A CONTINUOUS SPACE OF ATTENTION PROJECTION GEOMETRIES AT INITIALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Standard Transformer initializers set how large the query and key weights are, but not how those two matrices relate to each other. Attention scores, however, depend on their product. We treat that missing relation as a continuous family: a shared component, a residual orthogonal to the query row space, and an independent residual, all with the same expected weight energy. Closed-form finite-width formulas give the mean and variance of every attention score under this family. Monte Carlo checks match the formulas, and a sweep of the family at a fixed width and context produces several distinct initial attention patterns which are nearly uniform, sharply peaked, self-focused, and a mid-entropy band in between. The same coefficient values do not keep those patterns when width and context change. A one-block training comparison that first looks like a geometry effect is explained by scale: mid-entropy and diffuse points sit at small scale and train similarly, while concentrated and self-focused points were copied in at much larger scale and train worse. Joint query–key geometry is a real finite-width degree of freedom. The map we report describes one setting; it is not yet a portable initialization recipe.

1 INTRODUCTION

Modern-day transformers (Vaswani et al., 2017) compute the attention score between two tokens x_i and x_j as

$$\ell_{ij} = \frac{(W_Q x_i)^\top (W_K x_j)}{\sqrt{m}}, \tag{1}$$

where $W_Q, W_K \in \mathbb{R}^{m \times d}$ are the projection matrices used to obtain the query and key representations, respectively.

These projection matrices are typically initialized using Gaussian distributions, as in GPT and other transformer-based models (Radford et al., 2018; Brown et al., 2020). The two matrices are initialized independently, with their variances scaled according to the dimensions of the matrices (Glorot & Bengio, 2010; He et al., 2015). Thus, at initialization, W_Q and W_K are independent random matrices. However, in the attention mechanism, they do not appear independently; rather, the attention score depends on their product $W_Q^\top W_K$.

The relationship between the two matrices can result in very different attention at step 0. If $W_Q = W_K$, similar tokens receive higher attention scores, whereas $W_Q = -W_K$ would result in similar tokens having much lower scores. If $W_Q^\top W_K \approx 0$, then the attention will be near uniform for all the tokens. We study the relationship between W_Q and W_K by conditioning the initialization of W_K on the matrix W_Q. This gives us a continuous space of possible initializations for W_K, which we explore in the paper.

In this paper, we study the space of initializations by first deriving the mean and variance as functions of the parameters of the initialization space. Then we sweep across the space at a particular width and context length, and categorize the resulting attention patterns. Then we attempt to answer two questions: are the results transferable to different widths and context lengths, and can these patterns predict early training.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2 Related work

Initialization research has mostly controlled the scale of individual tensors. Glorot and He schemes match activation and gradient variance layer by layer (Glorot & Bengio, 2010; He et al., 2015). Orthogonal and dynamical-isometry methods go further for a single map, placing linear networks at a critical point with depth-independent signal propagation (Saxe et al., 2014; Hu et al., 2020). In Transformers, the dominant practical controls are still scale and normalization: pre-LayerNorm makes initial attention close to uniform (Xiong et al., 2020; Ba et al., 2016), residual reparameterizations stabilize deep stacks (Zhang et al., 2019; Huang et al., 2020), and query–key normalization or spectral constraints tame exploding attention scores and entropy collapse (Henry et al., 2020; Zhai et al., 2023). Giorlandino and Goldt (Giorlandino & Goldt, 2025) organize these effects into a signal-propagation diagram over attention scale and residual strength. None of this work, however, specifies the joint geometry of W_Q and W_K once their marginal scale is fixed.

A parallel theory literature studies random attention at initialization, usually under independent queries and keys. It has identified rank collapse with depth (Dong et al., 2021), spectral gaps as context grows (Nait Saada et al., 2025), Gaussian equivalence for the attention spectrum (Hayase et al., 2026), and infinite-width or infinite-depth limits (Noci et al., 2023; Bordelon et al., 2024). These analyses characterize what independent random projections do, but they do not treat the amount of query–key sharing as a parameter, so they cannot predict the diagonal bias of a tied initialization or the finite-width split between an orthogonal residual and an independent one.

A smaller set of papers does treat the product $W_Q^\top W_K$ directly, though each targets a single point in the space rather than the space itself. Mimetic initialization copies structures seen in trained models into that product (Trockman & Kolter, 2023). Bao et al. connect attention localization to concentration of the query–key eigenspectrum (Bao et al., 2024). Conditioned initialization targets the spectral conditioning of the attention Jacobian (Saratchandran & Lucey, 2026). Projection sharing makes $W_Q = W_K$ an architectural constraint rather than a point in a family (Kowsher et al., 2025; Kayyam et al., 2026). Our construction differs from all of these: we do not offer a single recommended product, and we do not tie the two matrices as an architecture choice. Instead, we turn the product into a continuous, energy-matched family, derive what it does to attention scores at finite width, and then measure the resulting regimes.

Empirically, trained Transformers develop specialized heads, including induction heads on in-context copying (Voita et al., 2019; Olsson et al., 2022). That literature describes where attention ends up; we ask a narrower prior question: where can attention start, if query–key geometry is allowed to vary? The small training probe later in the paper uses TinyStories and associative recall for that reason, since both are standard cheap tests of early language modeling and retrieval (Eldan & Li, 2023; Arora et al., 2024). A one-block decoder is a limited model, and one-layer Transformers are known to fail some retrieval tasks that deeper models solve (Sanford et al., 2024); we treat the probe as a first check, not as a substitute for a full language-model study.

3 The Query–Key Initialization Family

All derivations for this section are collected in Appendix A.

3.1 A Continuous Family of Query–Key Initializations

We fix a head with model width d and head width m, where $d > m$. We draw two independent Gaussian matrices $A, C \in \mathbb{R}^{m\times d}$ with i.i.d. entries of variance $1/d$.

We construct the third matrix, B, by projecting a fresh Gaussian matrix onto the orthogonal complement of A's row space. Let $G \in \mathbb{R}^{m\times d}$ have i.i.d. entries $\sim \mathcal{N}(0, (d-m)^{-1})$, and let U be the orthonormal basis for the row space of A obtained from the QR decomposition $A^\top = UR$. We set

$$B = G(I - UU^\top). \tag{2}$$

This construction ensures $AB^\top = 0$ exactly, while B retains the same expected energy as an unrestricted Gaussian of the same shape.

For any $s > 0$ and coefficients α, β, γ such that $\alpha^2 + \beta^2 + \gamma^2 = 1$, we define

$$W_Q = sA, \qquad W_K = s(\alpha A + \beta B + \gamma C). \tag{3}$$

The coefficient α is the amount of query–key tying. The remaining residual mass $1 - \alpha^2$ is split by an angle $\phi \in [0, \pi/2]$ as

$$\beta = \sqrt{1-\alpha^2}\cos\phi, \qquad \gamma = \sqrt{1-\alpha^2}\sin\phi. \tag{4}$$

Flipping the sign of β or γ does not change seed-averaged initialization statistics, so we restrict to $\phi \in [0, \pi/2]$ without loss of generality. The sign of α, however, does matter: it decides whether the model assigns larger scores to similar tokens or to dissimilar ones.

