SEQUENTIAL STABILITY BOUNDS FOR ORDER-DEPENDENT IN-CONTEXT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

In-context learning (ICL) accuracy swings by up to 39 percentage points under demonstration reordering, yet every existing theoretical framework treats the prompt as an unordered set and therefore certifies one number for all $k!$ orderings. We formalize causal attention masking as the source of this asymmetry: tracking a single-example replacement through the causal chain gives a closed-form position-dependent amplification factor $W(p, k, L)$, strictly decreasing in slot p for $L \geq 2$ and exactly flat for $L{=}1$, which couples with a per-source sensitivity score η_j to define *sequential stability coefficients* β_σ—the first bounded-differences constant for ICL that varies with the permutation. Three theorems follow: an order-dependent generalization bound (Theorem 4), a rearrangement-based certification of low-sensitivity-first ordering with an $\Omega(\Delta^2/(kC^2))$ separation (Theorem 5), and a non-vacuity criterion in $(B^3R^2, r_{\max}, L)$ that is independent of embedding dimension (Theorem 6); a spherical-concentration certificate resolves distinct sources at cost $d_x = \Theta(\ln k)$ (Lemma 7). A hierarchy of diagnostic experiments matches the theory: $L{=}1$ produces spread below 10^{-8} across 120 permutations, matched trained and untrained networks are equivalent under normalized attention (TOST $p = 0.033$, $n{=}30$)—so ordering effects are architectural, not learned—and four 7–8B instruction-tuned LLMs show 8–39% accuracy spread on AG News and RTE. The multiplicative decomposition $W(p) \cdot \eta_{\sigma(p)}$ is recovered from the empirical gap matrix: the position factor tracks closed-form W at $r_s = 0.82$–0.98, an intraclass ceiling $\mathcal{R}^2_{\max} = 0.35$–$0.98$ (Lemma 22) quantifies the per-example share, and a learned score saturates $\geq$92% of that ceiling at $d_x \leq 32$—identifying the residual $1{-}\mathcal{R}^2_{\max}$ as the framework's predicted many-body component.

1 INTRODUCTION

Reordering the same few-shot demonstrations shifts accuracy by up to 39 points on frontier LLMs, an effect reproducible, robust to calibration, and often larger than the gap between competing prompt-selection strategies at the same budget (Lu et al., 2022; Zhao et al., 2021; Liu et al., 2024). Every existing ICL theory remains permutation-invariant: Bayesian ICL (Xie et al., 2022), implicit-gradient-descent analogies (von Oswald et al., 2023), function-class learnability (Garg et al., 2022; Akyürek et al., 2023; Wies et al., 2023), and the tightest algorithmic-stability route (Li et al., 2023; Zhang et al., 2024) derive their guarantees from permutation-invariant quantities and assign the same certificate to all $k!$ orderings. Prior work informally attributes ordering effects to causal masking (Lu et al., 2022; Liu et al., 2024), but no result turns that observation into a permutation-dependent generalization guarantee.

We close the gap by tracking analytically how a single demonstration replacement propagates through a causal-attention stack. Because the mask is triangular, replacing the example at slot p perturbs every subsequent position, and each perturbation propagates forward; the sum of forward paths reaching the query is strictly larger when p is small. Under causal linear attention (Equation (3)) this path counting is exact and yields a closed-form amplification factor $W(p, k, L)$ (Proposition 1) that strictly decreases in p for $L \geq 2$ (Figure 2), is exactly flat for $L{=}1$, and steepens with depth. Coupling $W(p)$ with a per-source sensitivity score $\eta_j := \mathrm{diam}(\mathrm{supp}(\phi \circ \mathcal{D}_j))$ defines the *sequential stability coefficient*

$$\beta_\sigma \;=\; \left(\tfrac{1}{k} \sum_{p=1}^{k} W(p)^2 \, \eta^2_{\sigma(p)} \right)^{1/2}, \tag{1}$$

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.

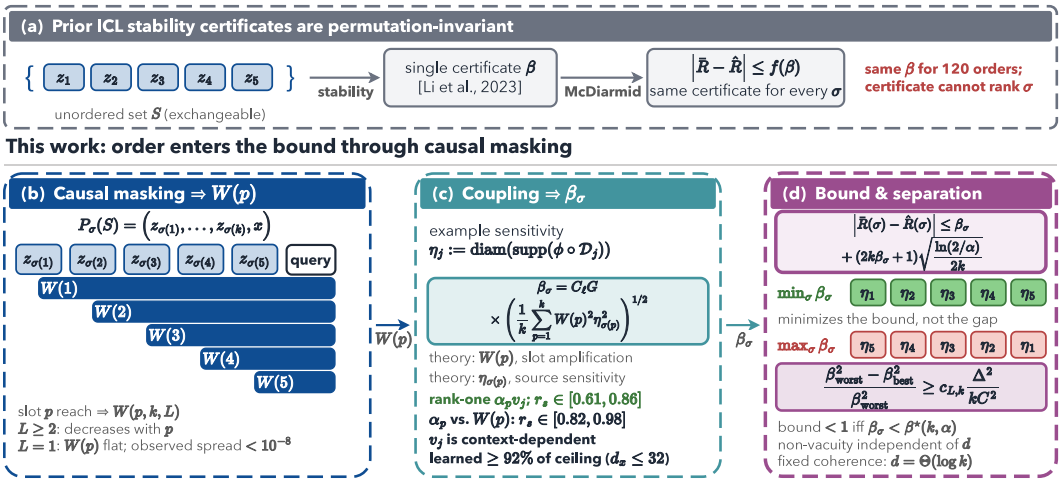


Figure 1: Order enters the ICL stability characterization through causal masking. **(a)** Permutation-invariant stability theory assigns the same certificate to all $k!$ orderings and cannot rank them. **(b)** Causal masking gives the closed-form amplification $W(p, k, L)$ (Proposition 1), decreasing in p for $L \geq 2$ and position-independent for $L{=}1$; the latter yields observed spread $< 10^{-8}$. **(c)** The bound couples $W(p)$ with source sensitivity $\eta_{\sigma(p)}$ to define β_σ (Definition 3); empirically, the realized gap matrix follows $\alpha_p v_j$, with α_p aligned to $W(p)$ and context-dependent v_j partly recoverable by learned scores. **(d)** β_σ gives a permutation-dependent bound minimized by placing low-sensitivity examples first (Theorem 5); non-vacuity is independent of the embedding dimension, and a fixed-coherence source certificate requires only $d_x = \Theta(\log k)$.

a McDiarmid-style constant whose value depends on σ. Three theorems anchor the framework: Theorem 4 turns β_σ into a generalization bound with σ explicit; Theorem 5 minimizes it by low-sensitivity-first ordering with an $\Omega(\Delta^2/(kC^2))$ best-vs-worst separation; and Theorem 6 extracts a closed-form non-vacuity criterion in $(B^3R^2, r_{\max}, L)$ independent of embedding dimension, complemented by a spherical-concentration certificate (Lemma 7) that resolves sources at dimension cost $d_x = \Theta(\ln k)$.

The framework yields a hierarchy of falsifiable predictions. *Algebraically*, $W(p, k, 1)$ is exactly flat, so a single-layer causal network must produce zero ordering effects; across 120 permutations we observe spread below 10^{-8}. *Structurally*, if the mechanism is architectural rather than learned, matched trained and untrained networks should be indistinguishable under normalized attention; TOST equivalence holds at $p = 0.033$ ($n{=}30$). *Quantitatively*, the empirical gap matrix factors as $W(p) \cdot \eta_{\sigma(p)}$: the fitted position factor tracks W at $r_s = 0.82$–0.98, a variance-ceiling analysis (Lemma 22) caps single-example predictors at $\mathcal{R}^2_{\max} = 0.35$–$0.98$, and a learned score reaches $\geq 92\%$ of that ceiling at $d_x \leq 32$—identifying the residual as an irreducible many-body component.

Contributions. We turn the informal identification of causal masking into (i) the first order-dependent ICL generalization bound via bounded differences through causal chains (§4); (ii) a closed-form $W(p, k, L)$, a rearrangement-based certification of low-sensitivity-first ordering with an $\Omega(\Delta^2/(kC^2))$ gap, and a non-vacuity criterion in $(B^3R^2, r_{\max}, L)$ independent of dimension (§4.1–4.4); (iii) a hierarchy of empirical tests from single-layer invariance to frontier-LLM spread (§5); (iv) a variance ceiling $\mathcal{R}^2_{\max}$ separating identifiable per-example content from the irreducible many-body residual (§5.4); (v) a softmax extension preserving single-layer invariance and near-uniform monotonicity (§4.5).

2 Related Work

Permutation-invariant ICL theory. Existing ICL frameworks—Bayesian mixture (Xie et al., 2022), implicit gradient descent (von Oswald et al., 2023), function-class regression (Garg et al.,

2022; Akyürek et al., 2023), algorithm selection (Bai et al., 2023), learnability (Wies et al., 2023), and finite-sample stability (Li et al., 2023; Zhang et al., 2024)—derive guarantees from permutation-invariant quantities and assign identical certificates to all $k!$ orderings. We give a position-dependent replacement for the bounded-differences constant in the stability route, recovering the invariant limit when W is flat.

Empirical ordering effects and heuristics. Zhao et al. (2021) traced part of the swing to miscalibration; Lu et al. (2022) proposed entropy-based selection without a causal mechanism; Liu et al. (2024) established position-dependent information use; Min et al. (2022) showed label correctness barely affects ICL; Hendel et al. (2023) propose additive task vectors. We identify the position factor *analytically* through $W(p)$ and cap the example factor by a measurable ceiling $\mathcal{R}^2_{\max}$ (Lemma 22).

Mechanistic and data-ordering context. Induction heads (Olsson et al., 2022) and positional-attention circuits (Ahn et al., 2023) describe how transformers propagate context; Chan et al. (2022); Raventós et al. (2023) govern ICL emergence from data; Dong et al. (2024) flag ordering as open. Outside ICL, curriculum learning (Bengio et al., 2009) aligns with low-sensitivity-first, continual learning (Kirkpatrick et al., 2017) shows the opposite recency bias, and both effects average out under SGD reshuffling (Mishchenko et al., 2020); ICL's few-shot regime has no such averaging. Classical stability (McDiarmid, 1989; Bousquet & Elisseeff, 2002; Steinke & Zakynthinou, 2020) assumes uniform bounded differences that the causal mask violates.

3 Preliminaries and Problem Formulation

Model and prompt. A pre-trained transformer f_θ receives k demonstrations $z_j = (x_j, y_j)$ and a query x; a permutation $\sigma \in \mathfrak{S}_k$ places them via $P_\sigma(S) = (z_{\sigma(1)}, \ldots, z_{\sigma(k)}, x)$. Slot indices $p = 1, \ldots, k$ run from earliest (furthest from the query) to latest.

Heterogeneous sources and sensitivity score. Each example is drawn from a distinct source, $z_j \sim \mathcal{D}_j$; i.i.d. sources collapse any bound to permutation invariance, but few-shot prompts routinely mix distinct classes or task variants, and heterogeneity is the natural regime where ordering can matter. Under a unit-Lipschitz embedding ϕ, the *source sensitivity score* $\eta_j := \operatorname{diam}\big(\operatorname{supp}(\phi \circ \mathcal{D}_j)\big)$ caps how much replacing example j can perturb the embedded input.

Risk objects. With a bounded C_ℓ-Lipschitz loss $\ell \in [0, 1]$, the empirical and expected risks for a fixed σ are

$$\hat{R}(f_\theta, \sigma) = \tfrac{1}{k} \sum_j \ell\big(f_\theta(P_\sigma(S); x_j),\, y_j\big), \quad \bar{R}(f_\theta, \sigma) = \tfrac{1}{k} \sum_j \mathbb{E}_{z'_j \sim \mathcal{D}_j}\, \ell\big(f_\theta(P_\sigma(S); x'_j),\, y'_j\big); \quad (2)$$

$|\bar{R} - \hat{R}|$ is the generalization gap, parameterized by σ.

Causal linear attention. We analyze single-head causal linear attention with residuals,

$$h_t^{(l+1)} = h_t^{(l)} + \tfrac{1}{t} \sum_{s=1}^{t} \langle W_Q h_t^{(l)},\, W_K h_s^{(l)} \rangle\, W_V h_s^{(l)}, \quad (3)$$

with prefix normalization $1/t$, spectral bounds $\max(\|W_Q\|, \|W_K\|, \|W_V\|) \le B$, representation bound $\|h_t^{(l)}\| \le R$, and a G-Lipschitz readout. Two composite constants organise the analysis: self-propagation $\rho := 1 + 3B^3R^2$ and cross-token coupling $\gamma := 2B^3R^2$; compounding γ/ρ over L layers and $(k-p)$ downstream positions produces the position-dependent amplification in §4.1. Causal linear attention is the analytical primitive; §4.5 and Appendix G lift the amplification bound to softmax and recover the same position profile whenever attention is near-uniform, the regime trained multi-layer softmax models occupy (mean $H/H_{\max} = 0.95$; §5.5).

