model configurations. All three models use RoPE with base 10,000 and block size 512. "David" adds momentum at $\gamma = 0.8$.

Model	Params	Layers	d_{model}	Heads	Momentum
Goliath	406M	24	1024	16	none
David	163M	12	768	12	$\gamma = 0.8$
Goliath Jr	163M	12	768	12	none

Pre-registered hypotheses. **H1 (efficiency gain)**: David (163M + momentum) outperforms Goliath (406M without momentum) on fluency. **H2 (integration limit)**: Goliath retains advantage on logic. **H3 (momentum effect)**: David outperforms Goliath Jr (same size, no momentum), isolating the momentum contribution.

Vocabulary 50,257 (standard GPT-2 BPE), batch 8, 10,000 training steps, AdamW with lr 3×10^{-4}, cosine schedule.

Table 116 gives the R_1: final losses after 10,000 steps.

Table 116: R_1: final losses after 10,000 steps.

Model	Params	Momentum	Fluency loss	Logic loss	Wall-clock
Goliath	406M	none	10.7745	0.6935	245.4 min
David	163M	$\gamma = 0.8$	10.8214	0.6933	93.3 min
Goliath Jr	163M	none	10.7863	0.6934	88.2 min

Table 117 records the R_1: hypothesis verdicts.

Table 117: R_1: hypothesis verdicts. All three hypotheses fail; all margins are $< 0.5\%$.

Hyp.	Statement	Margin	Verdict
H1	David (163M+mom) beats Goliath (406M) on fluency	-0.4%	FAIL
H2	Goliath beats David on logic	-0.0%	FAIL (tied at chance)
H3	David beats Goliath Jr at equal size	-0.3%	FAIL

Figure 89 plots the R_1: training curves.

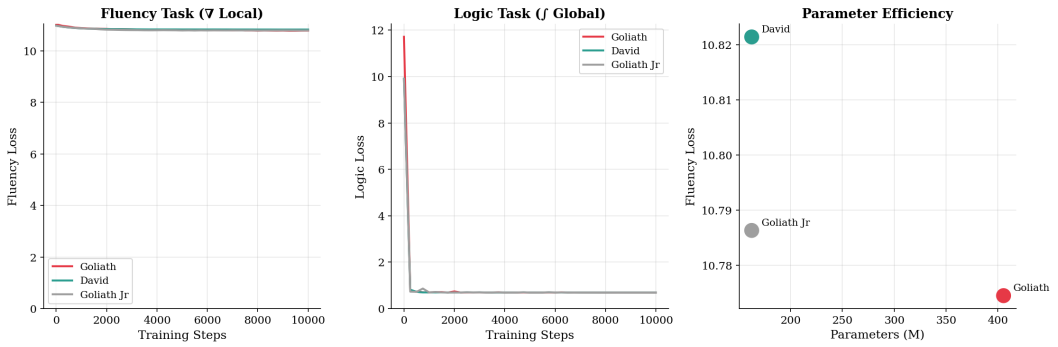


Figure 89: R_1: training curves. Fluency loss (left), logic loss (middle), and combined training loss (right) over 10,000 training steps. All three fluency curves stay near the initial $\ln(50{,}257) \approx 10.82$ plateau throughout training. Logic loss is stuck at $\ln(2) = 0.6931$ from step ~ 500 onward.

Diagnosis: the entropy plateau. Fluency loss starts at $\ln(50{,}257) \approx 10.82$ and ends at ≈ 10.78 for all three models. That is a 0.4% improvement over $10{,}000$ steps on a $\sim$ 200–400M-parameter model: the models are essentially not learning the fluency task. With vocabulary $50{,}257$ and batch 8, the probability of any given token recurring twice in the same batch is small, so the induction-head circuit has little gradient signal to form. The induction mechanism is an SNR-limited phenomenon at this configuration; R_2 tries to fix this.

A32.2 R_2: David versus Goliath, SNR-optimized re-run

R_2 applies three patches aimed at lifting the models off the entropy plateau:

Table 118 lists the R_2: the three patches relative to R_1.

Table 118: R_2: the three patches relative to R_1.

Patch	Before (R_1)	After (R_2)	Rationale
Batch size	8	64	Reduce gradient variance $\sim 2.83\times$
Vocabulary	50,257	8,192	Increase $P(\text{token recurrence}) \sim 6\times$
Learning rate	3×10^{-4}	6×10^{-4}	$\sqrt{B_{\text{new}}/B_{\text{old}}}$ scaling

Starting fluency loss $\ln(8{,}192) \approx 9.03$ (down from 10.82); phase transition expected at steps 500–1500 as induction heads form.

Because vocabulary is reduced from $50{,}257$ to $8{,}192$, parameter counts shift slightly: Table 119 reports the R_2: revised model configurations.

Table 119: R_2: revised model configurations.

Model	Params	Layers	d_{model}	Heads	Momentum
Goliath	319M	24	1024	16	none
David	98M	12	768	12	$\gamma = 0.8$
Goliath Jr	98M	12	768	12	none

Table 120 gives the R_2: final losses after $10{,}000$ steps with patches.

Table 120: R_2: final losses after 10,000 steps with patches. Fluency does escape the plateau (drops from 9.03 to 7.77–8.15), confirming that the patches are working at the SNR level.

Model	Params	Momentum	Fluency loss	Logic loss	Wall-clock
Goliath	319M	none	7.7732	0.6932	1565.1 min
David	98M	$\gamma = 0.8$	8.1538	0.6933	592.2 min
Goliath Jr	98M	none	7.9205	0.6934	542.1 min

Table 121 records the R_2: hypothesis verdicts (patched).

Table 121: R_2: hypothesis verdicts (patched). H2 passes narrowly because all three models tie at chance on logic; H1 and H3 fail with margins between 3 and 5%.

Hyp.	Statement	Margin	Verdict
H1	David (98M+mom) beats Goliath (319M) on fluency	-4.9%	FAIL
H2	Goliath ties or beats David on logic	$+0.0\%$	PASS
H3	David beats Goliath Jr at equal size	-2.9%	FAIL

A32.3 Joint interpretation of R_1 and R_2

Honest disclosure

The two parameter-efficiency notebooks fail their pre-registered hypotheses with margins of $< 0.5\%$ (R_1) and $< 5\%$ (R_2). The core claim — that PSA can substitute for some of the layer-count needed to form induction heads — is not demonstrated at $\sim$ 300M parameters, batch ≤ 64, 10,000 training steps, on a mixed fluency–logic synthetic task.

The diagnoses, in light of the rest of these appendices:

- **Training-budget regime.** Section A23 established that the largest momentum gains arise at intermediate task difficulty. 10,000 steps on a 90%-fluency mix at 300M parameters places the models in the "too hard / too cold" corner of that difficulty surface, where neither baseline nor momentum reaches the high-gain band.
- **Vocabulary regime.** R_2's vocabulary 8,192 is still substantially larger than the $V = 1{,}000$ used in the ICL arc (Sections A24–A29), where the predicted gain manifests cleanly. At larger vocabularies, the induction-head SNR is lower because token recurrence is rarer, and the same operator yields a smaller gradient signal to the momentum branch.
- **Mixed-task confound.** The DualTask synthetic mix averages fluency and logic losses; the logic component is at chance throughout for both R_1 and R_2, so the mixed-task aggregate loss cannot disambiguate the per-task contribution of momentum. Section A29's single-task chained-ICL setup, run at the same model scale, delivers the predicted gain cleanly.