Because B and C are centered, every admissible (α, β, γ) gives the same expected weight energy,

$$\mathbb{E}\|W_Q\|_F^2 = \mathbb{E}\|W_K\|_F^2 = s^2 m, \qquad \mathbb{E}[(W_K)_{ab}^2] = s^2/d, \tag{5}$$

so changing the geometry does not change the scale as a side effect (Lemma 1, proved in Appendix A).

This construction contains four named endpoints as special cases:

- independent Gaussian: $(\alpha, \beta, \gamma) = (0, 0, 1)$, the usual initializer;
- row-space orthogonal: $(\alpha, \beta, \gamma) = (0, 1, 0)$;
- tied: $(\alpha, \beta, \gamma) = (1, 0, 0)$, i.e. $W_K = W_Q$;
- anti-tied: $(\alpha, \beta, \gamma) = (-1, 0, 0)$, i.e. $W_K = -W_Q$.

3.2 What the Family Does to Attention Scores

Write $v_i = \|x_i\|^2/d$ and $c_{ij} = x_i^\top x_j/d$ for the normalized token energy and inner product. For fixed inputs, the attention score of a pair (i, j) has exact conditional moments.

Proposition 1 (Conditional score moments)**.** *Under the family above,*

$$\mathbb{E}[\ell_{ij} \mid X] = \alpha s^2 \sqrt{m}\, c_{ij}, \tag{6}$$

$$\mathrm{Var}(\ell_{ij} \mid X) = s^4 \left[\alpha^2(v_i v_j + c_{ij}^2) + \beta^2 g_{ij} + \gamma^2 v_i v_j\right], \tag{7}$$

where

$$g_{ij} = \frac{d\left[(d+1)v_i v_j - 2c_{ij}^2\right]}{(d-1)(d+2)}. \tag{8}$$

The proof is in Appendix A.

Only the shared channel contributes to the mean, since the orthogonal and independent residuals both average to zero. So a positive α pushes up the score for similar tokens – including the diagonal case $i = j$ – and a negative α pulls it down; this is what drives the self-locked and diffuse-toward-dissimilar regimes we see later.

The variance splits into three uncorrelated pieces, one per channel. A shared Gaussian row contributes $v_i v_j + c_{ij}^2$ by Isserlis' theorem. An independent row contributes only $v_i v_j$. The orthogonal row contributes g_{ij} instead, which comes from the second moment of a rank-m projector and only reduces to $v_i v_j$ as $d \to \infty$. At finite width, then, β and γ are not just two labels for the same thing, when we look at what changes if you swap one residual type for the other. The $1/\sqrt{m}$ attention scaling is what keeps the variance from picking up an extra factor of m, which is why $s = 1$ is the natural reference scale.

4 Experimental Setup

Here is what we actually ran, in four stages, each building on the one before it. Appendix B has the exact grids, thresholds, and internal stage names (Stage 1A–1D) for reproduction.

Stage 1A: a quality-control gate on the formulas. Before trusting any map built from Proposition 1, we check that the code actually computes what the formulas say. We compare the closed-form mean and variance against Monte Carlo draws, sweeping widths $d \in \{32, 64, 128\}$, head ratios $m/d \in \{1/8, 1/4, 1/2\}$, four token-pair correlations $c_{ij} \in \{-0.5, 0, 0.5, 1.0\}$, and 24 coefficient settings – 864 configurations, each drawn until the Monte Carlo standard error falls under 1% of the

predicted scale (a floor of 20,000 draws, a cap of 200,000). Alongside the mean and variance we verify the orthogonality identity $AB^\top = 0$, the component energies, the entrywise variance targets, and agreement between analytic and finite-difference gradients, with a Holm–Bonferroni correction at $\alpha = 0.01$ governing the 1,728 resulting hypothesis tests.

All 864 configurations pass every check: mean calibration slope 0.9999 (RMSE 0.0058), variance calibration slope 0.9976 (RMSE 0.0106), and the independent-Gaussian endpoint matches an i.i.d. Gaussian baseline by Kolmogorov–Smirnov test ($p = 0.574$, variance ratio 1.009; Figure 1). With this gate passed, later patterns in the atlas trace back to the family itself rather than to a coding error in the moment formulas.

Stage 1B: mapping the attention regimes. At a reference setting $(d, m, n) = (64, 16, 64)$ we sweep $\log_2 s \in [-2, 2]$ (17 values), $\alpha \in [-1, 1]$ (13 values), and $\phi \in [0, \pi/2]$ (9 values), giving 1,717 unique cells once the redundancy at $|\alpha| = 1$ is removed. Each cell runs with 64 paired seeds across eight input and masking strata – synthetic tokens at three pairwise correlations, post-LayerNorm activations from an independently initialized decoder, and both causal and unmasked attention – for 879,104 raw records and 13,736 stratum-level summaries.

We label each summary from 95% seed-cluster bootstrap intervals on entropy, peak mass, self-mass, and effective rank (exact cutoffs in Appendix B): *diffuse* rows are close to uniform over the allowed keys; *concentrated* rows put most mass on one or a few keys; *self-locked* rows are dominated by the query's own position; *intermediate* ("screened candidate" in the code) rows sit at mid-entropy once the three patterns above are ruled out; and *boundary* cells straddle a threshold and stay unlabeled. Ties are broken in that same order – self-locked first, since a token attending mainly to itself is worth naming ahead of ordinary peakedness – which matters for reading Table 1 and Figure 2.

Pooling all 13,736 summaries: 39.0% diffuse, 27.6% concentrated, 20.9% intermediate, 9.0% boundary, 3.4% self-locked, and one stratum gradient-starved. Real post-LayerNorm activations preserve these proportions with a modest shift toward concentration relative to synthetic Gaussian inputs. Shifting every labeling threshold by 0.05 changes the label on 9.4% of strata in one direction and 48.9% in the other, so we read the atlas as a map with wide, gradually shifting borders rather than sharply bounded countries.

Stage 1C: testing whether the map travels. To check whether Stage 1B's names hold at other widths and context lengths, we fit a classifier on the reference labels alone (20% held out; balanced accuracy 0.987, expected calibration error 0.015, confirming the map is smooth enough to interpolate) and use it to select 120 cells near a decision boundary or a theory/label disagreement. We remeasure those 120 points at both $(64, 16, 64)$ and $(128, 16, 128)$, with the same 64 seeds and eight strata as Stage 1B.

The preregistered bar for a confirmed cell – staying intermediate across all sixteen replicated conditions, at both settings – selects zero of the 120 points; only 30 keep the same majority label at all once width and context length change. Our preregistered fallback for this outcome keeps points labeled intermediate in a majority of the eight reference-setting strata alone, which selects 32 eligible points, eight of which (spread evenly over α and ϕ) carry the label into Stage 1D as a description of the reference setting specifically.

Stage 1D: does the map predict early training? Before any training loss is observed, we freeze 84 coefficient points into six groups: 15 diffuse, 15 intermediate, 15 concentrated, 15 self-locked, 20 boundary, and the four named endpoints at a matched scale. The diffuse, intermediate, and boundary groups come from the Stage 1C fallback above (median $s \approx 0.78$–0.80); since Stage 1C's sampler spent its budget on uncertain cells and never proposed a concentrated or self-locked point, those two groups are filled instead from the Stage 1B discovery sweep, at their own median scales of $s = 2.00$ and $s = 2.83$ – a difference in origin, and in scale, that we return to when interpreting the results.