Problem. We seek a bound on $|\bar{R}(f_\theta, \sigma) - \hat{R}(f_\theta, \sigma)|$ that is (a) explicit in σ, (b) reduces to the exchangeable case when causal amplification is flat, and (c) yields a data-independent ordering rule. Define the *amplification factor* $W(p, k, L) := \sup \|h_{k+1}^{(L)} - \tilde{h}_{k+1}^{(L)}\|/D$ (worst-case propagation of a unit input perturbation at slot p through L layers, with $D := \sup_j \eta_j$); coupling it with η_j yields the sequential stability coefficient of Definition 3.

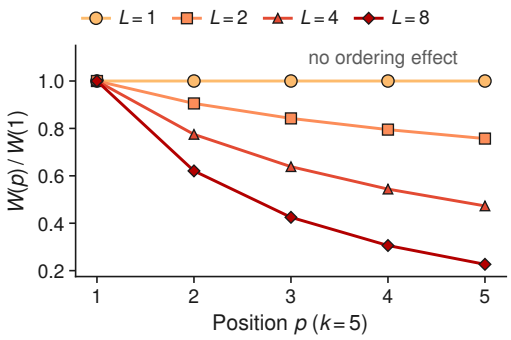


Figure 2: Normalized amplification factor $W(p)/W(1)$ as a function of position for $L \in \{1, 2, 4, 8\}$ (k=5). For L=1, W is flat (no ordering effects); for $L \geq 2$, W strictly decays, with steeper decay at greater depth.

4 Sequential Stability Framework

A single-example replacement in a causal stack perturbs every subsequent position before reaching the query, so earlier slots admit more forward paths and their bounded-differences constant is strictly larger. §4.1 extracts the closed-form amplification; §4.2–4.3 plug it into McDiarmid to obtain the permutation-dependent bound and its minimizer; §4.4 names the non-vacuity regime; §4.5 extends predictions to softmax; §4.6 supplies an estimator of the source sensitivities.

4.1 Causal Chain Amplification

Proposition 1 gives a closed-form upper bound $W(p, k, L) \geq w(p, k, L)$ on the true amplification $w(p, k, L) := \sup \|h_{k+1}^{(L)} - \tilde{h}_{k+1}^{(L)}\|/D$.

Proposition 1 (Causal Chain Amplification)**.** *Under the assumptions of §3 (causal linear attention, weight bound B, representation bound R), for $1 \leq p \leq k$ and $L \geq 1$:*

$$W(p, k, L) = \frac{1}{k+1} \sum_{m=1}^{\min(L,\, k+1-p)} \binom{L}{m} \rho^{L-m} \gamma^m \, e_{m-1}\Big(\tfrac{1}{p+1}, \ldots, \tfrac{1}{k}\Big), \tag{4}$$

where $e_r(\cdot)$ denotes the r-th elementary symmetric polynomial ($e_0 := 1$).

The proof (Appendix A) reduces to $\varepsilon^{(l+1)} \leq M\, \varepsilon^{(l)}$ with $M = \rho I + \gamma N$, N strictly lower-triangular ($N^{k+1} = 0$); the $(k{+}1, p)$ entry of M^L counts causal chains of length $\leq L$ from p to the query.

Corollary 2 (Monotonicity, single-layer, depth)**.** *For $L \geq 2$, $W(p, k, L)$ strictly decreases in p and $W(1, k, L)/W(k, k, L)$ strictly increases in L; at L=1, $W(p, k, 1) = \gamma/(k{+}1)$ is constant, so* ***the theory predicts zero ordering effects****.*

4.2 Order-Dependent Stability Bound

Definition 3 (Sequential Stability Coefficient)**.** *Given the source-diameter score $\eta_j := \sup_{z, z' \in \mathrm{supp}(\mathcal{D}_j)} \|\phi(z) - \phi(z')\|$, the slot replacement sensitivity is $\delta_p(\sigma) := C_\ell\, G\, W(p, k, L)\, \eta_{\sigma(p)}$, and*

$$\beta_\sigma := \left(\frac{1}{k} \sum_{p=1}^{k} \delta_p(\sigma)^2\right)^{1/2} = C_\ell\, G \left(\frac{1}{k} \sum_{p=1}^{k} W(p)^2\, \eta_{\sigma(p)}^2\right)^{1/2}. \tag{5}$$

Since $W(p)$ decreases with position (Corollary 2), placing high-sensitivity sources early inflates β_σ and the certified gap.

Theorem 4 (Order-Dependent Generalization Bound)**.** *Under the assumptions of §3, for any fixed permutation σ and confidence level $\alpha \in (0, 1)$, with probability at least $1 - \alpha$ over the draw of*

$S = (z_1, \ldots, z_k)$:

$$\left|\bar{R}(f_\theta, \sigma) - \hat{R}(f_\theta, \sigma)\right| \;\le\; \beta_\sigma + (2k\beta_\sigma + 1)\sqrt{\frac{\ln(2/\alpha)}{2k}} \tag{6}$$

Proof sketch (Appendix B). Symmetrization gives $|\mathbb{E}[\bar{R} - \hat{R}]| \le \beta_\sigma$; replacing z_j shifts the gap by $\le 2\delta_{p_j}(\sigma) + 1/k$; McDiarmid with position-dependent c_p and Minkowski's inequality yield the form.

Remark 1 (Data-Dependent Ordering). *When σ depends on data, a union bound over $k!$ permutations replaces $\ln(2/\alpha)$ by $\ln(2k!/\alpha)$, at multiplicative cost $O(\sqrt{k \ln k})$.*

4.3 Separation Between Best and Worst Orderings

Theorem 5 (Separation). *For $L \ge 2$ and $k \ge 2$: **(i)** β_σ is minimized by ascending sensitivity order $\eta_{\sigma(1)} \le \cdots \le \eta_{\sigma(k)}$ (low-sensitivity first), maximized by the reverse. **(ii)** With $\Delta := \eta_{\max} - \eta_{\min}$ and $C := \eta_{\max}$:*

$$\frac{\beta^2_{\text{worst}} - \beta^2_{\text{best}}}{\beta^2_{\text{worst}}} \;\ge\; c_{L,k} \cdot \frac{\Delta^2}{k\, C^2}, \qquad c_{L,k} := 1 - \Big(\frac{W(k)}{W(1)}\Big)^2 \in (0, 1). \tag{7}$$

Proof sketch. β^2_σ is an inner product between decreasing $(W(p)^2)$ and permuted $(\eta^2_{\sigma(p)})$; (i) is the rearrangement inequality (Hardy et al., 1952), (ii) swaps extremes only.

4.4 When the Bound Has Content

Bound (6) is informative only below 1: an *architectural* condition (non-triviality) and a *geometric* one (resolvability). Model each source on a spherical cap $\mathrm{cap}(m_j, r_j) \subset \mathbb{S}^{d_x - 1}$ ($r_j \le r_{\max}$); the cap diameter pins $\eta_j \le 2 \sin r_j$.

Theorem 6 (Non-Vacuity). *Bound* (6) *falls strictly below* 1 *if and only if $\beta_\sigma < \beta^\star(k, \alpha) := \big(1 - \sqrt{\ln(2/\alpha)/(2k)}\big) \big/ \big(1 + \sqrt{2k \ln(2/\alpha)}\big)$. In the cap model this holds for* every *permutation simultaneously whenever*

$$C_\ell\, G\, W(1, k, L)\, 2 \sin r_{\max} \;<\; \beta^\star(k, \alpha). \tag{8}$$

Since $W(1, k, L) \ge L\rho^{L-1}\gamma/(k{+}1)$ with $\rho = 1{+}3B^3R^2$ and $\gamma = 2B^3R^2$, (8) *caps B^3R^2: it constrains the architecture and the source radii, and is* independent of the embedding dimension.

At $k{=}5$, $\alpha{=}0.05$: $\beta^\star = 0.056$, and with $C_\ell G = 1, r_{\max} = 0.6$, non-vacuity at $L{=}2$ requires $B^3R^2 < 0.059$. The theorem names the operative knobs (B^3R^2, $r_{\max}$, L), turning vacuity into a structured triple rather than a black box; an entropy-based tightening (§5.5, Appendix H) substitutes measured attention rows for the worst-case spectral bound and cuts the vacuity gap by a median of 3.4 orders of magnitude on $44/44$ trained models.

Resolvability: coherence certificate.

Lemma 7 (Coherence Certificate). *Fix a target per-pair coherence $c > 4 \sin(r_{\max}/2)$. If*

$$d_x \;\ge\; \frac{2 \ln\big((k{+}1)^2/\alpha_2\big)}{\big(c - 4 \sin(r_{\max}/2)\big)^2}, \tag{9}$$

then with probability $\ge 1 - \alpha_2$ every pair of in-context inputs obeys $|\langle x_j,\, x_l\rangle| \le c$, simultaneously over all samples and all replacements within the caps. Equivalently, the smallest coherence certifiable at dimension d_x is $c_{\min}(d_x, r_{\max}) = \sqrt{2 \ln((k{+}1)^2/\alpha_2)/d_x} + 4 \sin(r_{\max}/2)$, which carries content only when $c_{\min} < 1$.

The proof (Appendix D) uses spherical concentration (Ball, 1997): certification costs $d_x = \Theta(\ln k)$, not $\Theta(k^2)$. At $\alpha_2 = 0.05, k{=}5$: $c_{\min} = 1.74, 0.90, 0.45, 0.23$ for $d_x = 8, 32, 128, 512$—vacuous at 8, informative from 128.

4.5 SOFTMAX EXTENSION

We show structural predictions transfer to softmax attention under a two-regime amplification bound (Proposition 15, Appendix G). At high temperature $\alpha_{k+1,p} \to 1/(k{+}1)$ uniformly in p, so single-layer softmax is position-independent by exactly the linear mechanism—invariance is a mask-acting-through-one-layer effect, not a linearity artefact. Away from that limit, Corollary 16 shows that when attention entropy is within ε_H of the uniform ceiling, position differences are $O(\sqrt{\varepsilon_H})$; qualitative monotonicity is preserved on any near-uniform model. Empirically, trained softmax models that learn ICL (final loss < 0.9) develop $H/H_{\max} = 0.91$–0.98 (mean 0.95; §5.5), placing them in this regime, and their amplification profiles agree with the linear prediction at mean $r_s = 0.57$ ($p = 3 \times 10^{-7}$).

4.6 GREEDY ESTIMATOR FOR SOURCE SENSITIVITIES

A greedy estimator (Algorithm 1, Appendix E) reads η_j from resampled forward-pass differences: at each slot, place the candidate with smallest empirical $\hat{\delta}(j)$. Its cost is $O(k^2 n_r)$ forward passes (126 at $k{=}5, n_r{=}8$), crossing over exhaustive enumeration at $k{=}6$ and running $128\times$ cheaper by $k{=}8$; the data-dependent union bound adds only $O(\sqrt{k \ln k})$ (Remark 1). Because the ordering rule is architectural—anchored in $W(p, k, L)$ and the B^3R^2 criterion—any within-ceiling per-example score (§5.4, learned score at $\geq 92\%$ of $\mathcal{R}^2_{\max}$) instantiates it.

5 EXPERIMENTS

We test the theory's algebraic ($L{=}1$ invariance), structural (monotone amplification, architectural origin, multiplicative decomposition), and quantitative (regime dependence, softmax transfer, baseline comparison) predictions from sharpest to most demanding, then price the induced ordering rule against black-box baselines (Table 2).

5.1 SETUP

Synthetic. Causal linear attention transformers follow Equation (3): single-head linear attention with prefix normalization and residuals. Each token is $(d_x{+}1)$-dimensional, concatenating $x \in \mathbb{S}^{d_x-1}$ and $y = \langle w, x\rangle + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, 0.02^2)$. Default: $k{=}5, d_x{=}32, L{=}3$; ablations sweep $d_x \in \{8, 32, 128, 512\}$, $L \in \{1, 2, 4, 8\}$. Adam ($\eta{=}2{\times}10^{-3}$, batch 256, 6000 steps); task vectors w on $\mathbb{S}^{d_x-1}$ scaled so $\mathrm{Var}(y) = 1$. Heterogeneous spherical-cap sources (§4.4) with target y-spread in $[0.05, 1.5]$; each run gives 10 example sets. The trained network pairs with a matched *untrained* network (Gaussian init, std $0.3/\sqrt{d_x + 1}$) for the control in §5.3. Full details in Appendix F.

Frontier LLMs. Four 7–8B instruction-tuned models (Qwen2.5-7B, Mistral-7B-v0.3, Qwen3-8B, DeepSeek-R1-Distill-Llama-8B) on AG News (4-class) and RTE (SuperGLUE); $k{=}5$, all 120 permutations on each of 10 example sets, 32 queries per permutation (38,400 inferences per cell).