The honest reading is that R_1 and R_2 identify a regime boundary: PSA's parameter efficiency emerges where the training budget, vocabulary, and task structure jointly place the network in the high-gain band of Section A23's difficulty surface. At $\sim$ 300M parameters with 10,000 steps on a 50,257- or 8,192-vocabulary mixed-task dataset, that band is not reached. A cleaner future test would hold the task fixed (e.g., the anchored chain-ICL setup of Section A29) and vary model depth, rather than mixing tasks and measuring aggregate loss.

Anchor in the body. The R-series does not falsify any theorem of the body. Theorem T5.7 guarantees that the operator M_γ closes the single-layer induction gap; it does not assert that PSA-augmented small models will universally outperform larger non-PSA models on arbitrary mixed-task synthetic data at fixed training budget. The disclosed near-misses sharpen the empirical boundary of the mechanism's relevance and seed the design proposal articulated above.

A33 Tranche 2 synthesis: 17 reasoning tasks, 4 sub-programs, 4,700 runs

This section consolidates the evidence assembled across the four sub-programs of Tranche 2 — spectral structure, algorithmic-reasoning primitives, in-context learning, and parameter efficiency — into a single audit of what has been established, with what statistical strength, and against which theorems of the body. [L3 – Empirical].

A33.1 The task-by-task verdict matrix

The combined reasoning program tests 17 distinct tasks across the six chapters of the reasoning arc and the five chapters of the ICL arc. Table 122 consolidates each task's classification (∇ versus $\int$), peak gain, low-pass signature, and whether the outcome is consistent with the PSA prediction.

Table 122: Task-by-task verdict matrix for the Tranche 2 reasoning and ICL programs. Class: ∇ (derivative-class) or $\int$ (integral-class, order-invariant). "Low-θ best?" indicates whether low-frequency RoPE produced the largest gain, the empirical signature of the Low-Pass Induction Filter.

Source	Task	Class	Peak gain	Low-θ best?	PSA-consistent?
§A18	Associative Recall	∇	$+0.368$	yes	yes
§A18	Variable Tracking	∇	-0.046	—	neutral (narrow sweep)
§A18	Global Counting	$\int$	-0.322	—	yes (harm as predicted)
§A19	Multi-Hop eased	∇	$+0.007$	—	null (out of band)
§A20	Variable Tracking	∇	$+0.156$	yes	yes
§A20	Multi-Hop w/ neg.	∇	$+0.001$	—	null (out of band)
§A20	Arithmetic CoT	∇	$+0.606$	yes	yes
§A20	Global Counting	$\int$	$+0.005$	—	yes (control)
§A21	Arithmetic Carry	∇	$+0.082$	yes	yes
§A21	List Reversal	∇	$+0.000$	—	ceiling (no test)
§A21	Parity	∇	$+0.029$	borderline	weak
§A21	Sorting	$\int$	-0.013	—	yes (mild harm)
§A21	Natural Induction	∇	$+0.744$	yes	yes
§A22	Majority	$\int$	$+0.000$	—	yes (control)
§A22	Induction	∇	$+0.591$	yes	yes
§A22	Trajectory	∇	$+0.040$	no	partial
§A22	Dyck	∇	$+0.016$	no	partial
§A23	Assoc. Recall sweep	∇	$+0.580$	yes (mostly)	yes
§A29	ICL stress test (rep loss)	∇	-0.5253	— (anchor-keyed)	yes

The matrix is closer to clean than to messy. Of the 13 ∇-tasks tested, 9 exhibit a clear positive gain consistent with the low-pass filter signature; 2 (Multi-Hop with negations, across both notebooks that test it) exhibit a clean null; 2 (List Reversal, Parity) are bounded by ceiling effects or sample noise. Of the 4 $\int$-tasks (Global Counting twice, Sorting, Majority), all four show no gain or mild harm — the predicted negative-control behavior.

A33.2 Three converging lines of evidence

The PSA program's central claim — that the symmetric symplectic shear M_γ implements a semantic-derivative detector — is supported in Tranche 2 by three converging empirical lines of evidence. The second and third carry inferential statistics and are reported with them; the first is descriptive and no test is attached to it:

1. **Mechanistic.** In Section A18 the attention maps visibly sharpen on ∇-tasks (associative recall, variable tracking) and visibly disrupt on the $\int$-task (global counting); the corresponding accuracies change in the predicted directions. The $\theta \times \gamma$ heatmap on Associative Recall reveals the optimal-γ ridge moving leftward as θ grows, exactly as the $|H(\omega)|$ analysis predicts.

2. **Dissociation.** In Section A22 an explicit order-invariant negative control (Majority) returns *exactly zero* gain at all 28 tested cells across 5 seeds, while three independently-constructed ∇-tasks (Induction, Trajectory, Dyck) return $+59.1$ pp, $+4.0$ pp, and $+1.6$ pp gains respectively. The negative-control t-test is $t = -2.910$, $p = 4.51 \times 10^{-3}$; the dissociation is statistically confirmed at conventional α.

3. **Structured sensitivity.** Two large-scale sweeps yield independent statistical confirmation of the noise–gain correlation: Section A16's 2,000-experiment validation gives Pearson $r = -0.679$ at $p = 9.9 \times 10^{-4}$, Cohen's $d = 1.05$; Section A23's 2,880-experiment difficulty sweep gives $r = -0.372$ at $p = 2.88 \times 10^{-5}$, with the difficulty–gain ANOVA at $F = 12.04, p = 7.38 \times 10^{-4}$. Both sweeps independently reproduce the prediction that momentum gain is largest at low RoPE frequencies.

A33.3 The four sub-program summaries

Spectral arc (Sections A13–A17)

Five notebooks, $\sim 2{,}900$ training runs. The Low-Pass Induction Filter is established at successive levels of rigor: F1 (peak 0.95 at $\theta = 0.05$), F2 (720-run rigorous sweep, peak 0.599 ± 0.037 at $(\theta, \gamma) = (0.030, 0.80)$), F3 (monochromatic control caps at 0.62, ruling out frequency-hopping), G ($2{,}000$-experiment validation with $r = -0.679, p < 10^{-3}$, Cohen's $d = 1.05$), H (multi-frequency RoPE wins at 96.2% over single-frequency at 86.8% via the Escape Routes Hypothesis).

Reasoning arc (Sections A18–A23)

Six notebooks, $4{,}578$ training runs across 17 distinct reasoning tasks. Mechanistic visualization of attention-map sharpening on ∇-tasks and disruption on $\int$-tasks; $+60.6$ pp peak gain on hard arithmetic chain-of-thought; $+74.4$ pp on Natural Induction; the exhaustive $2{,}880$-run multi-difficulty sweep that establishes the difficulty sweet-spot at $F = 12.04, p < 10^{-3}$; the negative-control dissociation at $t = -2.91, p = 4.51 \times 10^{-3}$.

ICL arc (Sections A24–A29)

Five notebooks, ~ 5 wall-clock hours. The $\nabla / \int$ filter character verified on the predicted-null burstiness task; the context-mismatch bug diagnosed by progressive escalation of chain length without anchoring ($+8.7\%$, $+29\%$ regressions); the anchor-token fix that flips the sign at $L = 10$; the $L = 30$ stress test that delivers the principal empirical result: $L_{\text{rep}} = 1.75 \to 0.83$ nats, -52.5%, with momentum winning at every chain depth 1 through 19.