We train a one-block decoder (width 128, four heads, context 128) for 2,000 AdamW steps at learning rate 3×10^{-4} with gradient clipping at 1.0, on associative recall and next-byte prediction on a fixed TinyStories subset. Each of the 84 points trains under eight paired seeds sharing every random tensor except the query and key weights, for $84 \times 2 \times 8 = 1{,}344$ jobs total, all completing without divergence. The primary statistic is the paired difference in loss decrease per token between the intermediate group and a pooled concentrated-plus-self-locked group, one per seed, combined via

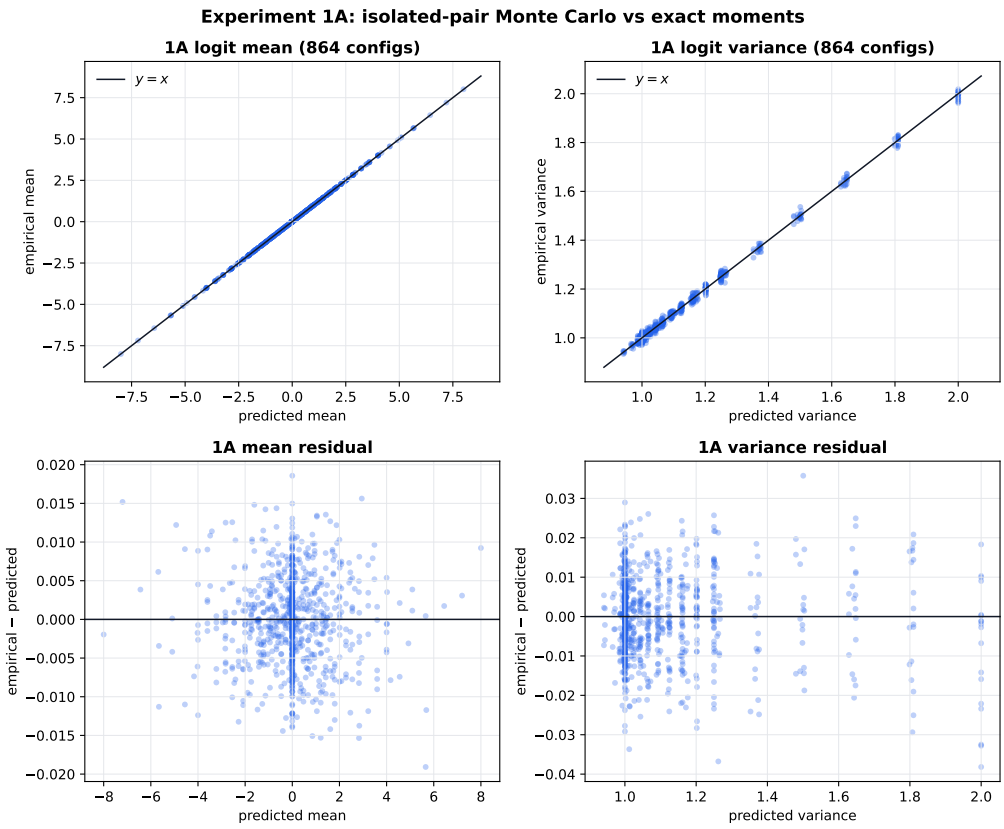


Figure 1: Closed-form score mean and variance versus Monte Carlo, over 864 width, head-size, token-pair, and coefficient settings. The test is for a single token pair under repeated weight draws.

a cluster bootstrap over the eight seeds (10,000 resamples); a task passes when the resulting 95% interval sits entirely above zero.

5 Results

5.1 The moment formulas hold, and several attention regimes emerge

All 864 Monte Carlo configurations pass. After Holm correction across the 1,728 mean and variance tests, nothing is rejected (Holm, 1979). Mean calibration has slope 0.9999 and RMSE 0.0058; variance calibration has slope 0.9976 and RMSE 0.0106. The independent endpoint is indistinguishable from a standard i.i.d. Gaussian draw (Kolmogorov–Smirnov $p = 0.574$, variance ratio 1.009). Figure 1 shows the comparison.

With the pairwise theory confirmed, we next examine the behavior across full sequences. At the reference geometry $(d, m, n) = (64, 16, 64)$, the 1,717-cell coefficient grid gives $879{,}104$ raw forward-pass records, which are grouped into $13{,}736$ stratum-level labels, one for each cell and input/masking condition. Across all strata, 39.0% are classified as diffuse, 27.6% as concentrated, 20.9% as intermediate, 9.0% as boundary, and 3.4% as self-locked. One additional stratum is labeled gradient-starved because it falls below the one-step backward-pass threshold.

The sweep shows that the attention patterns are not described by a single dominant behavior. On the canonical slice, synthetic inputs, unmasked attention, pairwise correlation 0.25, and $\phi = 0$, the different labels correspond to distinct attention statistics (Table 1). Diffuse rows have nearly uniform attention weights, while intermediate rows show a modest peak. Concentrated rows assign most of their mass to one or a few keys, whereas self-locked rows place almost all of their mass on the query's

Table 1: Attention statistics on the canonical reference slice. Entropy is normalized by $\log n$ so that 1 is uniform over the sequence.

Regime	Cells	Entropy	Peak mass	Self mass
Diffuse	99	.990	.028	.017
Intermediate	40	.829	.141	.038
Concentrated	60	.186	.761	.076
Self-locked	8	.022	.980	.971
Boundary	14	.613	.301	.044

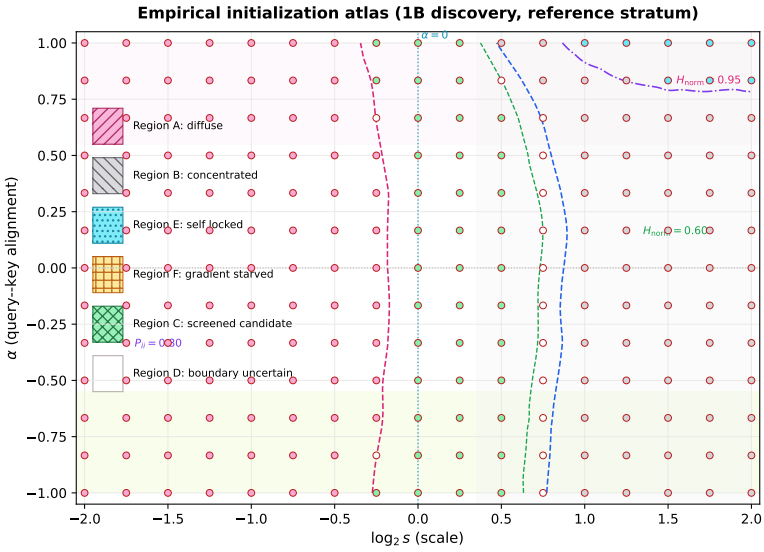


Figure 2: Initialization map at $d = 64$, $m = 16$, $n = 64$. Colors mark bootstrap-supported attention patterns over $(\log_2 s, \alpha)$. The surface is interpolated from measured cells for visualization.

own position. Real post-LayerNorm activations show the same broad range of behaviors, although they are somewhat more concentrated than the synthetic Gaussian inputs.