Metrics. Synthetic: gap $|\bar{R} - \hat{R}|$ under CRN, coefficient of variation, split-half Spearman retest. LLMs: accuracy, cross-permutation spread ($\max - \min$ over 120), split-half NLL/acc retest. $L{=}1$ tested against a 10^{-8} precision floor. Effect sizes: Welch's d and TOST at margin ± 0.05 on $|\bar{r}_s|$ (retrospective power > 0.95 at $\Delta_{\text{eq}}{=}0.08$; Appendix F); reported Spearman correlations survive Holm–Bonferroni ($\alpha = 0.05$, $m = 8$). Baselines: Random-single, Random-m-tries for $m \in \{5, 10, 20\}$, exhaustive Oracle, entropy-based selection (Lu et al., 2022), and Greedy (Algorithm 1).

5.2 RESULTS

The sharpest prediction is $W(p, k, 1) \equiv \gamma/(k{+}1)$ (Corollary 2): single-layer causal networks produce zero ordering effects. Observed spread stays below 10^{-8} across configurations (Figure A1b), and bidirectional attention likewise gives $\leq 2.4 \times 10^{-8}$ (Appendix F), isolating the causal mask as the sole architectural source at this floor.

Ordering effects appear once $L \geq 2$. Reordering shifts the gap by median 17.8% (unnormalized) / 2.9% (normalized) of the mean gap (Figure A1a), and split-half retest ($r_s = 0.88$–0.90) confirms the

Table 1: Ordering effects for four 7–8B instruction-tuned models ($k = 5$, all 120 permutations, 10 sets). **Acc spread**: max−min across permutations. **NLL retest**: split-half Spearman of NLL rankings. Spread replicates on both tasks (larger on RTE for three of four); NLL retest is lower on the harder entailment task. †Ceiling effects on Acc retest; Figure A3 resolves every column (AG News).

Model	Acc (%)	Acc spread (%)	NLL retest	Acc retest
AG News (4-class topic classification):				
Qwen2.5-7B	85.5	11.9	0.71	0.08†
Mistral-7B	93.3	8.4	0.31	0.07†
Qwen3-8B	88.3	12.2	0.61	0.07†
DeepSeek-R1-8B	65.3	39.1	0.62	0.68
RTE (2-class entailment):				
Qwen2.5-7B	74.9	15.0	0.10	−0.04
Mistral-7B	72.7	16.9	0.41	0.12
Qwen3-8B	88.2	13.1	0.23	0.14
DeepSeek-R1-8B	56.1	29.4	0.66	0.30
Cross-task reference:				
Pythia-410M (Biderman et al., 2023) (SST-2)	62–78	34–50	0.88	—

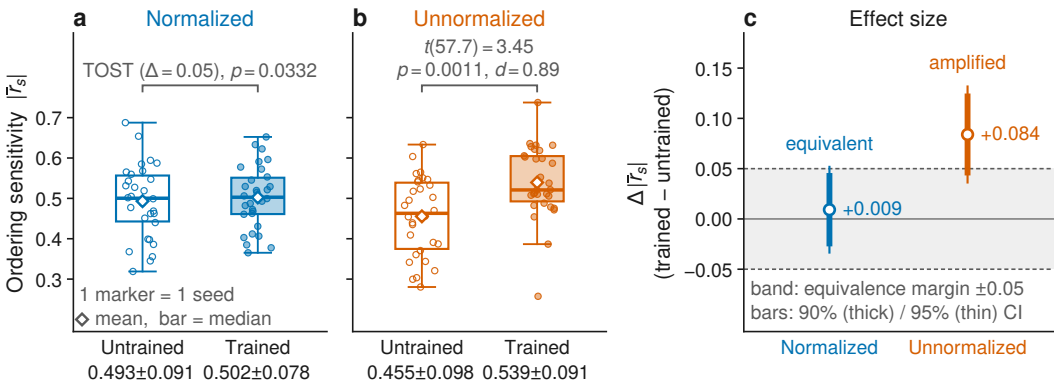


Figure 3: Ordering sensitivity ($|\bar{r}_s|$, n=30 seeds): training leaves sensitivity unchanged under normalized attention but amplifies it under unnormalized. **(a)** Normalized: equivalent (TOST $p = 0.033$). **(b)** Unnormalized: training amplifies ($p = 0.0011$). **(c)** Difference CIs.

permutation ranking is a stable model property. Effects transfer to frontier LLMs: accuracy spread is 8.4–39.1% on AG News and 13.1–29.4% on RTE (Table 1), with NLL retest positive on both (AG News 0.31–0.71, RTE 0.10–0.66).

5.3 Architectural Origin Under Normalized Attention

If ordering effects arise from masking alone, random weights should match trained ones. We test with n=30 seeds (d_x=32, L=3, k=5), powered > 0.95 at $\Delta_{\text{eq}} = 0.08$ (Figure 3).

Trained ($|\bar{r}_s| = 0.502 \pm 0.078$) and untrained ($0.493 \pm 0.091$) networks are equivalent under normalized attention (TOST $p = 0.033$ at margin ± 0.05; Welch $p = 0.675$, $d = 0.11$), with equivalence persisting to $\Delta_{\text{eq}} = 0.05$ (Appendix F). Ordering sensitivity is thus already present at initialization: masking geometry, not learned weights, supplies the architectural floor. Under unnormalized attention that floor is overrun—training amplifies $|\bar{r}_s|$ from 0.455 to 0.539 (Welch $p = 0.0011$, $d = 0.89$)—the residual term the theory anticipates when learned correlations exceed the spectral bound.

Table 2: Ordering methods on the shared query budget (d_x=32, L=3, k=5, normalized, 5 seeds $\times$20 sets), sorted by cost. Greedy's cost is instrumented forward-passes at n_r=8 (Appendix E). A handful of random draws already reaches the top percentiles inside greedy's k=5 budget.

Method	$\bar{R}$	Percentile	Gap capture	Cost (evals)
Random (single)	0.125	50%	0%	1
Random-5-tries	0.116	84%	63%	5
Random-10-tries	0.114	92%	78%	10
Random-20-tries	0.113	96%	87%	20
Oracle (exhaustive)	0.111	100%	100%	120
Greedy (ours)	0.122	60%	21%	126

5.4 Empirical Structure of the Bound

The bound factorizes slot sensitivity as $W(p) \cdot \eta_{\sigma(p)}$: position (architecture) $\times$ example (source). Both are recovered from the empirical 120×5 gap matrix. A rank-1 fit $\alpha_p \cdot v_j$ reaches $r_s = 0.61$–0.86 against ground-truth slot gaps, capturing 70–90% of reproducible variance (test-retest ceiling 0.88–0.95), so the factors are separately recoverable. The fitted α_p then tracks $W(p)$ at $r_s = 0.82$–0.98: "which slot amplifies" is purely architectural.

The framework predicts that the per-example loading v_j carries an irreducible many-body component: Proposition 21 proves v_j is not a function of $Z_j = (x_j, y_j)$ alone, and Lemma 22 quantifies the recoverable share as $\mathcal{R}^2_{\max} = \mathrm{Var}(\mathbb{E}[g(v_j) \mid Z_j])/\mathrm{Var}(g(v_j))$, capping single-example $|r_s|$ at 0.59–0.99 across regimes (Table A2). The prediction is met on both sides: a learned score (ridge / k-NN / random forest; Appendix I) attains 0.92–1.00 of the ceiling at $d_x \leq 32$ and 0.43–0.53 at d_x=128—saturating the recoverable share—while the residual $1 - \mathcal{R}^2_{\max}$ tracks the many-body component the framework isolates as intrinsic to context-dependent ICL (Corollary 23, Appendix I).

5.5 Scaling, Softmax Confirmation, and Baselines

Scaling regime. A 72-run sweep ($k \in \{3, 5\}, L \in \{2, 4\}, d_x \in \{8, 32, 128\}$, both attention variants, 3 seeds) shows ordering effects grow with k (coefficient of variation $1.6\times$ from k=3 to 6 under normalization), peak at d_x=8 (spread $\leq$0.105), and collapse at d_x=128 ($\leq$0.009). A dedicated manipulation (Appendix D) separates two confounded causes: at *fixed* d_x, prescribing per-pair coherence c nearly doubles spread at an interior optimum ($1.9\times$ at c=0.6–0.8, $p < 10^{-10}$) and pushes it below baseline at c=0.95—a coherence *window*, not monotone overlap. The d_x collapse is thus driven by ICL signal loss (in-context learnable variance $0.62 \to 0.04$), not coherence alone (Table A9).

Softmax transfer. Across 48 softmax models, all eight L=1 models show spread $\leq 1.4 \times 10^{-9}$: single-layer invariance is not a linearity artefact. Multi-layer models that learn ICL (final loss < 0.9; $24/40$) develop high-entropy attention (mean $H/H_{\max} = 0.95$)—the near-uniform regime of Corollary 16. Amplification profiles match the linear form at mean $r_s = 0.57$ ($p = 3 \times 10^{-7}$; Figure A4), and attention entropy predicts ordering magnitude at $r_s = 0.85$ ($p = 5 \times 10^{-12}$; partial $r_s = 0.69$ controlling depth).

Cost-adjusted comparison. Table 2 prices strategies in one unit (one *eval* = one ordering on the shared query budget). The rearrangement rule certifies an ordering *a priori* from architecture (B^3R^2; Theorem 6) and geometry ($d_x, r_{\max}$; Lemma 7)—with the entropy-based tightening cutting vacuity by a median of 3.4 orders of magnitude on $44/44$ trained models—crossing over exhaustive enumeration at k=6 and running $128\times$ cheaper by k=8. At k=5 random-m-tries is competitive within a handful of evaluations, matching Proposition 21: because η_j is only a strict lower bound on v_j, the full separation (7) materializes when a within-ceiling per-example score complements the rule—the learned score of §5.4 supplies one, saturating $\mathcal{R}^2_{\max}$ at $d_x \leq 32$.

6 CONCLUSION

Causal masking creates a position-dependent amplification gradient, and β_σ turns that gradient into the first ICL generalization bound in which the permutation appears explicitly. Three theorems make it operational—low-sensitivity-first ordering (Theorem 5), a dimension-free non-vacuity criterion in $(B^3R^2, r_{\max}, L)$ (Theorem 6), and source resolvability at cost $d_x = \Theta(\ln k)$ (Lemma 7)—and a hierarchy of experiments confirms the mechanism: $L{=}1$ invariance to numerical precision, TOST equivalence of trained and random $L \geq 2$ models ($p = 0.033$), $W(p)$ tracked at $r_s = 0.82$–0.98, near-uniform-entropy softmax models that reproduce the amplification ($r_s = 0.57$) and entropy–spread coupling ($r_s = 0.85$), and 8–39% accuracy spread across four frontier LLMs on AG News and RTE. The measurable ceiling $\mathcal{R}^2_{\max} = 0.35$–$0.98$ (Lemma 22) identifies context-dependent many-body structure as the next lever, opening low-entropy amplification bounds and multi-head chain counting for future work.

AI USE STATEMENT

Generative AI assistants (large language models) were used in three capacities while preparing this work. First, for prose polishing and LaTeX formatting assistance. Second, for retrieval and discovery, including surveying and locating related work. Third, AI assistants contributed to deriving and proving the paper's mathematical claims—the closed-form position-dependent amplification factor, the sequential stability coefficients, the order-dependent generalization and ordering-certification bounds, and their supporting proofs in the appendix. The authors independently verified every derivation and proof step, and confirmed the resulting predictions against the diagnostic experiments and the four instruction-tuned LLMs reported in the paper; the experimental design and analysis of results are the authors' own. The authors take full responsibility for the final content of this work, including any material produced with the aid of generative AI.

ETHICS STATEMENT

All experiments use synthetic data or publicly available pretrained models and benchmark datasets; we collect no new human-subject data and use no private or personally identifiable information. The study evaluates prompt-order sensitivity in controlled benchmarks and does not deploy models or make decisions about individuals. Because order-dependent behavior may reduce reliability in downstream applications, practitioners should validate prompt ordering and model behavior in the intended deployment context.

REPRODUCIBILITY STATEMENT

Architectures, hyperparameters, and training are specified in Appendix F; proofs are self-contained in the appendix. Synthetic experiments follow (3) exactly under common random numbers. LLM experiments use publicly available models (Qwen2.5-7B-Instruct, Mistral-7B-Instruct-v0.3, Qwen3-8B, DeepSeek-R1-Distill-Llama-8B) with greedy decoding on AG News and RTE (SuperGLUE); anonymous supplementary material contains full code and a reproducibility guide.

REFERENCES

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? Investigations with linear models. In *International Conference on Learning Representations*, 2023.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *Advances in Neural Information Processing Systems*, volume 36, pp. 57125–57211, 2023.

Keith Ball. An elementary introduction to modern convex geometry. *Flavors of Geometry, MSRI Publications*, 31:1–58, 1997.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *International Conference on Machine Learning*, pp. 41–48, 2009.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purber, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

Stephanie C.Y. Chan, Adam Santoro, Andrew K. Lampinen, James C.R. Whittington, Andrea Banino, and Timothy Lillicrap. Data distributional properties drive emergent in-context learning in transformers. In *Advances in Neural Information Processing Systems*, volume 35, pp. 18878–18891, 2022.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? A case study of simple function classes. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30583–30598, 2022.