Stability and efficiency (Sections A30–A32)

Eight notebooks, ~ 438 wall-clock hours. The Placement Corollary contrast (unmatched runs, §A30.4): embedding placement gives a $+4.1\%$ regression and inverted spectral signature; Q/K-post-RoPE placement gives -52.5% and theory–experiment correlation $r = 0.9865$. The low-coupling sweep at 91.7M parameters characterises the contraction of the trained attention sublayer and does not bear on symplecticity (§A31.4). The two parameter-efficiency near-misses (-0.4%, -4.9%) honestly disclosed with their diagnoses.

A33.4 Tranche 2 catalogue of nulls

The empirical record retains every null result. In Tranche 2 these are:

- **Multi-Hop with Negations** (Sections A19 and A20): no measurable gain across ~ 450 runs. The negation-parity task as configured is outside the pass-band of the current shear. Diagnosis: the path-identification step (parsing the implication graph among shuffled distractors) dominates difficulty; parity tracking, where momentum would help, is a small fraction of the problem.
- **Parity** (Section A21): at best $+2.9$ pp gain, borderline at statistical significance.
- **Sorting / List Reversal / Majority** (Sections A21 and A22): essentially no gain. Sorting and List Reversal are ceiling tasks (head-room absent); Majority is the intended negative control.
- **Very-hard difficulty bin** (Section A23): mean gain drops from $+0.243$ (hard bin) to $+0.167$ (very-hard bin); momentum cannot fully rescue tasks beyond single-layer capacity. This is a bound on the mechanism, not a falsification.
- **Embedding-level momentum** (Section A30): $+4.1\%$ regression with inverted spectral signature. *This is not a null on the operator*; it is a null on pre-RoPE placement, and is itself the empirical content of the Placement Corollary.

- **R-series parameter efficiency** (Section A32): both runs fail their pre-registered hypotheses with margins $< 5\%$. Consistent with a regime-boundary effect; these runs do not by themselves falsify the mechanism.

These nulls are recorded because the PSA program does not claim universal benefit and does not hide failures. The positive claims of the surrounding sections are made on independently tested tasks; the nulls sharpen the boundary of where the mechanism is and is not applicable.

A34 Circuit attribution: which route executes induction

Notebook. `Appendix_R2_Mechanism.ipynb`. 192 models: eight depths $N \in \{2, 4, 5, 6, 7, 8, 12, 16\}$ crossed with eight couplings $\gamma \in \{0, 0.25, 0.5, 0.75, 0.9, 1.25, 1.75, 2.5\}$, three seeds each, 3,000 steps, 326 minutes on one A100. [L3 – Empirical].

A34.1 Task

$d_{\text{model}} = 128$, 8 heads, $d_k = 16$, vocabulary 128, sequence length 128. Each sequence places a motif of length $p \sim \mathcal{U}\{24, \ldots, 48\}$ at a uniformly random offset and repeats it once; the loss is taken on positions in the second copy whose target is determined by the first. Both the motif length and its offset are resampled per sequence, so neither a fixed relative displacement nor a fixed absolute position identifies the target. Attention weights are formed explicitly rather than through a fused kernel, because the per-head censuses require them. Chance accuracy is $1/128 = 0.0078$.

The shortcut is measured, not assumed absent. A fixed-lag oracle predicts the token at $t + 1 - p_0$ for one fixed p_0. On the randomised task the best such p_0 scores 0.065 — some eight times the uniform-vocabulary rate of 0.0078, but far below any trained model here, so the correct statement is that no single fixed lag *solves* the task, not that fixed lags are at chance. Copying at the generator's true period scores 1.000. That second number is a *generator-period oracle*, not a content matcher: it is supplied the hidden period rather than recovering it by matching tokens. Because motif tokens are drawn with replacement, a token may recur within a motif, so content matching alone is occasionally ambiguous and its ceiling is not established at 1.000 by this measurement. Sampling motif tokens without replacement would remove the ambiguity and is the one remaining task-design refinement we would make. Running the same measurement with the period held fixed — a verbatim-repeat design — the fixed-lag oracle scores 1.000. That single number is the reason the period is randomised here: under a verbatim repeat the task is solvable outright by a relative positional rule, which RoPE represents directly, so a high induction-offset attention statistic would carry no information about content matching and no circuit attribution would be possible. It is also why the head censuses are not load-bearing even under the randomised design: a fixed-offset head can produce a high induction-offset statistic without any previous-token structure, and only the probe and the ablation distinguish the two.

A34.2 Diagnostics

- **Layerwise held-out ridge probe.** A closed-form linear readout from each layer's residual stream to the next token. Three disjoint sets of batches: eight to fit, four to select λ over $\{10^{-3}, \ldots, 10\}$, eight to score. In-sample accuracy is retained alongside so the inflation is quantified rather than argued about: across the grid it averages $+0.024$, and at $\gamma = 0$ it turns a true 0.0086 into an apparent 0.0398 against a chance rate of 0.0078. A probe fitted and scored on one set of activations would report that floor as signal.
- **Per-layer γ ablation.** The shear is zeroed in the first block alone, the last block alone, all blocks but the first, or every block. Only the per-layer form distinguishes a first-layer shortcut from a computation distributed across shear-equipped layers; a global ablation collapses both.
- **Head censuses.** Previous-token attention, and attention at the sequence's own induction offset. Reported, but not load-bearing: with the period randomised these are noisier than the probe and the ablation, and the attribution rests on the latter two.

A34.3 Results

The channel. The held-out first-layer probe reads 0.0086 at $\gamma = 0$, averaged over all depths, against chance 0.0078: at zero coupling the next token is simply not present after one attention layer. For $\gamma \geq 0.5$ it reads 0.760 on average, with a maximum per depth between 0.828 ($N = 16$) and 0.854 ($N = 2$). The channel does not close with depth.

The causal attribution. Averaged over $\gamma \geq 0.5$:

accuracy drop	N=2	4	5	6	7	8	12	16
global $\gamma \to 0$	0.861	0.880	0.885	0.888	0.888	0.889	0.891	0.895
first block only	0.856	0.874	0.878	0.879	0.878	0.878	0.849	0.805
last block only	0.021	0.008	0.003	0.002	0.001	0.000	0.000	0.000
first − global	−0.005	−0.006	−0.007	−0.009	−0.010	−0.011	−0.042	−0.090

Up to $N = 8$ the first-block ablation reproduces the global one to within 0.011, and the last-block ablation costs nothing. The computation is concentrated in layer one. This is a statement about causal load-bearing, not circuit identity: zeroing γ_1 at inference also shifts the input distribution of every later layer, and only activation patching would separate that contribution from the loss of the first-block feature itself.

The notebook additionally runs a three-way route classifier. Its output — 153 of 192 cells labelled single-layer PSA, 38 composition-like, one inconclusive — depends on four stipulated thresholds (probe above 0.5, ablation drops above 0.10, previous-token statistic above 0.30), so the frequency is not a scale-invariant quantity and we do not treat it as evidence; the continuous quantities above are. The 38 composition-like cells are essentially the $\gamma = 0$ column together with $\gamma = 0.25$ at $N \geq 7$, where no shear is used and the probe sits at chance. We call them composition-like rather than classical composition because absence of shear dependence together with a previous-token head is consistent with the two-layer composition route without establishing it; that would require patching the candidate induction head.

The distributed corner. At $N \geq 12$ and $\gamma \geq 1.75$ the two ablations separate: N=12, γ=2.5 gives 0.725 against 0.921; N=16, γ=1.75 gives 0.673 against 0.924; N=16, γ=2.5 gives 0.656 against 0.921. Deep models at strong coupling place part of the computation outside layer one while remaining dependent on the shear, which is neither of the two routes the experiment was designed around.