Figure 2 shows that these are extended regions in $(\log_2 s, \alpha)$, not isolated accidents. Large scale and large positive α move the map toward concentration and self-locking, as the mean formula (6) suggests: the expected self-versus-other gap grows like $\alpha s^2\sqrt{m}$. Small scale keeps attention diffuse, and the intermediate band sits between those extremes.

The map shows clear structure, but the boundaries between regions are not sharp. Shifting the labeling thresholds by 0.05 changes the labels of 9.4% of strata in one direction and 48.9% in the other. Thus, a substantial fraction of the atlas lies close to a classification boundary. This is not surprising: the underlying softmax statistics change continuously, while the labels are defined by thresholds that we choose. Figure 2 should therefore be read as showing broad regions with gradual transitions between them, rather than sharply separated categories.

There is one important distinction to make when interpreting the variance estimates. The sequence-level variance measured in this sweep is not the same quantity as the pairwise variance validated in Stage 1A. The former measures how scores vary across different pairs within a single realized attention matrix, whereas the latter measures how the score for a fixed pair varies across repeated draws of the weights. After applying the total-variance correction and averaging over the 64 seeds for each cell, the two estimates agree closely, with a seed-averaged $R^2 = 0.996$ for the mean. The full calibration is given in Appendix B. This agreement validates the connection between the two estimates, but the sequence-level measurement should not be viewed as an independent proof of Proposition 1 from a single forward pass.

5.2 THE REGIMES DO NOT TRANSFER WITH THE COEFFICIENTS

A natural question is whether these names describe something that holds beyond the one setting where they were measured. To test this, a classifier trained only on the reference-map labels (20% of cells held out) reaches a balanced accuracy of 0.987 and an expected calibration error of 0.015 at recovering those labels — confirming that the Stage 1B atlas is smooth enough to interpolate, though not yet that its labels travel to a new geometry. We use this classifier only to select 120 cells near a decision boundary or a theory/label disagreement, then remeasure those same cells at both $(64, 16, 64)$ and $(128, 16, 128)$.

The preregistered bar for a confirmed cell was strict: staying intermediate across all sixteen replicated conditions, at both settings together. That rule selects zero of the 120 candidate points. Relaxing it in stages shows how far short the strict rule falls: requiring intermediate status in a majority of all sixteen strata leaves only three eligible points, short of the eight needed; only when the bar drops to a majority of the eight *reference-geometry* strata alone do 32 points qualify, from which eight are drawn (spread evenly over α and ϕ) to carry the label into the training comparison below. Separately, comparing majority labels directly across the two geometries — without requiring agreement across every stratum — only 30 of the 120 points keep the same majority name at all.

That failure is visible in the formulas themselves, and in a threshold check besides: entries have standard deviation $s/\sqrt{d}$, while the typical score mean in (6) scales as $s^2\sqrt{m}$. Copying a numerical s to a new (d, m) therefore changes both the weight size and the softmax temperature at once, so a label attached to (s, α, ϕ) at one geometry describes a different initializer at another. The Stage 1C strata are also less forgiving of small perturbations than Stage 1B's: 27.8% of the 1,920 transfer summaries change label under a -0.05 threshold shift, and 73.1% change under $+0.05$. The map is tied to the reference setting at which it was measured, and any later use of it should remeasure at the target geometry rather than copy coefficients across.

5.3 THE TRAINING GAP IS A SCALE EFFECT

Before looking at training loss, 84 coefficient points are frozen into six groups — 15 diffuse, 15 intermediate, 15 concentrated, 15 self-locked, 20 boundary, and the four named endpoints — and trained for 2,000 steps under eight paired seeds on two tasks, for $1{,}344$ total jobs with zero divergences.

Associative recall does not separate the groups: gradient clipping fires on every step for every stratum, and every stratum finishes within 4.171–4.172 (total drop 1.83–1.89 nats). The probe is clip-saturated, so it speaks to nothing about geometry one way or the other.

TinyStories does separate the groups. All eight seed-level contrasts favor the intermediate group over the pooled concentrated-and-self-locked group, with an implied total-loss gap of 0.338 nats under the preregistered pooling rule (Table 2) — a genuine pass, and a true statement about that rule. A second, exploratory reading of the same numbers points to a different story, however. The diffuse, intermediate, and boundary groups were drawn from the Stage 1C fallback and sit at $s \approx 0.78$–0.80 (entry noise $\sigma \approx 0.07$ at the training width); the concentrated and self-locked groups were instead filled from the Stage 1B discovery sweep at median $s = 2.00$ and $s = 2.83$, which at the training geometry corresponds to $\sigma = 0.177$ and 0.250 and a logit-mean prefactor roughly $13\times$ larger than the low-scale band. Grouped by scale rather than by label, intermediate and diffuse finish at nearly the same loss (1.360 and 1.365), as do the four named corners at matched scale; concentrated and self-locked, at their much larger transplanted s, finish substantially worse (1.619 and 2.095). Notably, the high-scale runs clip *less* often than the low-scale band (clip frequency 0.013 versus 0.03–0.16), so the gap is a matter of improving less, not of instability. Figure 3 shows the copied scales; Figure 4 shows the resulting losses. Both readings of the TinyStories result are correct as far as they go, and they should be kept distinct: the preregistered pooling rule passes, and separately, the pass is carried by a scale difference inherited from how the pathology groups were sourced rather than by a geometry difference at matched scale.

Table 2: One-block primary comparison: intermediate minus pooled concentrated/self-locked loss decrease per token. The interval must lie above zero to pass.

Task	Estimate	95% CI	Pass
Associative recall	-1.16×10^{-7}	$[-4.68, 2.47] \times 10^{-7}$	no
TinyStories	$+2.06 \times 10^{-8}$	$[1.86, 2.32] \times 10^{-8}$	yes

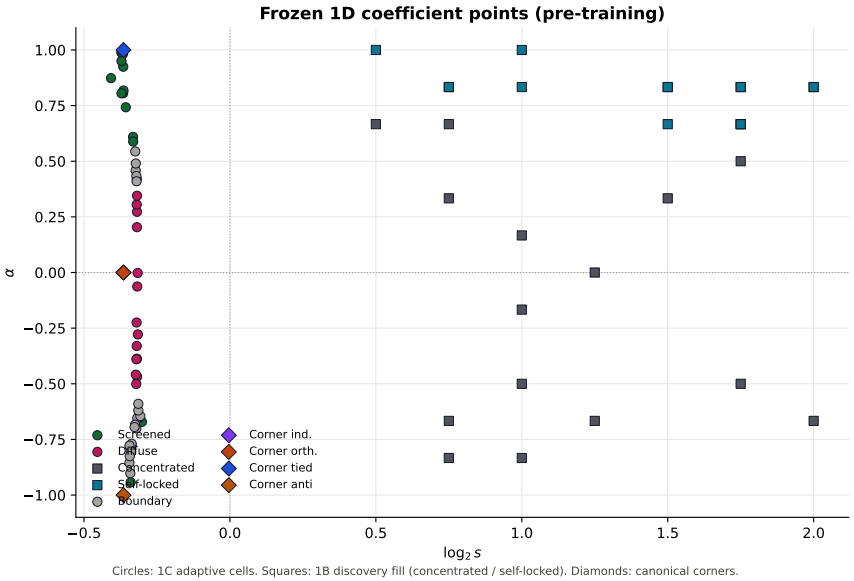


Figure 3: Frozen training points in coefficient space. Intermediate and diffuse points occupy a narrow low-scale band. Concentrated and self-locked points were filled from the discovery sweep at larger s.