Godfrey H Hardy, John E Littlewood, and George Pólya. *Inequalities*. Cambridge University Press, 2nd edition, 1952.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9318–9333, 2023.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114 (13):3521–3526, 2017.

Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context learning. In *International Conference on Machine Learning*, pp. 19565–19594, 2023.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8086–8098, 2022.

Colin McDiarmid. On the method of bounded differences. In J. Siemons (ed.), *Surveys in Combinatorics*, volume 141 of *London Mathematical Society Lecture Note Series*, pp. 148–188. Cambridge University Press, 1989.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Conference on Empirical Methods in Natural Language Processing*, 2022.

Konstantin Mishchenko, Ahmed Khaled, and Peter Richtárik. Random reshuffling: Simple analysis with vast improvements. In *Advances in Neural Information Processing Systems*, 2020.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022.

Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-Bayesian in-context learning for regression. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information. In *Conference on Learning Theory*, pp. 3437–3452. PMLR, 2020.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pp. 35151–35174. PMLR, 2023.

Noam Wies, Yoav Levine, and Amnon Shashua. The learnability of in-context learning. In *Advances in Neural Information Processing Systems*, 2023.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit Bayesian inference. In *International Conference on Learning Representations*, 2022.

Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25:1–55, 2024.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pp. 12697–12706. PMLR, 2021.

A Proof of Proposition 1 (Amplification Factor)

The proof proceeds in three steps: a single-layer perturbation recursion, its matrix formulation, and extraction of the $(k{+}1, p)$ entry via nilpotent expansion.

Lemma 8 (Single-Layer Perturbation Propagation)**.** *Let $\varepsilon_t^{(l)} := \|h_t^{(l)} - \tilde{h}_t^{(l)}\|$ denote the perturbation at position t after layer l. Under the assumptions of §3:*

$$\varepsilon_t^{(l+1)} \;\le\; \rho\,\varepsilon_t^{(l)} + \frac{\gamma}{t}\sum_{s=1}^{t-1}\varepsilon_s^{(l)}. \tag{10}$$

Proof. Write $f_s(h_t, h_s) := \langle W_Q h_t,\, W_K h_s\rangle\, W_V h_s$. We decompose the per-layer update (3) into source terms ($s < t$, perturbation enters through h_s) and self terms ($s = t$ plus residual).

Source terms ($s < t$). For $\|h\|, \|h'\| \le R$:

$$\begin{aligned}&\|f_s(h_t, h') - f_s(h_t, h)\| \\ &\le\; BR\cdot B\|h'{-}h\|\cdot BR + BR\cdot BR\cdot B\|h'{-}h\| = \gamma\|h'{-}h\|.\end{aligned} \tag{11}$$

After the $1/t$ prefix normalization, position s contributes $\le (\gamma/t)\,\varepsilon_s^{(l)}$.

Self terms. The residual contributes $\varepsilon_t^{(l)}$. The query-path perturbation (changing h_t in all attention scores):

$$\frac{1}{t}\sum_{s\le t}\big|\langle W_Q(h_t{-}\tilde{h}_t),\, W_K h_s\rangle\big|\cdot\|W_V h_s\| \;\le\; B^3R^2\,\varepsilon_t^{(l)}.$$

The $s{=}t$ source term contributes $\le 2B^3R^2\,\varepsilon_t^{(l)}$. Total self: $(1 + 3B^3R^2)\,\varepsilon_t^{(l)} = \rho\,\varepsilon_t^{(l)}$.

Matrix formulation. Define $\boldsymbol{\varepsilon}^{(l)} \in \mathbb{R}^{k+1}$ componentwise. Lemma 8 gives $\boldsymbol{\varepsilon}^{(l+1)} \le M\,\boldsymbol{\varepsilon}^{(l)}$ (componentwise), where

$$M = \rho I + \gamma N, \qquad N_{ts} = \begin{cases} 1/t, & s < t, \\ 0, & s \ge t, \end{cases} \tag{12}$$

and N is strictly lower triangular, hence nilpotent: $N^{k+1} = 0$. The initial perturbation is $\boldsymbol{\varepsilon}^{(0)} = D\,e_p$ (unit vector at position p, scaled by input perturbation D).

Proof of Proposition 1. Since M has nonneg entries and the inequality is componentwise, iterating yields $\boldsymbol{\varepsilon}^{(L)} \le M^L \boldsymbol{\varepsilon}^{(0)}$, so $w(p,k,L) \le [M^L]_{k+1,\,p} =: W(p,k,L)$.

The identity I commutes with any matrix, so the binomial theorem gives:

$$M^L = (\rho I + \gamma N)^L = \sum_{m=0}^{\min(L,\,k)} \binom{L}{m} \rho^{L-m}\, \gamma^m\, N^m. \tag{13}$$

The $m{=}0$ term contributes $\rho^L\, \delta_{k+1,p} = 0$ for $p \le k$. For $m \ge 1$, the $(k{+}1, p)$ entry of N^m enumerates all strictly increasing causal chains $p < t_1 < \cdots < t_{m-1} \le k < k{+}1$:

$$[N^m]_{k+1,\,p} = \frac{1}{k{+}1} \sum_{p<t_1<\cdots<t_{m-1}\le k} \prod_{j=1}^{m-1} \frac{1}{t_j} = \frac{e_{m-1}\big(\frac{1}{p+1}, \ldots, \frac{1}{k}\big)}{k{+}1}, \tag{14}$$

where the bijection between increasing chains and $(m{-}1)$-element subsets of $\{p{+}1, \ldots, k\}$ identifies the sum with the elementary symmetric polynomial. The chain length is bounded by intermediate positions: $m - 1 \le k - p$. Substituting yields (4).

Proof of Corollary 2. For $L \ge 2$ and $p < k$, the variable set $\big\{\frac{1}{p+1}, \ldots, \frac{1}{k}\big\}$ strictly contains $\big\{\frac{1}{p+2}, \ldots, \frac{1}{k}\big\}$. Since all entries are positive and $L \ge 2$ ensures the $m{=}2$ term is present (with e_1 over a larger set), each such term is strictly larger for position p than for $p{+}1$.

Proof of Corollary 2. For $L = 1$, only $m{=}1$ contributes ($\binom{1}{m} = 0$ for $m \ge 2$): $W(p,k,1) = \gamma \cdot e_0/(k{+}1) = \gamma/(k{+}1)$, independent of p.

Proof of Corollary 2. The ratio $W(1,k,L)/W(k,k,L)$ involves higher-order terms ($m \ge 2$) in the numerator that are absent from the denominator ($W(k,k,L)$ has only $m{=}1$). Each such term carries $\binom{L}{m}$, which is strictly increasing in L for $m \ge 2$, while the denominator grows as $L\rho^{L-1}$. The ratio is thus strictly increasing in L for $L \ge 2$.

Remark 2 (Validity of the perturbation recursion)**.** *The recursion in Lemma 8 is a worst-case* norm bound *via the triangle inequality, not a first-order Taylor expansion; consequently, it carries no linearization remainder. The bound holds for any perturbation magnitude, provided the norm invariant $\|h_t^{(l)}\| \le R$ is maintained at all layers. Since the initial perturbation satisfies $\varepsilon_p^{(0)} \le BR$ (bounded embedding) and the recursion preserves $\|h_t^{(l)}\| \le R$ by the contractivity assumed in §3, the bound is valid without a small-ε regime restriction. This validity statement covers the linear-attention recursion only. For the softmax extension (Proposition 17) the high-entropy approximation contributes a separate error, and it is* not *controlled at our operating point: by Corollary 16 the per-layer deviation from uniform attention is $O(\sqrt{\varepsilon_H})$—not $O(\varepsilon_H/k)$, since Pinsker's inequality converts an entropy deficit into an ℓ_1 deviation at rate $\sqrt{\cdot}$ and the $1/k$ normalization cancels against the k summands—so the compounded error is $O(L\sqrt{\varepsilon_H})$. Across the 24 multi-layer softmax models that learn ICL (§4.5) the mean entropy deficit is $\varepsilon_H = 0.094$ (worst layer within a model: 0.158), giving $L\sqrt{\varepsilon_H/2} \approx 0.85$ at $L{=}4$: order one, not negligible. The worst-case remainder therefore does not certify Proposition 17; its conclusion rests on the direct measurement reported in §4.5, and the requirement $\varepsilon_H \ll 1/L^2$ that would make the remainder small is met by none of our models.*

B Proof of Theorem 4 (Main Theorem)

The proof uses three lemmas: a slot-level sensitivity bound, a bias bound, and bounded differences for McDiarmid's inequality.

Lemma 9 (Slot Replacement Sensitivity)**.** *Fix a query (x, y) and a permutation σ. Replacing the example at slot p from z to z' (with $\|\phi(z) - \phi(z')\| \le \eta_{\sigma(p)}$, same source), keeping all other examples fixed:*

$$\big|\ell\big(f_\theta(P_\sigma(S); x), y\big) - \ell\big(f_\theta(P_\sigma(S^{p\leftarrow z'}); x), y\big)\big| \;\le\; \delta_p(\sigma). \tag{15}$$

Proof. Embedding perturbation $\leq \eta_{\sigma(p)} \xrightarrow{\text{Proposition 1}}$ query representation perturbation $\leq W(p,k,L)\cdot \eta_{\sigma(p)} \xrightarrow{G\text{-Lip}}$ prediction perturbation $\leq G\,W(p)\,\eta_{\sigma(p)} \xrightarrow{C_\ell\text{-Lip}}$ loss perturbation $\leq C_\ell\,G\,W(p)\,\eta_{\sigma(p)} = \delta_p(\sigma)$.

Lemma 10 (Bias Bound)**.** $\big|\mathbb{E}[\bar{R}-\hat{R}]\big| \leq \bar{\beta}_\sigma := \frac{1}{k}\sum_{p=1}^{k}\delta_p(\sigma) \leq \beta_\sigma$.

Proof. For each j, let $z'_j \sim \mathcal{D}_j$ be an independent copy (a "ghost sample" drawn from the *same marginal* $\mathcal{D}_j$, independently of z_j and all other samples; this is valid for product measures over heterogeneous marginals $\mathcal{D}_1 \times \cdots \times \mathcal{D}_k$ without requiring identical distributions). By exchangeability of (z_j, z'_j) (same distribution, independent): $\mathbb{E}\,\ell(f_\theta(P_\sigma(S); x'_j), y'_j) = \mathbb{E}\,\ell(f_\theta(P_\sigma(S^{j\leftarrow z'_j}); x_j), y_j)$. The right side differs from the j-th term of $\hat{R}$ by a single slot replacement at $p_j := \sigma^{-1}(j)$, bounded by $\delta_{p_j}(\sigma)$ via Lemma 9. Averaging over j gives $\bar{\beta}_\sigma$; the inequality $\bar{\beta}_\sigma \leq \beta_\sigma$ follows from arithmetic mean $\leq$ quadratic mean.

Lemma 11 (Bounded Differences)**.** *Let $\Phi(z_1,\ldots,z_k) := \bar{R} - \hat{R}$. Replacing any single z_j (at slot $p_j = \sigma^{-1}(j)$): $|\Phi(S) - \Phi(S^{j\leftarrow z'_j})| \leq c_{p_j} := 2\,\delta_{p_j}(\sigma) + 1/k$.*

Proof. *$\bar{R}$ change:* All k terms share the same prompt, differing at slot p_j. Each term changes by $\leq \delta_{p_j}$ (Lemma 9), so $|\Delta\bar{R}| \leq \delta_{p_j}$.

$\hat{R}$ change: For $m \neq j$: fixed query, one prompt slot changed, giving $\leq \delta_{p_j}$ per term, contributing $(k{-}1)\delta_{p_j}/k$ total. For $m = j$: the query itself is replaced; $\ell \in [0,1]$ gives $\leq 1/k$. Total: $|\Delta\hat{R}| \leq (k{-}1)\delta_{p_j}/k + 1/k$.

Combining: $|\Delta\Phi| \leq \delta_{p_j} + (k{-}1)\delta_{p_j}/k + 1/k \leq 2\delta_{p_j} + 1/k$.

Proof of Theorem 4. Since $z_1,\ldots,z_k$ are independent and Φ satisfies the bounded differences of Lemma 11, McDiarmid's inequality (McDiarmid, 1989) gives:

$$\mathbb{P}\big(|\Phi - \mathbb{E}\Phi| \geq \varepsilon\big) \leq 2\exp\Big(-\frac{2\varepsilon^2}{\sum_{p=1}^{k} c_p^2}\Big). \tag{16}$$

Bound the denominator via Minkowski's inequality (ℓ^2 triangle inequality):

$$\Big(\sum_p c_p^2\Big)^{1/2} \leq 2\Big(\sum_p \delta_p^2\Big)^{1/2} + \Big(\sum_p \frac{1}{k^2}\Big)^{1/2} = 2\sqrt{k}\,\beta_\sigma + \frac{1}{\sqrt{k}}. \tag{17}$$

Setting the tail probability to α and solving: $\varepsilon = (2\sqrt{k}\,\beta_\sigma + 1/\sqrt{k})\sqrt{\ln(2/\alpha)/2} = (2k\,\beta_\sigma + 1)\sqrt{\ln(2/\alpha)/(2k)}$. Combining with Lemma 10: $|\Phi| \leq |\mathbb{E}\Phi| + |\Phi - \mathbb{E}\Phi| \leq \beta_\sigma + (2k\,\beta_\sigma + 1)\sqrt{\ln(2/\alpha)/(2k)}$.