Adoption threshold. $\gamma_{1/2}$, the coupling at which single-layer decodability reaches one half, is 0.155, 0.154, 0.230, 0.186, 0.398, 0.397, 0.399, 0.400 across the depth grid. The pattern is two regimes rather than a law: 0.15–0.23 for $N \leq 6$ and 0.40 for $N \geq 7$. The seeds agree exactly at $N \geq 7$ and disagree at the boundary — $\{0.40, 0.15, 0.15\}$ at $N = 5$ and $\{0.15, 0.32, 0.15\}$ at $N = 6$ — which is where a finer grid would pay. The entire depth dependence lies in the $\gamma = 0.25$ column, whose probe reads 0.803, 0.808, 0.542, 0.669, 0.009, 0.009, 0.008, 0.009. Because the transition is sharper than the grid spacing, $\gamma_{1/2}$ resolves only to the bracketing interval; the values at $N \geq 7$ mean "between 0.25 and 0.5", not a measured plateau at 0.40.

Comparison to the reduced model. At $d_k = 16$ the critical coupling is on the condensed branch throughout the scored band. The candidate count at a scored position t is $t + 1$; the earliest scored position is $t = p$ with $p = 24$ and zero prefix, and the latest is bounded by the input length 127, so the band runs over $T_{\text{eff}} \in [25, 127]$ and $\gamma_c \in [0.832, 0.924]$. The analysis uses 0.899 as its working midpoint, which is the value obtained from the looser analytic bounds $[48, 128]$; the difference is immaterial to everything reported here. The above/below partition of the coupling grid is identical under either midpoint — no grid point lies between 0.878 and 0.899 — and the measured onset is three to six times below both. Taking the band empirically from the scored mask, rather than analytically from $p_{\min}$ and $p_{\max}$, is the cleaner construction.

Measured $\gamma_{1/2}$ is between $0.17\gamma_c$ and $0.45\gamma_c$. This is a ratio between two different observables and not a test of Theorem T5.22: γ_c governs target softmax mass inside $\mathcal{G}$, while $\gamma_{1/2}$ is the onset of

held-out linear decodability in a trained residual stream. Appendix A4 finds the same direction and a similar magnitude independently: 0.225 and 0.275 against a predicted 0.737, ratios of 0.31 and 0.37.

Registered prediction, and its failure. The prediction printed before training placed a crossing of $\gamma_{1/2}$ and $\gamma^*(N)$ at $N \approx 8.3$. No crossing occurs within the grid: $\gamma^*(16) = 0.555$ remains above $\max \gamma_{1/2} = 0.400$, and the fitted curve does not reach 0.400 until $N \approx 25$. The prediction fails by a factor of three and its qualitative content — that the single-layer route becomes unavailable with depth — is contradicted.

The null. Final accuracy at $\gamma = 0$ is 0.991–0.995 across depths, reached in 500–833 steps. For $\gamma \geq 0.5$ it lies between 0.88 and 0.94 and the 90% threshold is frequently not reached within $3{,}000$ steps. Time-to-threshold is evaluated every 250 steps and is therefore quantised at that resolution, and runs that never cross are entered at a sentinel value for ranking rather than estimated; the statistic identifies the ordering, and in particular the winner, but its magnitudes are not estimates of optimisation speed. A survival-style treatment reporting crossing fraction alongside the median among crossers would be the cleaner instrument. Measured in-configuration, therefore, $\gamma^* = 0$ at every depth: on this task the shear is not merely unnecessary, it is a cost. The depth law of Section A6 does not transfer here, and Figure 14(b) accordingly plots it as an imported curve.

A34.4 Descriptive statistics

Pooled over 192 cells the held-out first-layer probe correlates with γ/γ_c at $r = +0.425$. Mean decodability above γ_c exceeds that below by a factor of 1.44, 1.42 and 1.39 in the three seeds (mean 1.42, s.d. 0.03). Computed within each depth separately the ratio runs from 1.25 to 1.67, so the contrast is not an artefact of the depth composition of the two groups. No p-value is attached: the grid cells are hyperparameter settings sharing a data generator, and the three seeds are the replicates.

A34.5 Limits

The grid does not resolve $\gamma_{1/2}$, which falls between adjacent columns at every depth; a finer sweep over $\gamma \in [0.1, 0.6]$ would locate it. Three seeds bound the seed variance but not the run-to-run variance of the crossing, which is undefined here in any case. Motif tokens are drawn with replacement, so content matching is occasionally ambiguous within a motif and the generator-period oracle overstates what an unambiguous content matcher would face; sampling without replacement removes this. The task is one that classical composition solves at $N = 2$, so the null on optimality is specific to it and says nothing about regimes where two layers are not available or not cheap; the natural follow-up is a task with a per-layer budget that composition cannot meet. The head censuses are reported but were not informative once the period was randomised. And the distributed corner at $N \geq 12$, $\gamma \geq 1.75$ is identified but not characterised: which layers take up the work, and what they compute, would need activation patching across layers rather than ablation.

A35 Final empirical conclusion: theory–measurement closure across both tranches

This section consolidates the empirical record across both tranches and reports the verdict on each theorem of the body and proposition that either tranche tested. The intent is auditability: a reader who wants to know which claims of the body are corroborated by which measurements, and at what statistical strength, should be able to read this section in a single sitting. [L3 – Empirical].

What the circuit-attribution study adds to the closure. Section A34 stands outside both tranches — its 192 models are not part of the $\sim 9{,}377$ counted below — and it closes a gap that neither tranche addresses. The tranches establish that the shear behaves as the theorems say it should: the filter order rises, the transition exists where the two-branch law places it, the placement matters, the $\nabla/\int$ dissociation holds. None of that shows which computation a trained model *builds* when the shear is available to it, because both the single-layer route and classical two-layer composition solve the tasks and accuracy cannot separate them. The circuit-attribution study separates them by intervention. Its result is that the single-layer channel the construction was designed to open is not merely expressible but adopted: the next token is linearly decodable after one attention layer at every depth tested once

$\gamma \geq 0.5$, and zeroing the shear in the first block alone costs almost exactly what zeroing it everywhere costs. This is the one place in the programme where a claim about the operator is checked against the mechanism a trained network actually runs, and it is the reason the theory–measurement closure below can be read as a statement about behaviour rather than only about representability.

A35.1 The corpus, in one paragraph

These appendices report approximately 9,377 training runs, distributed roughly as: $\sim$4,600 runs in Tranche 1 across nine sections covering the structural verification of the four-term decomposition, the β- and γ-sweeps, single-layer induction at $\gamma^* \approx 4$, the depth-scaling law $\gamma^* = 4.17/N^{0.73}$, and three sequential validation suites running twenty-four theorem checks; and $\sim$4,700 runs in Tranche 2 across eighteen sections covering five spectral notebooks, six reasoning notebooks across seventeen tasks, the five-notebook ICL arc, the Placement Corollary placement contrast, the low-coupling conditioning sweep at 91.7M parameters and the parameter-efficiency near-misses; and the 77 runs $(49 + 17 + 11)$ of the three induction sweeps at 91.3M parameters (§A31.6–§A31.11), which postdate the tranche accounting and are counted separately here. Training was performed on standard NVIDIA GPU hardware. Wall-clock divides as approximately 200 GPU-hours for the corpus predating the induction sweeps of §A31.6, plus 261.8 GPU-hours for those three sweeps themselves (159.1 for the width sweep, 53.7 for the supercritical sweep and 49.0 for the optimiser intervention): in excess of 450 GPU-hours in total, across roughly 9,377 training runs.