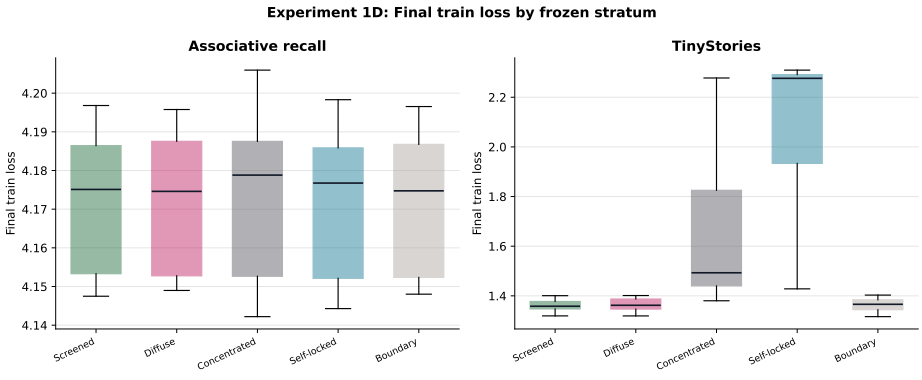


Figure 4: Final one-block training loss. Associative recall is flat. On TinyStories, the high-scale concentrated and self-locked points separate from the low-scale band; intermediate and diffuse points sit close together.

6 DISCUSSION

Writing query–key geometry as a continuous, energy-matched family does what it set out to do. Attention at initialization spreads across several regimes instead of sitting at one fixed operating point: the mean score tracks token similarity through $\alpha s^2\sqrt{m}$, and the orthogonal residual behaves as a genuinely separate finite-width channel from an extra independent Gaussian. These are claims about step zero, and the Monte Carlo checks together with the reference map back them up.

Two stronger readings would go beyond what we have shown. A point in coefficient space describes an attention pattern at the width and context where it was measured, and stays tied to that setting: the same (s, α, ϕ) at a new width and context yields a different softmax temperature, and every point that met our strict transfer bar failed it. Using the map elsewhere means remeasuring there, or changing coordinates so that entry scale and score scale are controlled separately, in the way width-transfer methods such as μP already do for other hyperparameters (Yang et al., 2022). And the one-block TinyStories comparison speaks to scale more than to geometry: intermediate geometry at $s \approx 0.8$ trains about the same as diffuse geometry at that same scale, and the gap we see instead traces to the much larger copied s of the concentrated and self-locked fill. Associative recall, in this implementation, stays clip-saturated across every stratum and so has nothing to say either way.

The next experiment worth running follows directly from that. Remeasure every training group at the target (d, m, n) before attaching a regime name to it. Compare intermediate, diffuse, tied, independent, and orthogonal initializations at a shared s, or at a shared measured score temperature, so scale and geometry finally vary one at a time. Retune the retrieval probe so it operates below the clipping threshold. With those three pieces in place, a multi-layer language-model comparison would be positioned to test whether joint geometry, rather than scale, is what changes optimization.

Even ahead of that test, the family itself is the useful object here. It turns an unspecified product $W_Q^\top W_K$ into a space that can be sampled, predicted, and mapped, and the map in Figures 2 and 6 gives one description of that space, at one width and context.

7 LIMITATIONS

The theory is exact for Gaussian factors at finite d and m, for a single pair of tokens, and for first and second moments. It does not give a law for softmax entropy, for the rank of the attention matrix, or for training. Scores in a sequence are correlated, and our sequence-level variance diagnostic is therefore a different estimand from Proposition 1; we rely on the isolated-pair Monte Carlo for the algebra and report the sequence diagnostic in Appendix B.

The empirical map is one architecture and one reference size. We did not sweep depth, residual scale, RoPE, grouped-query attention, or a second tokenizer. Transfer was tested for a single jump from $(64, 16, 64)$ to $(128, 16, 128)$. Training is a one-block decoder on two probes, one of which is saturated. Concentrated and self-locked training points were not scale-matched to the intermediate group, so the present comparison cannot estimate a geometry effect at fixed s. We also did not run the confirmatory multi-layer language-model study outlined in the original experimental plan.

REPRODUCIBILITY STATEMENT

The accompanying code builds the family (3) without forming a dense projector, evaluates the exact moments (6)–(8), initializes query and key slices head by head, and reruns each measurement from frozen grids and seeds. Formulas were checked before the map was interpreted; regime names were frozen before training; saturated and failed probes are kept in the released records. Appendix A proves the identities used in the main text. Appendix B lists grids, thresholds, model hyperparameters, and the transfer fallback rule.

ETHICS STATEMENT

This work studies random initialization of attention weights. It uses publicly available synthetic text (TinyStories) and a constructed associative-recall task. There are no human subjects, no private data,

and no deployed system. The paper does not propose an initializer for production models, and we see no direct dual-use risk beyond the ordinary use of language-model training code.

AI USE STATEMENT

Generative AI tools were used to aid and polish the writing, for research execution through writing and debugging experiment code, and for checking mathematical derivations against independent Monte Carlo tests. The topic, research ideas, and core contributions were developed by the authors. The authors reviewed and verified the mathematics, numerical results, citations, and prose, and take responsibility for the final content.

REFERENCES

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Ré. Zoology: Measuring and improving recall in efficient language models. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2312.04927.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Han Bao, Ryuichiro Hataya, and Ryo Karakida. Self-attention networks localize when QK-eigenspectrum concentrates. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. arXiv:2402.02009.

Blake Bordelon, Hamza Tahir Chaudhry, and Cengiz Pehlevan. Infinite limits of multi-head transformer dynamics. *arXiv preprint arXiv:2405.15712*, 2024.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Aravind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020. arXiv:2005.14165.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *Proceedings of the 38th International Conference on Machine Learning*, 2021. arXiv:2103.03404.

Ronen Eldan and Yuanzhi Li. TinyStories: How small can language models be and still speak coherent English? *arXiv preprint arXiv:2305.07759*, 2023.

Alessio Giorlandino and Sebastian Goldt. Two failure modes of deep transformers and how to avoid them: A unified theory of signal propagation at initialisation. *arXiv preprint arXiv:2505.24333*, 2025.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

Tomohiro Hayase, Benoît Collins, and Ryo Karakida. Gaussian equivalence for self-attention: Asymptotic spectral analysis of attention matrix. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, pp. 2512–2520, 2026. arXiv:2510.06685.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4246–4253, 2020. arXiv:2010.04245.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Wei Hu, Lechao Xiao, and Jeffrey Pennington. Provable benefit of orthogonal initialization in optimizing deep linear networks. In *International Conference on Learning Representations (ICLR)*, 2020. arXiv:2001.05992.

Xiao Shi Huang, Felipe Perez, Jimmy Ba, and Maksims Volkovs. Improving Transformer optimization through better initialization. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 4475–4483, 2020.