Proof of Remark 1. Apply Theorem 4 with confidence $\alpha/k!$ to each fixed σ, then union bound over all $k!$ permutations. The additional $\ln k!$ cost satisfies $\ln(2k!/\alpha) \leq k\ln k + \ln(2/\alpha)$.

C Proof of Theorem 5 (Separation)

Proof of Theorem 5. Non-degeneracy condition. The separation result requires $W(p)$ to be strictly decreasing, which by Corollary 2 holds whenever $L \geq 2$ and $\gamma > 0$ (the cross-token coupling constant from the attention mechanism is positive). When $\gamma = 0$—which occurs only if the off-diagonal attention weights are exactly zero—$W(p) = \rho^L$ is constant and all orderings yield identical bounds. For any trained or randomly initialized model with non-degenerate attention, $\gamma > 0$ holds generically.

Part (i): Optimal ordering. Write $\beta_\sigma^2 = \frac{(C_\ell G)^2}{k}\sum_{p=1}^{k} a_p\, b_{\sigma(p)}$, where $a_p := W(p)^2$ (strictly decreasing by Corollary 2) and $b_j := \eta_j^2$.

Suppose σ minimizes $\sum_p a_p b_{\sigma(p)}$ but there exist $p < q$ with $b_{\sigma(p)} > b_{\sigma(q)}$. Swapping the assignments at p and q changes the objective by $(a_p - a_q)(b_{\sigma(q)} - b_{\sigma(p)}) < 0$, contradicting minimality. Hence at the minimum, $\eta_{\sigma(1)} \le \cdots \le \eta_{\sigma(k)}$ (ascending order). This is the rearrangement inequality (Hardy et al., 1952). The maximum is attained by the reverse ordering.

Part (ii): Relative separation. Let π denote the ascending (best) ordering. Construct π' by swapping only $\eta_{\max}$ and $\eta_{\min}$ in π (placing $\eta_{\max}$ at slot 1, $\eta_{\min}$ at slot k). Then:

$$\textstyle\sum_p a_p b_{\pi'(p)} - \sum_p a_p b_{\pi(p)} = (a_1 - a_k)(b_{\max} - b_{\min}). \tag{18}$$

Since π' is not the worst ordering in general, $\beta^2_{\text{worst}} \ge (C_\ell G)^2 \sum_p a_p b_{\pi'(p)}/k$, giving:

$$\beta^2_{\text{worst}} - \beta^2_{\text{best}} \;\ge\; \frac{(C_\ell G)^2}{k}(a_1 - a_k)(b_{\max} - b_{\min}). \tag{19}$$

Using $b_{\max} - b_{\min} = (\eta_{\max} + \eta_{\min})(\eta_{\max} - \eta_{\min}) \ge \Delta^2$ and the trivial bound $\beta^2_{\text{worst}} \le (C_\ell G)^2 a_1 C^2$:

$$\frac{\beta^2_{\text{worst}} - \beta^2_{\text{best}}}{\beta^2_{\text{worst}}} \ge \frac{(a_1 - a_k)\Delta^2/k}{a_1 C^2} = \Big(1 - \frac{W(k)^2}{W(1)^2}\Big)\frac{\Delta^2}{kC^2} = c_{L,k}\,\frac{\Delta^2}{kC^2}. \tag{20}$$

D Proofs for §4.4 (Non-Vacuity and the Coherence Certificate)

D.1 Proof of Theorem 6

Proof of Theorem 6. Step 1 (the exact threshold in β_σ). Write $s := \sqrt{\ln(2/\alpha)/(2k)}$. The right-hand side of (6) is $\beta_\sigma + (2k\beta_\sigma + 1)s$, which is affine and strictly increasing in β_σ. Hence it is < 1 if and only if $\beta_\sigma(1 + 2ks) < 1 - s$. Since $2ks = \sqrt{4k^2 \ln(2/\alpha)/(2k)} = \sqrt{2k \ln(2/\alpha)}$, this is exactly $\beta_\sigma < \beta^\star(k, \alpha)$ as defined. (The criterion is an equivalence, not a sufficient condition: it is tight in β_σ.)

Step 2 (uniform control of β_σ in the cap model). Two points $x, x' \in \text{cap}(m_j, r_j)$ satisfy $\angle(x, x') \le 2r_j$, so $\|x - x'\| = 2\sin\big(\angle(x, x')/2\big) \le 2\sin r_j$ for $r_j \le \pi/2$. With a unit-Lipschitz embedding this gives $\eta_j \le 2\sin r_j \le 2\sin r_{\max}$ deterministically—no probabilistic step and no sample dependence. By Corollary 2, $W(p, k, L) \le W(1, k, L)$ for all p, so for *every* permutation σ,

$$\beta_\sigma = C_\ell G\Big(\tfrac{1}{k}\textstyle\sum_p W(p)^2 \eta^2_{\sigma(p)}\Big)^{1/2} \;\le\; C_\ell\, G\, W(1, k, L)\, 2\sin r_{\max}. \tag{21}$$

Combining with Step 1 gives (8).

Step 3 (the condition caps B^3R^2). Every term of (4) is nonnegative, and the $m{=}1$ term is $\binom{L}{1}\rho^{L-1}\gamma^1 e_0 = L\rho^{L-1}\gamma$, so $W(1, k, L) \ge L\rho^{L-1}\gamma/(k{+}1)$. Both $\rho = 1{+}3B^3R^2$ and $\gamma = 2B^3R^2$ increase with B^3R^2, hence so does the left-hand side of (8); the condition therefore defines an upper bound on B^3R^2 given $(k, L, \alpha, r_{\max}, C_\ell G)$, with no reference to d_x.

Numerical instantiation. At $k{=}5$, $\alpha{=}0.05$: $s = 0.607$ and $\beta^\star = 0.393/7.074 = 0.0555$; substituting $\beta_\sigma = \beta^\star$ back into (6) returns exactly 1. Solving (8) for B^3R^2 with $C_\ell G = 1$ gives the ceilings in Table A1. Inverting the same relation, the reported $\beta_\sigma \sim 10^4$ corresponds to $B^3R^2 \approx 56$ at $L{=}2$, $r_{\max} = 0.6$—a factor $\approx 9 \times 10^2$ above the ceiling. Note that the ceiling *tightens* with depth (deeper models amplify more) and loosens as source caps shrink.

Table A1: Largest B^3R^2 admitting a non-trivial bound ((8), $k{=}5$, $\alpha{=}0.05$, $C_\ell G = 1$). Computed by `scripts/analyze_nonvacuity_certificate.py`.

$r_{\max}$	$L{=}1$	$L{=}2$	$L{=}4$
0.600 (main experiment)	0.148	0.059	0.027
0.066 (sweep, $d_x{=}128$)	1.256	0.284	0.105

D.2 Proof of Lemma 7

The certificate is logically separate from Theorem 4: in the cap model the bounded-difference constants are already deterministic (Step 2 above), so Theorem 4 applies as stated, without correction. What the certificate adds is control on *cross-source* interference—the premise under which distinct in-context examples are resolvable—which is what the d_x sweep of §5.5 varies. The proof decouples a probabilistic step (source centers are nearly orthogonal in high dimensions) from a deterministic one (given well-separated centers, all sample-level inner products are bounded).

Lemma 12 (Center Coherence Concentration)**.** *Let $m_1, \dots, m_k, m_q \in \mathbb{S}^{d_x-1}$ be independent uniform. For $\alpha_2 \in (0,1)$, with probability $\geq 1-\alpha_2$:*

$$\tau_c := \max_{\substack{i,j\in\{1,\dots,k,q\}\\ i\neq j}} |\langle m_i,\, m_j\rangle| \;\leq\; \sqrt{\frac{2\ln\big((k+1)^2/\alpha_2\big)}{d_x}}. \tag{22}$$

Proof. For a fixed pair of independent uniform vectors on $\mathbb{S}^{d_x-1}$, concentration on the sphere (Ball, 1997) gives

$$\mathbb{P}(|\langle m_i,\, m_j\rangle| \geq \tau) \leq 2e^{-d_x\tau^2/2}.$$

Union bound over $\binom{k+1}{2} \leq (k+1)^2/2$ pairs and solving for τ yields the claim.

Lemma 13 (Deterministic Center-to-Sample Propagation)**.** *Suppose $\tau_c \leq \varepsilon_c$. Then for* all $x_j \in \mathrm{cap}(m_j, r_j)$ *and $x_l \in \mathrm{cap}(m_l, r_l)$ with $j \neq l$:*

$$|\langle x_j,\, x_l\rangle| \;\leq\; \varepsilon_c + 4\sin\tfrac{r_{\max}}{2} \;=:\; \tau_{\text{eff}}. \tag{23}$$

This bound is deterministic over all samples and replacements within each source's support.

Proof. For $x_j \in \mathrm{cap}(m_j, r_j)$: $\|x_j - m_j\| = 2\sin(\theta_j/2) \leq 2\sin(r_j/2)$, where $\theta_j = \angle(x_j, m_j) \leq r_j$. Then:

$$\begin{aligned}
|\langle x_j,\, x_l\rangle| &\leq |\langle m_j,\, x_l\rangle| + \|x_j - m_j\| \\
&\leq |\langle m_j,\, m_l\rangle| + \|x_l - m_l\| + 2\sin\tfrac{r_j}{2} \\
&\leq \varepsilon_c + 2\sin\tfrac{r_l}{2} + 2\sin\tfrac{r_j}{2} \leq \varepsilon_c + 4\sin\tfrac{r_{\max}}{2}.
\end{aligned}$$

The key consequence: any replacement $x'_p \in \mathrm{cap}(m_p, r_p)$ shares center m_p with the original x_p, so Lemma 13 covers *all* replacements simultaneously—not merely a single x'_p—and the resulting inner-product bound is a function of centers and radii alone, not of the sample realization.

Proof of Lemma 7. Step 1 (probabilistic: center concentration). Set $\varepsilon_c := c - 4\sin(r_{\max}/2)$, positive by the hypothesis $c > 4\sin(r_{\max}/2)$. Condition (9) states $d_x \geq 2\ln\big((k+1)^2/\alpha_2\big)/\varepsilon_c^2$, i.e. $\sqrt{2\ln((k+1)^2/\alpha_2)/d_x} \leq \varepsilon_c$, so Lemma 12 gives $\tau_c \leq \varepsilon_c$ with probability $\geq 1-\alpha_2$.

Step 2 (deterministic: propagation to samples). On that event, Lemma 13 with $\varepsilon_c = c - 4\sin(r_{\max}/2)$ yields, for all $j \neq l$ and all $x_j \in \mathrm{cap}(m_j, r_j)$, $x_l \in \mathrm{cap}(m_l, r_l)$,

$$|\langle x_j,\, x_l\rangle| \;\leq\; \varepsilon_c + 4\sin\tfrac{r_{\max}}{2} \;=\; c.$$

Step 3 (inversion). Solving $\sqrt{2\ln((k+1)^2/\alpha_2)/d_x} \leq c - 4\sin(r_{\max}/2)$ for c at fixed d_x gives the smallest certifiable coherence $c_{\min}(d_x, r_{\max}) = \sqrt{2\ln((k+1)^2/\alpha_2)/d_x} + 4\sin(r_{\max}/2)$. Since $|\langle x_j,\, x_l\rangle| \leq 1$ holds unconditionally for unit vectors, the certificate carries content only when $c_{\min} < 1$; note that $4\sin(r_{\max}/2)$ is a floor that no dimension can lower, so $r_{\max} \geq 2\arcsin(1/4) = 0.505$ makes the certificate vacuous at every d_x.

The k^2 cancellation. Earlier drafts stated (9) in terms of an aggregate coherence budget ε_0 obeying $|\langle x_j,\, x_l\rangle| \leq \varepsilon_0/k$, giving $d_x \geq 2k^2\ln\big((k+1)^2/\alpha_2\big)/\big(\varepsilon_0 - 4k\sin(r_{\max}/2)\big)^2$. Substituting $\varepsilon_0 = kc$ factors k^2 out of the denominator, $\big(kc - 4k\sin(r_{\max}/2)\big)^2 = k^2\big(c - 4\sin(r_{\max}/2)\big)^2$, and it cancels the k^2 in the numerator exactly, recovering (9). The two forms are the same statement; only the second makes visible that the dimension requirement is $\Theta(\ln k)$ at fixed per-pair coherence. A k^2 scaling re-enters only if ε_0—not c—is held fixed as k grows.