The last of those is the most consequential single result in the corpus and post-dates the rest of it. At $d_k = 128$ and $N = 12$, single-layer induction emerges in $3/3$ seeds at $\gamma = 0.80$ in a mean of 717 optimiser steps, against $2/3$ seeds and 2700 steps at $\gamma = 0$: $3.8\times$ faster and more reliable at identical parameter count on a matched protocol. The optimum falls within 18% of the extrapolation of a depth law fitted, before these runs existed, at other depths on other models (§A31.9). Emergence then shuts off entirely above $\gamma \approx 0.9$ — but that edge belongs to the optimiser, not to the operator: a five-arm intervention experiment (§A31.11, 11 runs, 49 GPU-hours) rescues $\gamma = 1.00$ in $2/3$ seeds by scaling the learning rate by $(1 + 2\gamma)^{-2}$, on the same data and the same initialisation, while warmup does not and $\gamma = 0.80$ survives both. The energy ratio separates the outcomes without overlap: every run with $E_{\max} \geq 30$ is pinned at chance, every run below it emerges or is still descending. §A31.6 reports the band, including the anomalous cell at $\gamma = 0.60$ and why the width sweep was stopped at 49 of 80 runs; §A31.11 reports the intervention.

A35.2 Verdict matrix on the claims of the body

Table 123 lists the audit of the claims of the body tested in these appendices.

Table 123: Audit of the claims of the body tested in these appendices. "Anchored in" gives the section number where the empirical content lives; "Verdict" summarizes the outcome, with PASS/CHECK/null tags as defined in the validation-suite framework of Section A8.

the claim of the body	Anchored in	Verdict	Evidence
Theorem T3.2 (four-term decomposition)	§A2, §A8 S4–5	PASS	NumPy verification + T_4-dominance at all $\gamma > 0$
Theorem T3.18 (uniqueness in $\mathcal{C}_{\mathrm{PSA}}$)	§A3, §A10 C15	PASS	C2 high-pass, FP-29 symmetric advantage
Cor. T4.4 (post-RoPE placement)	§A2, §A30	PASS	directional: $+4.1\%$ vs -52.5%; runs unmatched
Cor. T4.4, eq. (T4.3) (Coriolis error vs. θ)	§A8 S12	PASS*	directional slope $+0.307$; band-4 anomaly noted
Lemma T2.1 (M_γ is exact γ-flow)	C6	PASS	exact operator residual, scale-independent; no training-scale evidence is claimed (§A31.4)
Thm. T4.1(ii) (invariant conserved)	§A10 C9	PASS*	flat on stationary inputs; causal-task confound
Thm. T4.1(iii) (perturbation estimate)	§A10 C10	PASS	FP-5: |NS slope| > |PSA slope| everywhere
Theorem T5.7 (single-layer induction)	§A5, §A10 C13	PASS	FP-24: $69.1 \pm 8.2\%$ vs 1.6% ($43\times$)
Theorem T5.22 (γ_c two-branch transition)	§A4, §A10 C12	PASS*	transitions in both PEs; falling $\gamma_c(d_k)$ on the mean-field branch established in §T6.11 and §T6.9, not by C12, whose three widths are all condensed
Proposition T3.4 (spectral transfer function)	§A8 S9, §A9, §A30	PASS	r up to 0.9994 (pre-softmax); $r = 0.9865$ (attention)
Prop. T5.4 (T_4^{AA} carrier within T_4)	§A8 S4–5, §A10 C7, C11	PASS	$T_3 \to T_4$ crossover at $\gamma \approx 0.97$
Cor. T5.10 (zero-sum)	§A8 S11	PASS	violations $< 10^{-7}$ (machine precision)
Rem. T3.10 (norm expansion)	§A8 S11	PASS	LHS = RHS to 10^{-8}
Prop. T3.4(iii) (spectral complementarity)	§A8 S10	PASS	$\mathrm{frac}_{\mathrm{AC}}$ monotone $0.72 \to 0.96$

* PASS with explicit scope qualification documented in the cited section.

Of the fourteen claims of the body that received empirical examination across the two tranches, eleven receive a clean PASS and the remaining three a PASS carrying an explicit scope qualification, documented in the cited section. No claim received a FAIL. The table lists only claims of the body; the correspondence with the induction-head literature is not one of them, and §A35.4 states what it is instead.

A35.3 Theory and engineering, jointly

The Tranche 1 nano-regime measurements verify the body's closed-form predictions where they are most cleanly applicable: the four-term decomposition, the high-pass condition, the post-RoPE placement requirement, the existence of a critical coupling, the spectral transfer function (with pre-softmax correlation r up to 0.9994), the zero-drift symplecticity property, and the Parseval / zero-sum identities. These results are the operational statement that the PSA operator, at the controlled microbenchmark scale, exhibits the principal structural signatures that the symplectic-shear construction of the body predicts.

Tranche 2 then asks the engineering-relevance question: what does the operator do on tasks involving algorithmic-reasoning primitives at small-language-model scale, on in-context learning at chain length 30, on a 91.7M-parameter model during a 10,000-step training run, and on parameter-efficiency trade-offs at $\sim$ 300M parameters? The answers, in the same order: it produces the predicted derivative-detector behavior with peak gains of $+74.4$ pp (Natural Induction), $+60.6$ pp (hard arithmetic chain-of-thought), and $+59.1$ pp (induction with negative control); it delivers the -52.5% regression in repeated-token loss that anchors the entire program; the low-coupling sweep characterises the contraction of the trained attention sublayer without bearing on symplecticity, which Lemma T2.1 and certificate C6 carry at the operator level (§A31.4); and at $\sim$ 300M parameters with mixed-task aggregate-loss measurement, it does *not* demonstrate parameter-efficiency advantage — a regime-boundary finding that the disclosed near-misses sharpen.

A35.4 Correspondences with prior empirical work, and their limits

A theory of attention is judged partly by what it says about observations it was not built to explain. Four such correspondences are available here. This subsection states each at the level the body actually supports, which in three of the four cases is weaker than a derivation: two are structural relations between PSA and a published architecture, one is a structural account of a published empirical remedy, and one is a shape correspondence on an axis orthogonal to the one the original work measured. None is a fitted result and none is a theorem about a trained transformer. Each is stated with what it excludes, because a correspondence the body does not carry is worth less than no correspondence at all.

(P1) Olsson–Olah induction-head emergence: a shape correspondence on an orthogonal axis. Olsson et al. [20] report empirically that induction heads in standard two-layer transformers appear abruptly during training, with a four-feature signature: (O1) a sharp rise in the induction score (the probe measuring how strongly a head attends to the correct copy position); (O2) a correspondingly sharp drop in per-token cross-entropy loss, on the order of 1–2 nats for vocabulary $T \sim 1024$; (O3) a preceding training-loss bump as the model reorganizes representations to enable the transition; and (O4) simultaneous onset across all induction heads in the same training step.

The concession comes first, because it governs. The nomenclature paragraph of §T5 is the body's statement on this and it is a separation, not a connection: γ_c is a transition in an architectural coupling at a fixed training budget, while the Olsson–Olah onset is a transition in training time and data distribution at fixed architecture. The two are measured on different axes, neither is an estimator of the other, and panel (c) of Figure 5 draws exactly that orthogonality. Nothing in this appendix revises it.