Ali Kayyam, Anusha Madan Gopal, and M Anthony Lewis. Do transformers need three projections? systematic study of QKV variants. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2606.04032.

Md Kowsher, Nusrat Jahan Prottasha, Chun-Nam Yu, Ozlem Garibay, and Niloofar Yousefi. Does self-attention need separate weights in transformers? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pp. 535–543, 2025. arXiv:2412.00359.

Thiziri Nait Saada, Alireza Naderi, and Jared Tanner. Mind the gap: A spectral analysis of rank collapse and signal propagation in attention layers. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. arXiv:2410.07799.

Lorenzo Noci, Chuning Li, Mufan Bill Li, Bobby He, Thomas Hofmann, Chris J. Maddison, and Daniel M. Roy. The shaped transformer: Attention models in the infinite depth-and-width limit. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2306.17759.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018. URL `https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf`.

Clayton Sanford, Daniel Hsu, and Matus Telgarsky. One-layer transformers fail to solve the induction heads task. *arXiv preprint arXiv:2408.14332*, 2024.

Hemanth Saratchandran and Simon Lucey. Conditioned initialization for attention. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2609.07086.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations (ICLR)*, 2014. arXiv:1312.6120.

Asher Trockman and J. Zico Kolter. Mimetic initialization of self-attention layers. In *Proceedings of the 40th International Conference on Machine Learning*, 2023. arXiv:2305.09828.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. arXiv:1706.03762.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 5797–5808, 2019. arXiv:1905.09418.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning*, 2020. arXiv:2002.04745.

Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs V: Tuning large neural networks via zero-shot hyperparameter transfer. In *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2203.03466.

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua M. Susskind. Stabilizing transformer training by preventing attention entropy collapse. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 40770–40803, 2023. arXiv:2303.06296.

Hongyi Zhang, Yann N. Dauphin, and Tengyu Ma. Fixup initialization: Residual learning without normalization. In *International Conference on Learning Representations (ICLR)*, 2019. arXiv:1901.09321.

A PROOFS

Throughout, $A, C \in \mathbb{R}^{m\times d}$ have i.i.d. entries $\mathcal{N}(0, 1/d)$, G has i.i.d. entries $\mathcal{N}(0, 1/(d-m))$, $B = G(I - P_A)$, and (W_Q, W_K) follow (3). Inputs X are treated as fixed. Write Π for P_A when the subscript is clear.

Lemma 1 (Energy matching)**.** *For every admissible* (α, β, γ)*,*

$$\mathbb{E}\|W_Q\|_F^2 = \mathbb{E}\|W_K\|_F^2 = s^2 m, \qquad \mathbb{E}[(W_K)_{ab}^2] = \frac{s^2}{d},$$
$$\mathbb{E}[W_Q^\top W_K] = s^2\alpha\frac{m}{d} I_d, \qquad \mathbb{E}[W_Q W_K^\top] = s^2\alpha I_m.$$

Moreover $AB^\top = 0$ *almost surely whenever* A *has full row rank.*

Proof. Each row of A has expected squared norm 1, so $\mathbb{E}\|A\|_F^2 = m$ and $\mathbb{E}\|W_Q\|_F^2 = s^2 m$. The same holds for C. Conditional on A, the rows of G are independent of Π, and

$$\mathbb{E}[BB^\top \mid A] = \frac{\mathrm{tr}(I-\Pi)}{d-m} I_m = I_m,$$

hence $\mathbb{E}\|B\|_F^2 = m$. Rotational invariance of a uniformly random rank-m subspace gives $\mathbb{E}[(I - \Pi)_{jj}] = (d-m)/d$, so $\mathbb{E}[B_{ab}^2] = 1/d$. The product $AB^\top$ vanishes because each row of B lies in the orthogonal complement of the rows of A. Cross terms that involve C vanish in expectation by independence and mean zero. Expanding W_K and taking expectations therefore yields $\mathbb{E}\|W_K\|_F^2 = s^2 m$ and $\mathbb{E}[(W_K)_{ab}^2] = s^2/d$. Finally $\mathbb{E}[A^\top A] = (m/d) I_d$ and $\mathbb{E}[AA^\top] = I_m$, which give the two cross-moment identities. □

A.1 PROOF OF PROPOSITION 1

Proof. Decompose the score into three channels,

$$\ell_{ij} = \frac{s^2}{\sqrt{m}}\Big(\alpha\,(Ax_i)^\top(Ax_j) + \beta\,(Ax_i)^\top(Bx_j) + \gamma\,(Ax_i)^\top(Cx_j)\Big).$$

Call the three summands $\ell^{(\alpha)}$, $\ell^{(\beta)}$, and $\ell^{(\gamma)}$.

Mean. Conditional on A, each row of B is centered, so $\mathbb{E}[\ell^{(\beta)} \mid A, X] = 0$. The factor C is independent of A and centered, so $\mathbb{E}[\ell^{(\gamma)} \mid X] = 0$. For a single row $a \sim \mathcal{N}(0, I_d/d)$,

$$\mathbb{E}[(a^\top x_i)(a^\top x_j)] = x_i^\top \mathbb{E}[aa^\top] x_j = c_{ij}.$$

Summing m independent rows and multiplying by $s^2\alpha/\sqrt{m}$ produces (6).

Vanishing cross-covariances. The three channels are uncorrelated. The β channel is a linear function of G given A, hence has mean zero given A and is uncorrelated with any function of A, including $\ell^{(\alpha)}$. The γ channel is a linear function of C, independent of (A, B) and centered. Variance therefore adds across channels.

Shared channel. Let $a \sim \mathcal{N}(0, I_d/d)$. Isserlis' theorem gives

$$\mathbb{E}[(a^\top x_i)^2 (a^\top x_j)^2] = v_i v_j + 2c_{ij}^2,$$

so

$$\mathrm{Var}\big((a^\top x_i)(a^\top x_j)\big) = v_i v_j + c_{ij}^2.$$

The m rows of A are independent, and $\ell^{(\alpha)}$ multiplies their sum by $s^2\alpha/\sqrt{m}$, hence

$$\mathrm{Var}(\ell^{(\alpha)} \mid X) = s^4\alpha^2(v_i v_j + c_{ij}^2).$$

Independent channel. Given A, the inner products $c_t^\top x_j$ are independent $\mathcal{N}(0, v_j)$ and independent of A. Thus

$$\mathrm{Var}\big((a_t^\top x_i)(c_t^\top x_j)\big) = \mathbb{E}[(a_t^\top x_i)^2]\, v_j = v_i v_j.$$

Summing m rows and scaling by $s^2\gamma/\sqrt{m}$ produces $s^4\gamma^2 v_i v_j$.