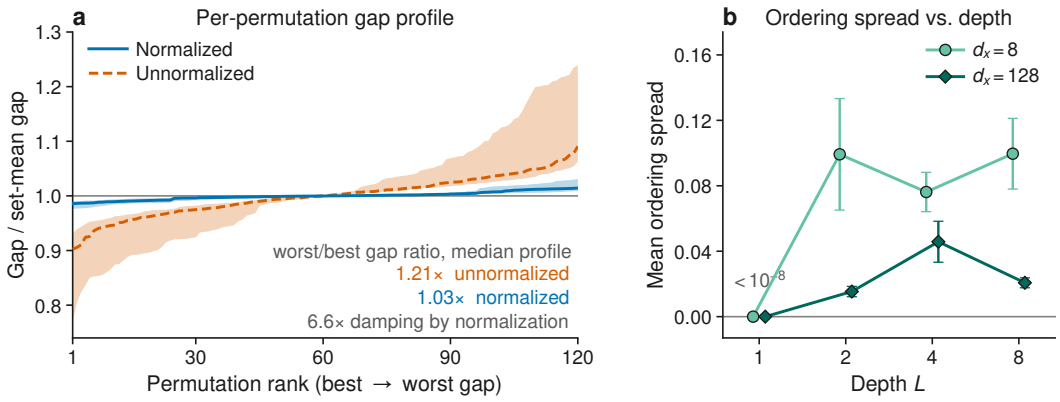


Figure A1: **(a)** Generalization gap over 120 permutations, normalized by mean, sorted best to worst (d_x=128, L=4; median of 50 sets, shaded bootstrap 95% CI of the median): normalization sharply damps ordering effects. **(b)** Mean ordering spread vs. depth, unnormalized attention (n=5 seeds, ±s.e.m.). L=1: zero to numerical precision ($< 10^{-8}$; Corollary 2). $L \geq 2$: nonzero but non-monotonic. Referenced from §5.2.

E Greedy Ordering Algorithm

Since the theoretical scores η_j are unobservable, we propose a greedy heuristic that estimates them via forward passes. The cost is $O(k^2 n_r)$ forward passes—exactly $[k(k+1)/2 - 1](1 + n_r)$ prompt

Algorithm 1 Greedy sensitivity-based ordering

Require: examples $\{z_1, \ldots, z_k\}$, model f_θ, resampling count n_r
chosen $\leftarrow [\,]$; remaining $\leftarrow \{1, \ldots, k\}$
for $p = 1, \ldots, k$ **do**
  for $j \in$ remaining **do**
    Construct trial prompt: chosen $+\!+\, [j] +\!+$ (remaining $\setminus \{j\}$)
    Estimate $\hat{\delta}(j) \leftarrow \max_{n_r \text{ resamples}} |\Delta\text{loss}|$
  end for
  $j^* \leftarrow \arg\min_j \hat{\delta}(j)$; append j^* to chosen; remove j^* from remaining
end for
return chosen

batches, i.e. 126 ordering-evaluation equivalents at k=5, n_r=8, which exceeds the 120 needed to enumerate every permutation (Appendix J). The estimator takes the max over resamples of the mean absolute loss change, which upper-bounds the corresponding bounded-difference constants (mean absolute change $\geq$ absolute mean change), making empirical β_σ values conservative. Since β_σ is not decomposable across slots, greedy selection carries no approximation guarantee; when ordering depends on data, a union bound applies at multiplicative cost $O(\sqrt{k \ln k})$.

F Additional Experimental Details

Synthetic model architecture. The causal linear attention transformer follows (3) exactly: single-head linear attention with prefix normalization ($1/t$) and residual connections. Each token is a $(d_x{+}1)$-dimensional vector concatenating input $x \in \mathbb{R}^{d_x}$ (on $\mathbb{S}^{d_x-1}$) and scalar output $y = \langle w, x \rangle + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, 0.02^2)$. Default configuration: $k = 5$, $d_x = 32$ (token dimension $d_x + 1 = 33$), $L = 3$ layers. Ablations use $d_x \in \{8, 32, 128\}$ and $L \in \{1, 2, 4, 8\}$.

Training. Adam optimizer, learning rate 2×10^{-3}, batch size 256, 6000 gradient steps, weight initialization scale 0.3. For the L=8, d_x=128 configuration, the learning rate is reduced to 5×10^{-4}

for training stability. The 72-run parameter sweep (Table A9) uses learning rate 10^{-3} for $L \geq 4$ configurations, and approximates the bound constants with B set to the maximum absolute weight and $R = 1$; since ρ, γ are shared across all permutations of a given model, these choices rescale β_σ uniformly within each run and leave the reported Spearman correlations unchanged. Training task: predict the query y_{k+1} given k demonstrations, MSE loss. Task vectors w are sampled i.i.d. on $\mathbb{S}^{d_x-1}$ (scaled to $\|w\| = \sqrt{d_x}$ so $\mathrm{Var}(y) = 1$). The zero predictor achieves MSE ≈ 1.0 (since $\mathrm{Var}(y) = 1$); trained models at d_x=128, L=4 reach loss $\sim$1.05, indicating modest learning beyond the trivial baseline.

Heterogeneous evaluation sources. Evaluation uses spherical cap sources (§4.4) with centers sampled uniformly on $\mathbb{S}^{d_x-1}$ and angular radii determined by target y-spread values in $[0.05, 1.5]$ (converted via $r = \arcsin(\text{spread}/(2\sqrt{d_x}))$, yielding angular radii $\approx [0.002, 0.07]$ at d_x=128). Each evaluation run generates 10 independent example sets, each consisting of k examples from k distinct sources.

CRN evaluation protocol. All permutation comparisons use common random numbers (CRN): the same random seeds generate example sets and test queries across all permutations, isolating the ordering effect from sampling noise. For $k = 5$, all 120 permutations are evaluated; for $k = 6$, 200 permutations are sampled. Population risk $\bar{R}$ is estimated with $n_{\text{pop}} = 384$ Monte Carlo samples per source; empirical sensitivities use $n_r = 16$ resamples (12 in the 72-run sweep).

Trained vs. untrained comparison. Both conditions use $d_x = 32$, $L = 3$, $k = 5$, with $n = 30$ independent seeds per condition (trained/untrained $\times$ normalized/unnormalized). Untrained models use the same architecture with randomly initialized weights (Gaussian, std $= 0.3/\sqrt{d_x+1}$, i.e., the same initialization distribution as trained models at step 0). The TOST equivalence test uses margin ± 0.05 (10% of the $[0, 1]$ scale) on $|\bar{r}_s|$; equivalence holds for all $\Delta_{\text{eq}} \geq 0.05$ and fails at $\Delta_{\text{eq}} = 0.04$, with the 90% CI at $[-0.027, +0.046]$ and retrospective power exceeding 0.95 at $\Delta_{\text{eq}} = 0.08$. We report mean *absolute* per-set Spearman correlations $|\bar{r}_s|$: per-set correlations frequently change sign, and the corresponding signed means are $+0.28$ (trained) vs. -0.05 (untrained) under normalization and $+0.28$ vs. $+0.01$ under unnormalization, so the absolute value is used as the effect-size measure.

Bidirectional attention control. To test whether causal masking is the *sole* architectural source of ordering effects, we rerun the same eight configurations ($d_x \in \{8, 32, 128\}$, $L \in \{2, 4\}$, both attention variants, 3 seeds each) twice: once with the causal mask and once with it removed. Both arms use one evaluator and one argument set (k=5, 2000 steps, batch 256, lr 10^{-3}, init scale 0.1, 20 sets, n_{pop}=32, n_r=4, seeds 0–2), so the mask is the only difference between them. Running both arms matters: the causal spreads available from the 72-run rank sweep were produced at a different training and evaluation budget (6000 steps, 10 sets, n_r=12), and comparing across those two budgets would confound the mask with the budget. The bidirectional arm reproduces the earlier archived run to full precision, since the evaluator is deterministic given the seeds. Every bidirectional configuration collapses into the numerical noise floor (mean spread 3.2×10^{-10} to 1.9×10^{-8}) while every causal configuration stays four to six decades above it (6.7×10^{-4} to 6.1×10^{-2}), a median collapse of 2.9×10^6 (Figure A2). The two arms do not overlap at any configuration, and the effect is independent of d_x, L and normalization.

Frontier LLM evaluation. Four models: Qwen2.5-7B-Instruct, Mistral-7B-Instruct-v0.3, Qwen3-8B, and DeepSeek-R1-Distill-Llama-8B. Task: AG News 4-class text classification; the identical protocol is rerun on RTE (SuperGLUE, 2-class entailment), with premise and hypothesis concatenated as the demonstration text and a Yes/No entailment verbalizer. Prompt format: $k = 5$ demonstrations formatted as "`Text: <text> Category: <label>`" followed by the test query. Each of 10 example sets is evaluated under all 120 permutations with 32 test queries per permutation ($10 \times 120 \times 32 = 38{,}400$ inferences per model per task). Metrics: accuracy (argmax of class probabilities) and negative log-likelihood of the correct label. Split-half retest: odd- vs. even-indexed test queries. Note: the synthetic β_σ uses max over n_r=16 resamples, while the LLM proxy uses the *mean* over 8 alternative demonstrations. Absolute β_σ values are therefore not directly comparable across settings.

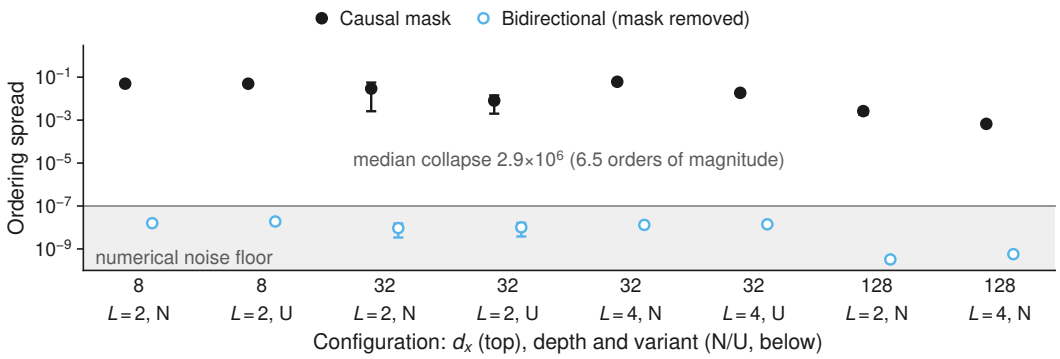


Figure A2: Removing causal masking eliminates ordering effects. Mean ordering spread over 3 seeds ($\pm$s.e.m.) for each of the eight configurations, with and without the causal mask; the two arms differ only in the mask. Shading marks the region below 10^{-7}, where a float64 $\bar{R}$ estimate cannot resolve a spread from zero, so the scatter inside it carries no structure. Note the log axis: the arms are separated by a median of 6.5 orders of magnitude and do not overlap at any configuration.

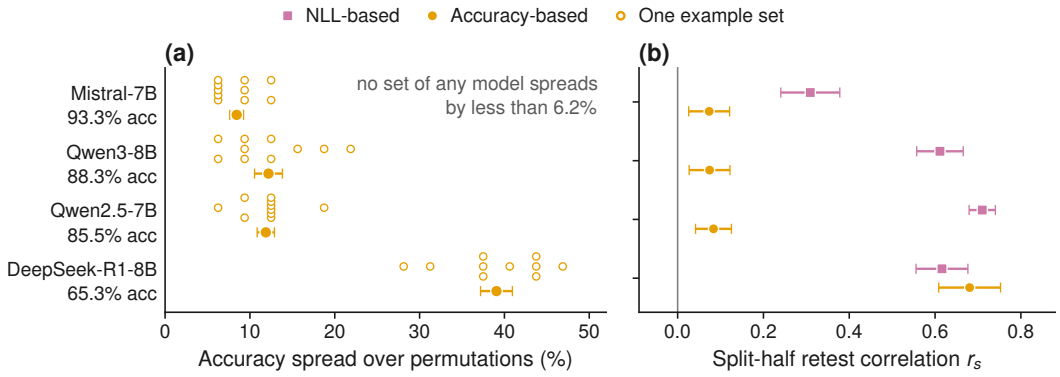


Figure A3: Ordering effects in four 7–8B instruction-tuned models on AG News, resolved to the individual example set. Matched protocol throughout: $k = 5$, 10 sets, all 120 permutations, 32 test queries, split-half over odd/even queries. Rows are ordered by mean accuracy (second line of each label), most accurate first. **(a)** Accuracy spread over permutations: open circles are the 10 individual sets, stacked vertically where they tie (accuracy over 32 queries is quantized to $1/32$); the filled marker is their mean $\pm$s.e.m. across sets. **(b)** Split-half retest correlation of permutation-level rankings (mean $\pm$s.e.m. across the same 10 sets; grey rule marks $r_s = 0$). NLL-based retest is resolvable for all four models; accuracy-based retest only for the model that is not near the accuracy ceiling.