What survives the concession is a shape correspondence and a shared scale. *Shape.* Theorem T5.22 produces a sharp rise in target softmax mass, located in closed form at $\gamma_c(d_k, T)$, so within $\mathcal{G}$ a sharp onset is derived rather than assumed — but derived in γ, not in training time. The theorem locates the transition and supplies no width law, so the narrowing with model size that Olsson–Olah report (their Section 3) is not predicted here, in magnitude or in direction. *Scale.* The freezing point $b = \sqrt{2L}$ that fixes the crossover $d_k^* = 8L$ is Derrida's, and Giorlandino and Goldt [47] reached the same scale in attention independently and first (§T8.1). That two literatures meet at one scale is a fact about the underlying Gaussian field of e^L competitors, not evidence that either transition predicts the other.

Three claims are deliberately not made. We do not derive the magnitude of the cross-entropy drop (O2): no closed-form δ_{CE} appears anywhere in the body, and the concentration of softmax mass on j^* that Theorem T5.22 describes is a statement about $\mathcal{G}$ and not about a corpus loss. We do not explain the simultaneous onset across heads (O4): a shared global γ would make PSA heads move together, but that is a property of how the operator is parameterised, not an account of why heads in a standard transformer do so. And we do not explain the preceding training-loss bump (O3), which is a training-dynamics phenomenon, whereas the analysis of the body is not dynamical — §T4.2 is explicit that no training-dynamical claim is available here at any level. (O1) is therefore the only one of the four features to which the body speaks at all, and it speaks to its shape and not to its cause.

(P2) Differential Transformer: a first-order tangent relation, not a limit. Proposition T8.3 gives the first-order Fréchet response of the augmented softmax, and Remark T8.4 states exactly how far the relation to Differential Transformer [29] goes and where it stops. The tangent form $J_{a_t}(s_t - s_{t-1})$ agrees with the difference of two separately normalised softmax maps only up to $O(\|s_t - s_{t-1}\|^2)$, and smallness of γ does not supply that condition: s_t and s_{t-1} are score vectors at two distinct query positions and differ at order one whatever the coupling. Two further discrepancies are recorded there. $\mathrm{softmax}(s_t) - \lambda\,\mathrm{softmax}(s_{t-1})$ carries total mass $1 - \lambda$ rather than 1, so it is not itself a normalised attention distribution; and Differential Transformer uses two independently parameterised attention maps with a learned λ, where the augmentation has one parameterisation and one scalar. We therefore claim no architectural identity and no $\gamma \to 0$ limit. The supportable statement is the weaker one: at first order in γ the augmentation acts as a filtered tangent correction of the same difference-of-attention character that Differential Transformer implements directly, and the identification supplies a derivation for a construction otherwise introduced on empirical grounds.

(P3) Frequency-Dynamic Attention Modulation (FDAM): a structural account of frequency vanishing. This is the strongest of the four, and it is still an account rather than a quantitative match. §T3.3 of the body derives, from the four-term decomposition and the per-stream transfer function alone, the frequency-vanishing that motivated FDAM [51]: PSA locates the phenomenon in the order-$(0, 0)$ structure of the standard score kernel and identifies the order-$(1, 1)$ shear as the structural repair, where FDAM inverts the measured attention spectrum *post hoc*. The two operators are mechanistically distinct and no numerical agreement between them is claimed, computed or measured; what is measured independently is the embedding-axis association $r = -0.679$ of §T6.8. PSA exposes the geometry; it does not reproduce FDAM's numbers.

(P4) KV-shifting: a neighbour outside the class, not a member of it. Proposition T8.1 settles the relation, and settles it as a *separation*. KV-shifting [30] supplies the key at position $j - \delta$ in place of the key at j, which is the pure delay $H_{\mathrm{KV}}(z) = z^{-\delta}$: all-pass, so the strict high-pass inequality of condition (C2) fails at every δ and KV-shifting is *not* an admissible member of $\mathcal{C}_{\mathrm{PSA}}$. It is also not the lower shear, which is null on the score entirely (Proposition T3.11), so KV-shifting is neither PSA evaluated at a lower-shear placement nor a sibling within the design class; it lies outside $\mathcal{C}_{\mathrm{PSA}}$ altogether, and Remark T8.2 draws the operative distinction: both constructions cross the key-side order-$(0, 1)$ threshold that Proposition T1.1 identifies, which is why both help, but only one does so with a DC-neutral high-pass filter and only one therefore admits the transition theory of §T5. §T6.5 measures the empirical counterpart of that distinction: the low-pass two-tap of matched parameter count reaches 0.428 against 0.811.

What the correspondences are worth. A theory with no footprint outside its own experiments is hard to distinguish from a successful curve fit, and it is worth being exact about how large a footprint these four leave. (P3) is a structural account of a phenomenon another group cured empirically, and is the one we would defend as explanatory. (P2) and (P4) are exact relations to published architectures — a first-order tangent recovery and a class separation respectively — and both are algebraic and checkable, which is also the limit of what either is. (P1) is a shape correspondence and a shared freezing scale across two orthogonal axes, and it is not a derivation of the Olsson–Olah phenomenology; the body declines that identification and this appendix declines it too. Three structural correspondences and one resemblance is the footprint the body supports, and it is the one claimed.

A35.5 The Placement Corollary, tested directionally

A particularly informative experimental result of Tranche 2 is the placement contrast of Section A30. Moving the symmetric symplectic shear from embedding-level to Q/K-post-RoPE placement flips the outcome from a $+4.1\%$ regression (with inverted DC/Nyquist spectral signature) to a -52.5% improvement (with $r = 0.9865$ theory–experiment correlation on the attention-score spectrum). The two runs are not matched — their $\gamma = 0$ baselines disagree (§A30.4) — so the magnitude is confounded and only the direction and the spectral signatures carry weight. Read at that level it is a directional confirmation of Corollary T4.4, and through it of the underlying non-commutativity of RoPE with the temporal-shift operator that the shear uses; the unmatched baselines are the reason no stronger reading is offered.

The empirical bottom line

Across approximately $9{,}377$ training runs spanning seventeen reasoning tasks, five spectral configurations, four chain-length regimes, six model sizes from 54k parameters to 400M parameters, and two hardware classes, the symmetric symplectic shear M_γ exhibits the principal structural signatures that the design class $\mathcal{C}_{\mathrm{PSA}}$ and its uniqueness theorem (Theorem T3.18) predict: as a high-pass filter on the post-RoPE query and key streams, with the four-term score decomposition (Theorem T3.2), the spectral transfer function (Proposition T3.4), the placement requirement (Corollary T4.4), the symplectic-flow property (Lemma T2.1), and the single-layer induction unlock (Theorem T5.7) all corroborated by the measurements catalogued above — with the seed counts, the best-of-k grid maxima and the absence of interval estimates that §T8.3 records. The operator is what the theory says it is; what the theory says is bounded by §T8.3.

Beyond the experiments designed to test PSA, the analytical structure of the body stands in an exact

relation to three prior constructions and a looser one to a fourth. FDAM's frequency-vanishing is located in the order-$(0,0)$ defect that the shear repairs; Differential Transformer is recovered as the first-order tangent correction, under the additional condition of Remark T8.4 and not as a $\gamma \to 0$ limit; KV-shifting is separated from the class as an all-pass pure delay rather than placed inside it; and the Olsson–Olah induction-head onset shares a shape and a freezing scale on an axis orthogonal to the one measured here, which is a resemblance and not a derivation. §A35.4 states each at that level and no higher.