Orthogonal channel. Conditional on A, $b_t^\top x_j = g_t^\top (I-\Pi)x_j$ has variance $x_j^\top (I-\Pi)x_j/(d-m)$, and the rows of G are independent. Therefore

$$\mathrm{Var}(\ell^{(\beta)} \mid A, X) = s^4\beta^2 \cdot \frac{\|Ax_i\|^2}{m} \cdot \frac{x_j^\top (I-\Pi)x_j}{d-m}. \tag{9}$$

It remains to take the expectation over A. The row space of a Gaussian matrix A is uniform on the Grassmannian and independent of the singular values, and $\mathbb{E}[A^\top A] = \mathbb{E}[\Pi] = (m/d)I_d$, so $\mathbb{E}[A^\top A \mid \Pi] = \Pi$. Hence $\mathbb{E}[\|Ax_i\|^2 \mid \Pi] = x_i^\top \Pi x_i$. Substituting into (9) yields

$$g_{ij} = \frac{1}{m(d-m)} \mathbb{E}\big[(x_i^\top \Pi x_i)\,(x_j^\top (I-\Pi)x_j)\big].$$

For a uniformly random rank-m orthogonal projector on $\mathbb{R}^d$,

$$\mathbb{E}[\Pi_{pq}\Pi_{rs}] = a\,\delta_{pq}\delta_{rs} + b\,(\delta_{pr}\delta_{qs} + \delta_{ps}\delta_{qr}), \tag{10}$$

with

$$a = \frac{m(md+m-2)}{d(d-1)(d+2)}, \qquad b = \frac{m(d-m)}{d(d-1)(d+2)}.$$

Contracting (10) against $x_i \otimes x_i \otimes x_j \otimes x_j$ gives

$$\mathbb{E}[(x_i^\top \Pi x_i)(x_j^\top \Pi x_j)] = a\,d^2 v_i v_j + 2b\,d^2 c_{ij}^2,$$

while $\mathbb{E}[x_i^\top \Pi x_i] = m v_i$. Therefore

$$\begin{aligned}\mathbb{E}\big[(x_i^\top \Pi x_i)\,(x_j^\top (I-\Pi)x_j)\big] &= md\, v_i v_j - ad^2 v_i v_j - 2bd^2 c_{ij}^2 \\ &= md\left(v_i v_j \left(1 - \frac{m(d+1)-2}{(d-1)(d+2)}\right) - \frac{2(d-m)}{(d-1)(d+2)} c_{ij}^2\right).\end{aligned}$$

The term in parentheses simplifies because

$$(d-1)(d+2) - m(d+1) + 2 = (d-m)(d+1).$$

Dividing by $m(d-m)$ produces (8). Adding the three variances completes the proof. □

If $i = j$, then $c_{ii} = v_i$ and $g_{ii} = dv_i^2/(d+2)$. For fixed normalized tokens, $g_{ij} \to v_i v_j$ as $d \to \infty$, so the orthogonal and independent channels agree asymptotically and differ at finite width.

A.2 Additional exact identities

The following identities are used as implementation checks rather than as training theorems. Proofs are included for completeness.

Proposition 2 (Score gradients)**.** *Let $\mathcal{L}$ be any scalar function of the score matrix L, and write $D = \partial\mathcal{L}/\partial L$. Then*

$$\frac{\partial \mathcal{L}}{\partial W_Q} = \frac{1}{\sqrt{m}}(W_K X)\, D^\top X^\top, \qquad \frac{\partial \mathcal{L}}{\partial W_K} = \frac{1}{\sqrt{m}}(W_Q X)\, D X^\top.$$

Proof. $L = X^\top W_Q^\top W_K X/\sqrt{m}$. Differentiating the bilinear map $(W_Q, W_K) \mapsto L$ in Frobenius inner product against D yields the two displayed formulas. □

Proposition 3 (Cross-head mean-bias statistic)**.** *For two independent heads with coefficients (s_r, α_r), $r = 1, 2$, write $M_r = W_Q^{(r)\top} W_K^{(r)}$ and $f(\alpha) = 1 + \alpha^2(m+1)/d$. Then*

$$\mathbb{E}[M_r] = s_r^2 \alpha_r \frac{m}{d} I_d, \qquad \mathbb{E}\|M_r\|_F^2 = s_r^4 m f(\alpha_r),$$

and

$$\frac{\mathbb{E}\langle M_1, M_2\rangle_F}{\sqrt{\mathbb{E}\|M_1\|_F^2\, \mathbb{E}\|M_2\|_F^2}} = \frac{\alpha_1 \alpha_2 m}{d\sqrt{f(\alpha_1) f(\alpha_2)}}.$$

Proof. The first display is Lemma 1. Expanding $\|M\|_F^2 = \|W_Q W_K^\top\|_F^2$ and using Gaussian fourth moments of the rows of A, together with $AB^\top = 0$ and independence of C, produces $\mathbb{E}\|M\|_F^2 = s^4 m f(\alpha)$. Independent heads have $\mathbb{E}\langle M_1, M_2\rangle_F = \langle \mathbb{E}M_1, \mathbb{E}M_2\rangle_F$, which is the remaining identity. The ratio is a ratio of expectations, not $\mathbb{E}[\cos(M_1, M_2)]$. □

B Experimental details

B.1 Moment check

The Monte Carlo gate uses $d \in \{32, 64, 128\}$, $m/d \in \{1/8, 1/4, 1/2\}$, $c_{ij} \in \{-0.5, 0, 0.5, 1\}$, and 24 coefficient settings, for 864 configurations. Each configuration starts at 20,000 draws and may grow to 200,000 until both standard-error targets are met. Mean and variance are tested with Holm–Bonferroni control at 0.01 over 1,728 tests. Auxiliary checks require $\|AB^\top\|_F/(\|A\|_F\|B\|_F) < 10^{-6}$, component energies within 1% of target, entry variances within 2% of s^2/d, entrywise query–key correlation within 0.02 of α, and relative finite-difference gradient error below 10^{-5}. All of these passed; the independent endpoint also matched an ordinary Gaussian initializer (variance ratio 1.009).

A sequence of tokens is a different estimand from a single pair. The variance across pairs in one forward pass includes an extra term

$$\mathrm{Var}_{ij}(\ell_{ij}) = \mathbb{E}_{ij}\big[\mathrm{Var}_W(\ell_{ij} \mid x)\big] + \mathrm{Var}_{ij}\big(\mathbb{E}_W[\ell_{ij} \mid x]\big). \tag{11}$$

After averaging over 64 seeds, the sequence-mean prediction has $R^2 = 0.996$. The uncorrected pairwise variance diagnostic is worse, as (11) requires. Figure 5 shows the correction; it is a diagnostic, not a second proof of Proposition 1.

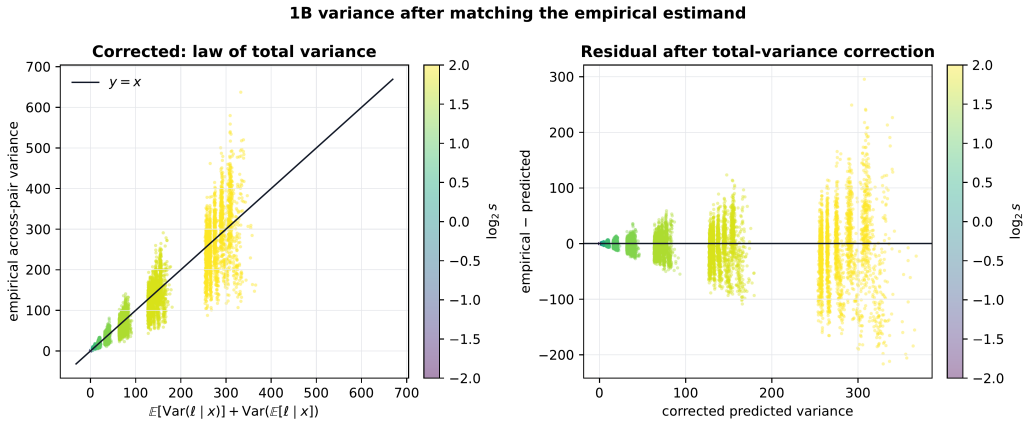


Figure 5: Sequence-level score variance after adding the second term in (11). Calibration improves relative to the uncorrected scatter, but the confirmatory test remains the isolated-pair Monte Carlo in Figure 1.