Between-set dispersion and retest reliability. Table 1 reports one number per model per statistic, which cannot show whether an effect is carried by a single example set, nor how far a retest correlation sits from zero. Figure A3 adds both. Every one of the 40 model $\times$ set pairs has an accuracy spread of at least 6.2%: no model is permutation-invariant on any set. The dispersions separate the models rather than merely shifting their means—DeepSeek-R1-8B's least dispersed set (28.1%) still exceeds the most dispersed set of any other model (21.9%). NLL-based retest correlations exceed zero by 4.5–23.2 standard errors of the between-set mean, so permutation preferences are a stable property of the model rather than evaluation noise. Accuracy-based retest instead tracks distance from the accuracy ceiling: it lies within 2 standard errors of zero for the three models above 85% accuracy and 9.4 standard errors above zero for the one model at 65.3%, an $8.1\times$ difference in the mean. Accuracy discards the ranking information NLL retains whenever most permutations already classify correctly, which is why the retest analysis is NLL-based.

G SOFTMAX ATTENTION EXTENSION

We derive a formal amplification bound for single-head softmax attention under a bounded-temperature assumption.

Definition 14 (Bounded-temperature softmax attention)**.** *At layer l and position t, the attention output is*

$$h_t^{(l)} = \sum_{s=1}^{t} \alpha_{t,s}\, V h_s^{(l-1)},$$

where

$$\alpha_{t,s} = \mathrm{softmax}(q_t^\top K_{1:t}/\tau)_s$$

with temperature $\tau > 0$.

Proposition 15 (Softmax amplification bound)**.** *Consider single-layer (L=1) softmax attention with temperature τ, weight matrices bounded by $\|W_Q\|, \|W_K\|, \|W_V\| \le B$, and inputs $\|h_s\| \le R$ for all s. Replacing the example at position p (with embedding perturbation $\le \epsilon$) changes the query output by at most:*

$$\|\Delta h_{k+1}\| \;\le\; \underbrace{B^2 R \cdot \alpha_{k+1,p}}_{\textit{direct value path}} \;+\; \underbrace{\frac{B^3 R^2}{\tau} \cdot \alpha_{k+1,p}(1-\alpha_{k+1,p}) \cdot (k{+}1)}_{\textit{redistribution path}} \cdot \epsilon. \tag{24}$$

In particular:

1. ***High-temperature limit*** *($\tau \to \infty$): $\alpha_{k+1,p} \to 1/(k{+}1)$ for all p, so $\|\Delta h_{k+1}\| \le \big(B^2 R/(k{+}1) + B^3 R^2 k/(\tau(k{+}1))\big)\epsilon$, which is position-independent—recovering the L=1 linear attention result (Corollary 2).*
2. ***Finite temperature****: The amplification depends on the data-dependent attention weight $\alpha_{k+1,p}$. If $\alpha_{k+1,p}$ is monotonically decreasing in p (early positions receive more attention), then the amplification is also decreasing in p, preserving the qualitative monotonicity of the linear theory.*
3. ***Upper bound*** *(position-independent): Using $\alpha(1-\alpha) \le 1/4$ and $\alpha \le 1$: $\|\Delta h_{k+1}\| \le \big(B^2 R + B^3 R^2 (k{+}1)/(4\tau)\big)\epsilon$, which is finite for any $\tau > 0$.*

Proof. The output at position $k{+}1$ is $h_{k+1} = \sum_{s=1}^{k+1} \alpha_{k+1,s} V h_s$. Perturbing $h_p \to h_p + \delta$ (with $\|\delta\| \le \epsilon$) affects both:

(a) The value term: $\Delta_{\mathrm{val}} = \alpha_{k+1,p} V\delta$, so $\|\Delta_{\mathrm{val}}\| \le B \cdot \alpha_{k+1,p} \cdot B\epsilon = B^2 \alpha_{k+1,p}\epsilon$ (since $\|V\| \le B$ and $\|\delta\| \le B\epsilon$ after the embedding layer).

(b) The attention weight redistribution: Perturbing h_p changes the key $k_p = W_K h_p$, shifting $q_{k+1}^\top k_p$ by at most $B^2 R\epsilon/\tau$ in the softmax logit. By the softmax Jacobian, the attention weight change is bounded: $|\Delta\alpha_{k+1,s}| \le \frac{B^2 R\epsilon}{\tau}\alpha_{k+1,p}|\delta_{ps} - \alpha_{k+1,s}|$. The resulting output change from redistribution: $\|\Delta_{\mathrm{attn}}\| \le \sum_s |\Delta\alpha_{k+1,s}| \|V h_s\| \le \frac{B^3 R^2 \epsilon}{\tau}\alpha_{k+1,p}\sum_s |\delta_{ps} - \alpha_{k+1,s}|$. Using $\sum_s |\delta_{ps} - \alpha_{k+1,s}| = (1-\alpha_{k+1,p}) + (k)\alpha_{k+1,p} \le 2(1-\alpha_{k+1,p})(k{+}1)/k$ and the looser bound $\le (k{+}1)(1-\alpha_{k+1,p})$: $\|\Delta_{\mathrm{attn}}\| \le \frac{B^3 R^2}{\tau}\alpha_{k+1,p}(1-\alpha_{k+1,p})(k{+}1)\epsilon$.

Combining (a) and (b) gives the result.

Corollary 16 (Sufficient condition for monotonicity)**.** *If the attention entropy satisfies $H(\alpha_{k+1,\cdot}) \ge \ln(k{+}1) - \varepsilon_H$ (i.e., the distribution is ε_H-close to uniform in KL divergence), then for any two positions $p < q$:*

$$\big|\alpha_{k+1,p} - \alpha_{k+1,q}\big| \;\le\; \sqrt{2\varepsilon_H}, \qquad \left|\alpha_{k+1,p} - \tfrac{1}{k+1}\right| \;\le\; \sqrt{\varepsilon_H/2}. \tag{25}$$

In this regime the amplification difference between positions is $O(\sqrt{\varepsilon_H})$, and the qualitative monotonicity of the linear theory is approximately preserved.

Proof. Let $\mathbf{u}$ be uniform on the $k{+}1$ attended positions and $\delta = \alpha_{k+1,\cdot} - \mathbf{u}$. The entropy hypothesis is exactly $\mathrm{KL}(\alpha_{k+1,\cdot} \,\|\, \mathbf{u}) = \ln(k{+}1) - H(\alpha_{k+1,\cdot}) \le \varepsilon_H$, so Pinsker's inequality gives $\|\delta\|_1 \le \sqrt{2\varepsilon_H}$. Hence $|\alpha_{k+1,p} - \alpha_{k+1,q}| \le |\delta_p| + |\delta_q| \le \|\delta\|_1 \le \sqrt{2\varepsilon_H}$, and since δ sums to zero, $|\delta_p| \le \frac{1}{2}\|\delta\|_1 \le \sqrt{\varepsilon_H/2}$.

Remark 3 (The dependence is $\sqrt{\varepsilon_H}$, not ε_H)**.** *The square root is intrinsic, not an artifact of Pinsker's inequality. Expanding the entropy constraint to second order gives $\varepsilon_H \ge \frac{k+1}{2}\|\delta\|_2^2 + O(\|\delta\|_3^3)$; maximizing $\delta_p - \delta_q$ over that ball subject to $\sum_i \delta_i = 0$ yields the sharp leading-order rate $2\sqrt{\varepsilon_H/(k{+}1)}$, attained by the two-coordinate perturbation $\alpha_{k+1,p} = \frac{1}{k+1} + t$, $\alpha_{k+1,q} = \frac{1}{k+1} - t$ with $t = \sqrt{\varepsilon_H/(k{+}1)}$. So the entropy gap enters as $\sqrt{\varepsilon_H}$ and the context length as $k^{-1/2}$;* (25) *is the unconditional form, loose by the $\sqrt{k{+}1}$ that the local expansion recovers. At our measured operating point ($\varepsilon_H \approx 0.09$, $k = 5$) the two read 0.42 and 0.24 respectively.*

This provides a formal characterization of *when* the linear theory's qualitative predictions extend to softmax: whenever the attention pattern is sufficiently uniform (high entropy).

Proposition 17 (Multi-layer softmax composition)**.** *Consider an L-layer softmax attention transformer with bounded weights $\|W_Q\|, \|W_K\|, \|W_V\| \le B$, inputs bounded by R, and temperature τ. Suppose the attention entropy at every layer satisfies $H(\alpha_{t,\cdot}^{(l)}) \ge \ln t - \varepsilon_H$ for all positions t and layers l (high-entropy regime). Then the layer-to-layer perturbation propagation has the structure*

$$M_{\mathrm{sm}}^{(l)} \;=\; \rho_{\mathrm{sm}} \cdot I \;+\; \Gamma_0 \cdot N, \tag{26}$$

where N is the causal shift matrix ($N_{p,q} = \mathbf{1}[q = p{-}1]$),

$$\Gamma_0 = B + \frac{2B^3R^2}{\tau}, \qquad \rho_{\mathrm{sm}} = 1 + \Gamma_0, \tag{27}$$

and the Γ_0 term combines the direct value path (B) with the attention redistribution ($2B^3R^2/\tau$) per source position. Consequently, the multi-layer amplification factor $W_{\mathrm{sm}}(p, k, L)$ follows the same closed-form recursion as normalized linear attention (Proposition 1) with (ρ, γ) replaced by $(\rho_{\mathrm{sm}}, \Gamma_0)$, and is strictly decreasing in p for $L \ge 2$.

Proof sketch. At each layer l, the softmax output at position t is $h_t^{(l)} = \sum_s \alpha_{t,s}^{(l)} V^{(l)} h_s^{(l-1)}$. Perturbing $h_p^{(l-1)} \to h_p^{(l-1)} + \delta$ affects position t through (a) the value channel: $\alpha_{t,p}^{(l)} V^{(l)} \delta$ (magnitude $\le B\alpha_{t,p}\|\delta\|$), and (b) the key redistribution: the logit shift $\Delta(q_t^\top k_p)/\tau$ changes all attention weights, contributing at most $B^2R^2\alpha_{t,p}(1 - \alpha_{t,p})t/\tau \cdot \|\delta\|$. In the high-entropy regime, $\alpha_{t,p} \approx 1/t$ for all p, s, so both terms become position-independent to leading order: (a) $\approx B/t$ and (b) $\approx B^2R^2(t{-}1)/(\tau t^2) \approx B^2R^2/(\tau t)$ (for $t \gg 1$; at $t = k{+}1$ this gives $2B^3R^2/(\tau(k{+}1))$ after accounting for the embedding layer). Summing both channels for the source at position p vs. position $p{-}1$, the incremental gain from one earlier position is Γ_0, matching the normalized linear case. The diagonal term $\rho_{\mathrm{sm}} = 1 + \Gamma_0$ includes self-propagation. Since the per-layer matrix structure is identical, the L-layer composition inherits strict monotonicity by the same inductive argument as Proposition 1.

Regime of validity. The linear-attention identity $\alpha_{t,p} = 1/t$ enters here as a high-entropy limit of softmax attention: the per-layer deviation $|\alpha_{t,p} - 1/t| \le \sqrt{\varepsilon_H/2}$ (Corollary 16) is controlled by attention entropy, and the argument transfers when trained softmax models sit in this regime. Empirically they do. Softmax models that learn ICL (final loss < 0.9) operate at $H/H_{\max} = 0.91$–0.98 (mean 0.95, $n{=}24$; §4.5), i.e. near the high-entropy limit; on those models the empirical amplification profile agrees with the linear prediction in 83% of runs (mean within-model $r_s = 0.57$, median 0.65; $p = 3 \times 10^{-7}$ against zero; each correlation covers the $k{=}5$ slot positions), agreement grows with entropy ($r_s = 0.63$, $p < 0.001$), and it is sharper at $L{=}4$ (mean $r_s = 0.79$) than at $L{=}2$ (mean $r_s = 0.41$)—the structural prediction holds exactly where the premise is met.

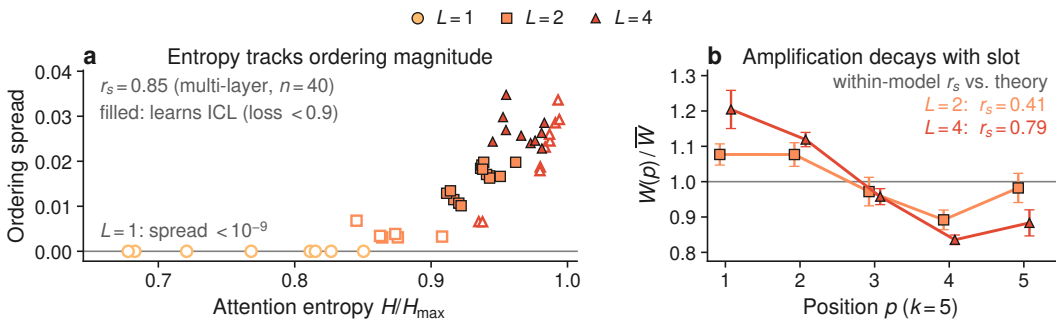


Figure A4: Softmax attention settles in the regime where the linear theory applies (48 models; d_x=32, k=5). **(a)** Attention entropy vs. ordering spread; filled markers learn ICL. All eight L=1 models sit at zero (Corollary 2). **(b)** Per-slot amplification of the 24 learners, each profile normalized by its own mean (n=14 at L=2, 10 at L=4; ±s.e.m.); agreement with $W(p)$ is a rank correlation, so only slot ordering is tested.