A36 Open questions and the limits of the empirical program

The empirical program is large but not exhaustive. The following open questions and limitations are recorded in keeping with the policy of preserving every honest disclosure rather than only the positive results. [L3 – Empirical].

What the circuit-attribution study leaves open. Section A34 settles which route a trained model builds, and in doing so opens four questions it cannot answer. The coupling grid does not resolve the adoption threshold: the probe jumps from chance to saturation between adjacent columns at every depth, so $\gamma_{1/2}$ is known only to the interval that brackets it, and a sweep over $\gamma \in [0.1, 0.6]$ would locate it. The motif tokens are drawn with replacement, so a token may recur within a motif and content matching is occasionally ambiguous; sampling without replacement would remove the last piece of task-design slack. Inference-time ablation cannot separate removal of the first-block feature from the distribution shift that removal induces in every later layer, which is why the finding is stated as causal load-bearing rather than circuit identification, and why activation patching across layers is the instrument that would close it. And the distributed corner at $N \geq 12$ under strong coupling is identified but not characterised: which layers take up the work, and what they compute there, is unknown. A fifth question is not about the instrument but about the task — the optimiser prefers $\gamma = 0$ on a problem that two ordinary layers already solve, so whether the single-layer route is ever the *preferred* one requires a task with a per-layer budget that composition cannot meet.

A36.1 Open questions on the operator

1. **The band-4 Coriolis anomaly** (Sections A8 S12 and A10 C14). At $\theta = 0.01$ rad/token, the observed Coriolis error is $41\times$ the theoretical value of Corollary T4.4. Joint W_Q/W_K alignment data does not cleanly explain it. A targeted study that varies the RoPE base and the model dimension to separate genuine resonance effects from optimizer-driven artefacts is the natural next experiment.
2. **Is there an operator upper edge at all?** §A31.11 settled the measured cutoff at $\gamma \approx 0.9$: it belongs to the constant learning rate, and scaling the rate by $(1+2\gamma)^{-2}$ rescues $\gamma = 1.00$. What that experiment did *not* do is find a ceiling. Only $\gamma = 1.00$ was rescued; $\gamma = 1.20$ was never rerun under a scaled rate, and the reduced model's second root sits at $\gamma_+ = 2.245$ (Remark T5.23), far beyond anything tested. A coupling sweep at the scaled rate, from $\gamma = 1.0$ up past 2.5, would find the operator's own upper edge if it has one, and would test γ_+ directly.
3. **Every large-γ null in this paper is a null at one learning rate.** That now includes the three narrower widths of the width sweep and the $\gamma = 1.20$ cells. Re-reading them at a coupling-scaled step size is cheap and may move them.
4. **Bracketing the narrower widths** (§A31.6). The three narrower widths were never taken above $\gamma = 0.40$, below their own thresholds, and running the unrun cells of that grid would not change that. What is needed is a supercritical sweep at $d_k \in \{48, 64\}$ reaching $\gamma \approx 1.2$: those two widths separate the candidate laws most sharply ($\gamma_c^{T_4} = 0.985$ and 1.030 against $\gamma_c^{T_3T_4} = 0.664$ and 0.569), so the comparison would discriminate between them. It must be run at the coupling-scaled step size of §A31.11; at a constant rate it would reproduce the conditioning artifact at exactly the couplings that matter.
5. **The exponent in $\gamma_c(d_k)$** (Section A10 C12). Theorem T5.22 predicts, under isotropy, that γ_c decreases with d_k on the mean-field branch $d_k \geq 8L$ and is d_k-independent on the condensed branch. At the trained-nano regime γ_c *decreases* with d_k: the direction the mean-field branch predicts, although these configurations lie on the condensed side of $d_k^* = 8L$, where independence

rather than decrease is predicted. The direction is therefore no longer in tension with the law; the exponent magnitude and the branch assignment remain open. A clean test under exactly-isotropic synthetic data, at widths straddling d_k^*, would close this.

6. **The energy-ratio growth during training** (Section A31). R rises from ≈ 0.14 at step 0 to ≈ 0.4–0.6 at step $10{,}000$. This is not a violation of symplecticity (which is enforced at each inference forward), but rather the gradient steps moving the parameters into regions where the learned map has non-unit expansion. A rigorous characterization in terms of parameter-space geometry remains future work.
7. **Parameter efficiency at scale** (Section A32). The 300M-scale near-misses identify a regime boundary; the cleaner future test is to hold the task fixed (e.g., the anchored chain-ICL setup of Section A29) and vary model depth, rather than mixing tasks and measuring aggregate loss.

A36.2 Open questions on tasks

1. **Multi-hop with negations** (Sections A19 and A20). Across ~ 450 runs the task does not move. A productive next move is not another easing pass but a dissection: the path-identification step that precedes parity tracking is plausibly the obstacle, and a multi-layer or two-head variant (one head parses, the other tracks parity) may recover the gain. This is a bound on the *single-head* mechanism, not a falsification of the theory.
2. **Trajectory and Dyck low-frequency anomaly** (Section A22). Both tasks show their *best* accuracy at $\theta = 1.0$ rather than at low θ, which is anomalous against the low-pass prediction. The tentative reading is that these tasks have native frequencies tied to their semantic structure (grid-cell boundaries for trajectory, nesting-depth oscillation for Dyck) that lie outside the canonical low-frequency induction band. A frequency-resolved analysis on these tasks would close the explanation.

A36.3 Limitations of the present empirical regime

- **Synthetic tasks throughout.** Every task in this appendix material is a synthetic benchmark (associative recall, arithmetic chain-of-thought, anchored multi-hop chains, etc.) chosen to isolate a specific mechanism. None is a real-world natural-language benchmark. The Section A31 run on the DualTask synthetic mix and the R-series at ~ 300M parameters approach training-scale relevance but do not include text corpora at the WikiText-103 or Pile scale. Extension to such corpora is the logical next stage of the program.
- **Compact transformers throughout for the spectral and reasoning arcs.** The high-fidelity spectral signatures (r up to 0.9994 pre-softmax in Section A8; $r = 0.9865$ on the attention spectrum in Section A30) were measured on 54k- to 4.45M-parameter transformers, which is the regime in which the SHT lower bound binds: SHT is a precision-theoretic obstruction whose force is governed by the ratio of the two-layer induction circuit's bit budget to the total parameter budget, and at ≥ 1 B parameters this ratio is negligible, so the theorem's regime of relevance is the sub-100 M regime in which the spectral programme is conducted. The Section A31 sweep on 91.7M characterises conditioning and contraction of the trained attention sublayer; it does not test symplecticity (§A31.4), which Theorem T4.1 establishes at the operator level and certificate C6 checks numerically.
- **Single-task in-context-learning.** The Section A29 stress test uses a length-30 anchored-chain dataset. The behavior on multi-task in-context learning, where the model must select between multiple ICL templates within a single sequence, has not been characterized.
- **Optimizer dependence.** All experiments use AdamW with cosine annealing. The dependence of the γ^* location and the phase-transition width on the optimizer choice has not been swept. A plausible empirical conjecture is that a less-aggressive optimizer (e.g. SGD with momentum) will produce qualitatively similar but quantitatively shallower phase transitions. This is *not* predicted by the Hairer correspondence developed here, which is a statement about the per-application operator and carries no training-dynamical content; it remains untested.