B.2 Map, names, and transfer

The reference grid is 17 values of $\log_2 s \in [-2, 2]$, 13 values of $\alpha \in [-1, 1]$, and 9 values of $\phi \in [0, \pi/2]$, giving 1,717 unique cells. Each cell is evaluated in eight strata with 64 paired seeds. Names use 95% seed-cluster bootstrap intervals and the precedence in Section 4. A cell is diffuse if the lower interval on median normalized entropy is at least 0.95 and the upper interval on distance from uniform is at most 0.10. It is concentrated if the upper entropy interval is at most 0.50, or the lower interval on median peak mass is at least 0.80. It is self-locked if the lower interval on median self-mass is at least 0.80. It is intermediate if none of those rules fire, normalized entropy lies between 0.60 and 0.95, and attention-matrix effective rank exceeds a pilot floor of 0.275. Jacobian and update-to-weight floors from the pilot are 0.0556 and 9.75×10^{-6}; gradient starvation almost never occurred (one stratum summary). An interval that crosses a cutoff remains unlabeled.

The discrete labels in Figure 2 come from those thresholds. Figure 6 shows the underlying field: median normalized row entropy as a continuous surface over $(\log_2 s, \alpha)$ on the canonical slice. The high plateau is the diffuse region; entropy then falls toward large scale and large positive α, into the concentrated and self-locked parts of the family. As in the atlas, the surface is interpolated from measured cells for visualization.

The transfer classifier was trained only on reference names, with 20% of cells held out (balanced accuracy 0.987, log loss 0.0355, expected calibration error 0.015). It selected 120 cells near decision boundaries for measurement at both the reference setting and $(128, 16, 128)$. Table 3 records the decision rule used to freeze training heads. Label sensitivity near those boundaries is high by construction: 27.8% of the 1,920 transfer summaries change name under a -0.05 cutoff shift, and 73.1% under a $+0.05$ shift.

Table 3: Decision rule for freezing training heads. The strict intersection was empty.

Rule	Eligible points	Used
Intermediate in all 16 replicated strata	0	no
Intermediate in a majority of all 16 strata	3	no (< 8)
Intermediate in a majority of eight reference strata	32	yes
Intermediate in any reference stratum	117	no

B.3 One-block training

The decoder has width 128, four heads, byte vocabulary 256, zero dropout, context 128, batch size 64, 2,000 AdamW steps at 3×10^{-4}, and clip threshold 1.0. Table 4 lists the frozen groups. Primary

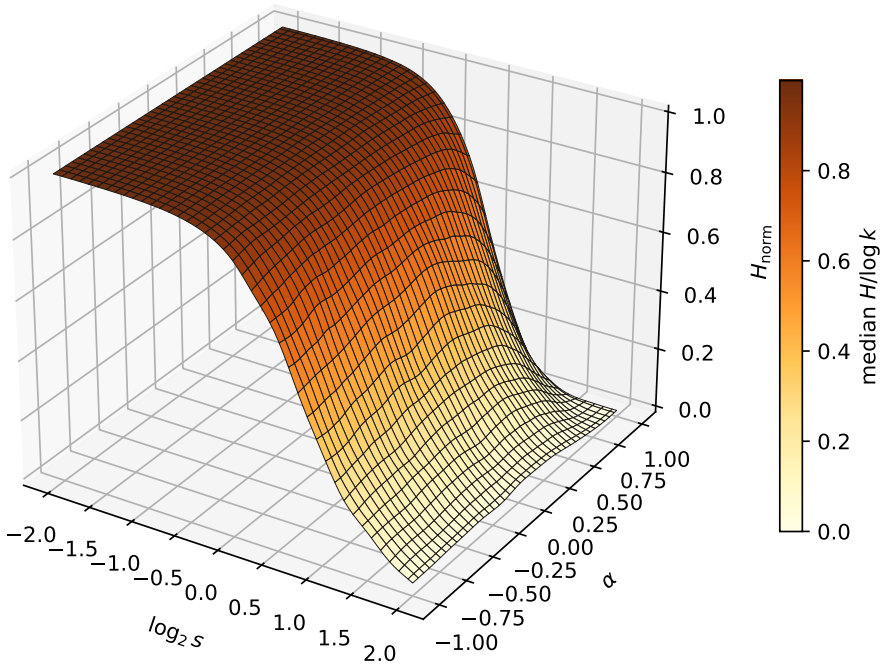


Figure 6: Normalized row entropy over the measured family. The high-entropy plateau and the drop into concentrated and self-locked regions are the same landscape as Figure 2, viewed as a continuous surface.

analysis uses 10,000 bootstrap draws over the eight seed-level contrasts. All 1,344 jobs completed, with no divergences. Figure 7 shows representative traces: recall reaches a shared clipped plateau, while TinyStories separates high scale from low scale throughout training.

Table 4: Frozen training groups. Median s is the copied coefficient, not a remeasured temperature at width 128.

Group	Source	Points	Median s
Diffuse	reference majority	15	.802
Intermediate	reference majority	15	.777
Boundary	reference majority	20	.799
Concentrated	discovery fill	15	2.000
Self-locked	discovery fill	15	2.828
Named endpoints	matched-scale controls	4	.777

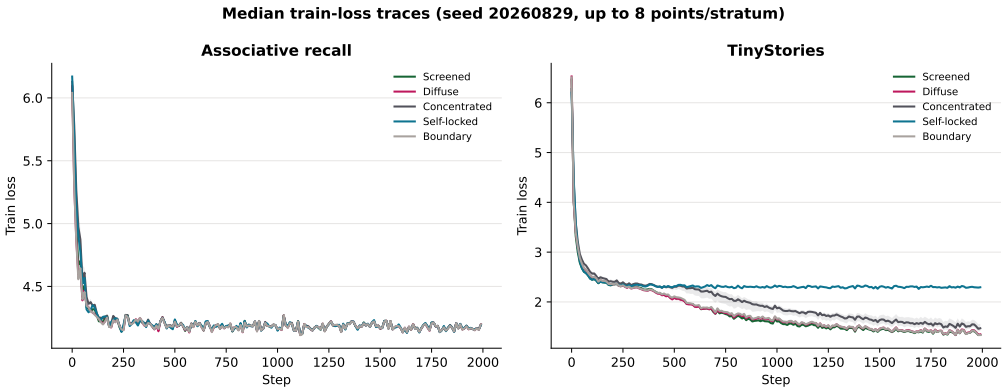


Figure 7: Median learning curves. Associative recall converges to a common clipped regime. TinyStories separates the high-scale fill from the low-scale band for the whole run.