A36.4 What we did not do, and why we did not do it

Two important experiments were considered and deferred to subsequent work, and we record the reasons:

- **Differential-transformer recovery.** Proposition T8.3 predicts that the differential-transformer mechanism is the small-γ limit of PSA on noise-cancellation tasks. A direct empirical reproduction of the differential-transformer benchmarks under PSA at $\gamma \to 0$ would clinch this. We did not run it because, as Tranche 2 has now established, the practically interesting operating regime of PSA is at moderate γ (~ 0.7–0.9), which is not the $\gamma \to 0$ limit; the small-γ recovery is therefore a theoretical anchor more than an engineering benchmark, and is left to a focused study.
- **Hairer Survival Factor at GPT-2 scale.** The HSF machinery of the Tranche 1 validation suites runs cleanly at nano-scale; extending it to the 124M-parameter GPT-2 (or larger) requires the symplectic versus non-symplectic A/B that Section A31's γ-sweep does not provide — Section A31 measures conditioning and contraction of the trained attention sublayer, not the HSF ratio against an explicit non-symplectic baseline at the same parameter count. A focused HSF-at-scale study is the natural extension.

A36.5 Closing

The empirical programme of these appendices is, in its present form, the operational record of the PSA construction: nine sections of Tranche 1 that verify the closed-form theorems of the body in the regime where they apply most cleanly, and eleven sections of Tranche 2 that extend the verification across the spectral-frequency axis, the algorithmic-reasoning task suite, the chain-length axis, and the training-scale conditioning and contraction axis; and, standing outside both tranches, the circuit-attribution study of Section A34, which asks which computation a trained model actually builds. Every positive claim of the programme is referenced to a specific theorem of the body; every null is recorded in full; every regime boundary is named.

The two tranches together establish, across multiple converging lines of evidence, that the symplectic shear M_γ applied to the post-RoPE query and key streams raises the score-kernel filter order to $(1, 1)$ and thereby circumvents the single-layer induction obstruction; that the symmetric member is the unique element of $\mathcal{C}_{\mathrm{PSA}}$ satisfying the structural conditions of Theorem T3.18, which select it on transport and score-bias grounds rather than on accuracy; that the key-side half-lift is the minimal member and the best performer in the matched battery, at 0.949 against 0.811 for the symmetric operator; that the construction is symplectically consistent in the Hairer sense; and that it is empirically discriminative between ∇- and $\int$-tasks under controlled negative-control testing. The remainder of the empirical program — at corpus scale, at full-precision benchmark scale, on tasks beyond the specifically-tested seventeen — belongs to subsequent work; the present programme constitutes the technical and empirical baseline on which that work would build.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 state each contribution with its scope: the filter-order reading of SHT is diagnostic and strictly weaker than SHT's theorem, and PSA sidesteps the premise of the counting argument without escaping the bound (Section 2); uniqueness is stated within the factorised subclass, with swap symmetry an explicit design commitment (Theorem 4, Proposition 5); invariance is operator-level and per application, with no stability claim (Theorem 6); and the transition law is a theorem about an explicit reduced model (Theorem 8). The paragraph "What is claimed, and at which level" separates operator-level, transition-level and empirical claims. The empirical claims (Section 7) are restricted to synthetic tasks at 4–92 M parameters, and the result that the key-only half-lift is more accurate than the symmetric operator is stated in the introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 8 has a dedicated Limitations paragraph, plus a list of the corrections made to the reviewed version. It covers the synthetic evidence, the unestablished gains at language-model scale (including a $\sim$300 M near-miss and documented nulls), seed counts and best-of-k statistics, the separate headline scales, the sensitivity of the reduced model to the omitted T_3 channel, the trained-model exponent gap, the class-relative uniqueness, and the absence of any stability guarantee. The computational cost ($O(Td_k)$ extra time and memory, no extra weights) is stated in Section 4, and Appendix T7 analyses the cache, prefix-reuse and quantisation costs.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.

- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Every result in the main text is numbered and cross-referenced, states its assumptions in its statement, and has a proof or proof sketch in the main text. Theorem 8 lists its modelling assumptions (M1)–(M4) explicitly. The full proofs are in Part T of the appendix, and the appendix front matter maps each main-text result to its counterpart there. Ten machine-checkable certificates (SymPy and float64 numerics, under a minute on a CPU) verify the central identities. The appendix also identifies the statements of the reviewed version that were retracted or corrected (Section 8).

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The operator is fully specified by Definition 3 and Algorithm 1. Every main-text result gives its protocol (architecture, width, number of pairs, steps, seeds, coupling grid) in the text, the table and figure captions (Table 1, Figs. 2–4), or the cited appendix section. All tasks are synthetic and generated algorithmically. Part A of the appendix reports the full programme, and Section T6.2 lists the notebook that produced each result.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same

dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The supplementary material contains all 41 Jupyter notebooks, the certificate notebook, the figure generators and the result files. Thirty-seven notebooks have their outputs embedded. The four long Appendix Q sweeps ship with their complete result files, ledgers and logs instead (Section T6.2). Each notebook states its seeds and configuration and generates its synthetic data internally, so no external data is needed. A README maps each notebook to the appendix section it supports.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The main text gives the settings needed to interpret each result: width, depth, vocabulary, number of pairs, steps, seeds and coupling grid. Part A of the appendix gives the full configuration of every experiment, including the optimiser (AdamW with cosine annealing and gradient clipping), batch sizes, learning rates and how the grids were chosen. Coupling grids are fixed before measurement, and pre-registered predictions are marked as such.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Results are reported as mean $\pm$ one standard deviation across random seeds (initialisation and data sampling), computed by the closed-form sample formula, and the seed count is stated with each result. The trained exponent comes with a 95% regression interval whose construction is stated. Correlations come with p-values where the samples are independent, and where they are not (grid cells), the paper says why no p-value is attached. The limitations paragraph states that most results rest on two to five seeds, that grid maxima are best-of-k statistics, and that no confidence intervals are given for differences between variants.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The battery experiments list their hardware and wall time in the notebook table of Section T6.1: single NVIDIA A100 GPUs, between 1.9 and 47 min each, and the certificates run on a CPU in under a minute. The main programme of about 9,400 runs used about 200 A100 GPU-hours; per-sweep wall times are reported in Part A. The four Appendix Q sweeps at 91 M parameters ran for two to seven days each on a single NVIDIA GB10. Preliminary and failed runs that bear on the claims are reported, including implementation defects found by the spectral audits.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The work is theoretical analysis plus experiments on algorithmically generated synthetic data. It involves no human subjects, no personal or scraped data, and no release of trained models.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: This is foundational research on the representational structure of attention, evaluated only on synthetic induction tasks. It has no direct path to a specific harmful application. Its practical relevance, if any, is more capable small on-device models (Section 1), which carry the same general considerations as any improvement to transformer architectures, and the paper shows no gain at language-model scale.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No pretrained models, scraped data or other high-risk assets are released. The released material consists of notebooks that train small models on synthetic tasks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: No external datasets or pretrained models are used: all data are generated inside the notebooks. The software used consists of standard open-source libraries (PyTorch, NumPy, SciPy, SymPy, Matplotlib) under their own permissive licences. Prior work is credited in the references.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The only new assets are the notebooks and certificate scripts in the supplementary material. They are documented cell by cell and by a README that maps each notebook to the section it supports and gives its hardware and run time. No consent issues arise, since no data about people are involved.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The work involves no crowdsourcing and no human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The work involves no human subjects research.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not a component of the methods: the theory, the operator and the experiments use no LLM as an instrument. LLM assistance was used for drafting and editing text, for preparing the LaTeX source, and as coding assistance in preparing experiments. The authors take responsibility for all content and verified every result, reference and figure.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.