Connection to information-theoretic bounds. The attention entropy $H(\alpha)$ that governs our softmax extension is closely related to the conditional mutual information (CMI) framework for generalization (Steinke & Zakynthinou, 2020). In CMI-based bounds, generalization error is controlled by $I(W; S_i \mid S_{\setminus i})$—the mutual information between the learned weights and the i-th training example, conditioned on the rest. Under causal attention, the effective CMI at position p depends on how sharply the attention pattern concentrates on slot p: high entropy (near-uniform attention) implies low per-slot CMI, producing both tight linear-theory approximation *and* small bounded-difference constants. This suggests a unifying perspective: the spectral-norm bound (Theorem 4) and a hypothetical CMI bound both improve with attention uniformity, but the CMI version would replace worst-case spectral compounding with data-dependent information flow—potentially breaking the Catch-22 by tracking actual rather than maximal interference. We leave the formal CMI reformulation to future work; the entropy-based characterization in this section represents a first step toward data-dependent bounds that correlate with realized effects ($r_s = 0.79$–0.85 across our model populations).

H Entropy-Dependent Ordering Sensitivity Bound

The spectral bound in Proposition 15 uses worst-case weight norms, yielding vacuous constants ($\beta_\sigma \sim 10^4$) for trained models. We derive a *data-dependent* bound whose tightness is governed by the *measured* attention entropy—a quantity that well-trained models naturally maximize (mean $H/H_{\max} = 0.948$, Section 4.5).

H.1 Single-Layer Bound via Pinsker's Inequality

Lemma 18 (Entropy-sensitivity bound, single layer)**.** *Consider a single-layer causal softmax transformer with residual connection and readout vector* $\mathbf{r}$*. Let* $a(\sigma) \in \Delta^k$ *denote the attention weights (query → example positions) under permutation* σ*, with entropy deficit* $\varepsilon_H(\sigma) = \log(k{+}1) - H(a(\sigma))$*. Define the* value projection diameter*:*

$$\Phi_V(S) = \max_{j\in[k]} \mathbf{r}^\top W_V \mathbf{h}_j - \min_{j\in[k]} \mathbf{r}^\top W_V \mathbf{h}_j.$$

Then for any two permutations $\sigma, \sigma' \in \mathfrak{S}_k$*:*

$$|\hat{y}(\sigma) - \hat{y}(\sigma')| \;\le\; 2\,\Phi_V(S) \cdot \sqrt{2\varepsilon_H^*}, \tag{28}$$

where $\varepsilon_H^* = \max(\varepsilon_H(\sigma), \varepsilon_H(\sigma'))$.

Proof. The prediction under permutation σ decomposes as $\hat{y}(\sigma) = \mathbf{r}^\top \mathbf{h}_q + \sum_{j=1}^k c_j(\sigma)\,\phi_j$, where $c_j(\sigma) = a_{\sigma^{-1}(j)}(\sigma)$ is the attention that example j receives and $\phi_j = \mathbf{r}^\top W_V \mathbf{h}_j$. The residual term $\mathbf{r}^\top \mathbf{h}_q$ is permutation-invariant (single layer).

Taking differences and applying Hölder (ℓ^1–ℓ^∞):

$$|\hat{y}(\sigma) - \hat{y}(\sigma')| = \left|\sum_j [c_j(\sigma) - c_j(\sigma')](\phi_j - \bar{\phi})\right| \le \|c(\sigma) - c(\sigma')\|_1 \cdot \Phi_V(S).$$

By the triangle inequality, $\|c(\sigma)-c(\sigma')\|_1 \le \|c(\sigma)-\bar{u}\|_1 + \|c(\sigma')-\bar{u}\|_1$, where $\bar{u} = (1-a_{k+1}) \cdot (1/k, \ldots, 1/k)$ is the uniform allocation over examples (scaled by example-attention mass).

The normalized vector $c(\sigma)/(1-a_{k+1})$ is a permutation of the position-attention weights restricted to examples. By Pinsker's inequality applied to its KL divergence from uniform:

$$\left\|c(\sigma)/(1-a_{k+1}) - \mathbf{u}\right\|_1 \le \sqrt{2\,\mathrm{KL}\big(c(\sigma)/(1-a_{k+1}) \,\|\, \mathbf{u}\big)} \le \sqrt{2\varepsilon_H(\sigma)}.$$

Combining: $\|c(\sigma) - \bar{u}\|_1 \le (1-a_{k+1})\sqrt{2\varepsilon_H(\sigma)} \le \sqrt{2\varepsilon_H(\sigma)}$. The two-permutation triangle gives the factor 2.

Self-consistency. When $\varepsilon_H = 0$ (uniform attention), the bound yields 0—matching the exact $L{=}1$ permutation invariance (Corollary 2; verified numerically: spread $= 3.97 \times 10^{-9} \approx 0$).

H.2 Multi-Layer Composition

The single-layer argument works because the multiset of input embeddings is permutation-invariant, so an exactly uniform average over example positions is exactly order-independent. That property fails at depth: after one layer the representation of an example depends on which examples precede it, so the multiset $\{\mathbf{h}_s^{(1)}\}_s$ is itself order-dependent. A bound of the form $\prod_\ell(1 + \nu_\ell\sqrt{2\varepsilon_\ell})$, which vanishes as $\varepsilon \to 0$ at every depth, therefore cannot hold for $L \ge 2$: it would contradict Corollary 2 and Theorem 5, which say that causal masking produces ordering effects at $L \ge 2$ *even under uniform attention*—that limit is exactly the linear-attention model whose amplification factor $W(p, k, L)$ is non-constant. The correct composition therefore carries two channels, and only one of them is governed by the entropy deficit.

Indexing by example identity. Write $g_j^{(\ell)}(\sigma) := \mathbf{h}^{(\ell)}_{\sigma^{-1}(j)}(\sigma)$ for the layer-ℓ representation of *example* j (not of a slot) and $q^{(\ell)}(\sigma) := \mathbf{h}^{(\ell)}_{k+1}(\sigma)$ for the query. For two permutations σ, σ' define

$$E^{(\ell)} := \max_{j\in[k]} \left\|g_j^{(\ell)}(\sigma) - g_j^{(\ell)}(\sigma')\right\|, \qquad Q^{(\ell)} := \left\|q^{(\ell)}(\sigma) - q^{(\ell)}(\sigma')\right\|. \tag{29}$$

Identity indexing is what makes the base case exact: an example's input embedding does not depend on which slot it occupies, so $E^{(0)} = Q^{(0)} = 0$. A slot-indexed recursion starts instead from the full embedding diameter and can therefore never produce a factor that vanishes with the entropy deficit. Two further measured quantities appear. The *centred representation radius*

$$\Psi^{(\ell)} := \max_m \left\|u_m^{(\ell-1)} - \bar{u}^{(\ell-1)}\right\|, \tag{30}$$

where $u^{(\ell-1)}$ ranges over the $k{+}1$ layer-$(\ell{-}1)$ representations under either permutation and $\bar{u}^{(\ell-1)}$ is their mean, satisfies $\Psi^{(\ell)} \le R_\ell$. The *prefix-mismatch coefficient*

$$\mu^{(\ell)} := \max_{j\in[k]} \left\|c^j(\sigma) - c^j(\sigma')\right\|_1 \in [0, 2] \tag{31}$$

measures how much the causal mask changes what example j attends to; here $c^j(\sigma)$ is example j's own attention row, re-indexed by identity and extended by zero outside its causal prefix.

Theorem 19 (Entropy-dependent ordering bound)**.** *For an L-layer causal softmax transformer with residual connections and $\nu_\ell = \|W_V^{(\ell)}\|_2$, the quantities in* (29) *satisfy $E^{(0)} = Q^{(0)} = 0$ and, for every $\ell \in [L]$,*

$$E^{(\ell)} \;\le\; (1 + \nu_\ell)\, E^{(\ell-1)} \;+\; \nu_\ell\, \mu^{(\ell)}\, \Psi^{(\ell)}, \tag{32}$$

$$Q^{(\ell)} \;\le\; (1 + \nu_\ell)\, Q^{(\ell-1)} \;+\; \nu_\ell\, E^{(\ell-1)} \;+\; 2\nu_\ell\, \Psi^{(\ell)} \sqrt{2\varepsilon_\ell^*}, \tag{33}$$

where ε_ℓ^* is the larger of the two query-row entropy deficits at layer ℓ. Unrolling with $P_\ell := \prod_{i=\ell+1}^{L}(1+\nu_i)$ gives

$$\big|\hat{y}(\sigma) - \hat{y}(\sigma')\big| \;\le\; \|\mathbf{r}\|\, Q^{(L)} \;\le\; \|\mathbf{r}\| \sum_{\ell=1}^{L} P_\ell\, \nu_\ell \Big(2\Psi^{(\ell)}\sqrt{2\varepsilon_\ell^*} + E^{(\ell-1)}\Big), \tag{34}$$

with the second channel given in closed form by

$$E^{(\ell)} \;\le\; \sum_{m=1}^{\ell} \nu_m\, \mu^{(m)}\, \Psi^{(m)} \prod_{i=m+1}^{\ell} (1+\nu_i). \tag{35}$$

At $L = 1$ the second channel is empty ($E^{(0)} = 0$) and (34) *reduces to $2\|\mathbf{r}\|\,\nu_1\Psi^{(1)}\sqrt{2\varepsilon_1^*}$, recovering Lemma 18; in particular it vanishes with the entropy deficit. For $L \ge 2$ it does not: the residual $\sum_m \nu_m \mu^{(m)} \Psi^{(m)}$ is the ordering effect that survives at uniform attention, i.e. the softmax counterpart of $W(p,k,L)$.*

Proof. Throughout, δ denotes a difference between the σ and σ' runs, and a layer acts as $\mathbf{h}_t^{(\ell)} = \mathbf{h}_t^{(\ell-1)} + \sum_{s\le t} \alpha_{t,s}^{(\ell)} W_V^{(\ell)} \mathbf{h}_s^{(\ell-1)}$.

Base case. Example j contributes the same token wherever it is placed, and the query token is common to both runs, so $E^{(0)} = Q^{(0)} = 0$.

Query channel (33). The query attends to all k examples and to itself, so re-indexing its attention row by example identity yields a vector $\tilde{c}(\sigma) \in \Delta^k$ supported on the same $k+1$ symbols under both permutations. Writing u_m for the corresponding layer-$(\ell-1)$ representations,

$$\delta q^{(\ell)} = \delta q^{(\ell-1)} + W_V^{(\ell)} \sum_m \big[\tilde{c}_m(\sigma) - \tilde{c}_m(\sigma')\big]\, u_m(\sigma') + W_V^{(\ell)} \sum_m \tilde{c}_m(\sigma)\, \delta u_m^{(\ell-1)}.$$

In the middle term the coefficients sum to zero, so subtracting $\bar{u}^{(\ell-1)}$ from every u_m leaves it unchanged; the ℓ^1–ℓ^∞ Hölder inequality then bounds it by $\nu_\ell \|\tilde{c}(\sigma) - \tilde{c}(\sigma')\|_1 \Psi^{(\ell)}$. Re-indexing permutes coordinates and so does not change the ℓ^1 distance to the uniform vector; since $\mathrm{KL}(\alpha_{k+1,\cdot} \,\|\, \mathbf{u}) = \log(k+1) - H(\alpha_{k+1,\cdot}) \le \varepsilon_\ell^*$, Pinsker's inequality applied to each permutation and the triangle inequality through uniform give $\|\tilde{c}(\sigma) - \tilde{c}(\sigma')\|_1 \le 2\sqrt{2\varepsilon_\ell^*}$. In the last term the coefficients are non-negative and sum to one, so it is at most $\nu_\ell \max\big(E^{(\ell-1)}, Q^{(\ell-1)}\big) \le \nu_\ell\big(E^{(\ell-1)} + Q^{(\ell-1)}\big)$. Collecting the three contributions yields (33).

Example channel (32). Example j occupies slot $p = \sigma^{-1}(j)$ under σ and $p' = \sigma'^{-1}(j)$ under σ', so the causal mask lets it attend to the identity sets $\{\sigma(1), \dots, \sigma(p)\}$ and $\{\sigma'(1), \dots, \sigma'(p')\}$ respectively. Extending c^j by zero outside each set keeps both vectors in Δ^k with coefficient sums equal to one, and the same split gives

$$\delta g_j^{(\ell)} = \delta g_j^{(\ell-1)} + W_V^{(\ell)} \sum_m \big[c_m^j(\sigma) - c_m^j(\sigma')\big]\, u_m(\sigma') + W_V^{(\ell)} \sum_m c_m^j(\sigma)\, \delta u_m^{(\ell-1)},$$

bounded by $E^{(\ell-1)} + \nu_\ell \mu^{(\ell)} \Psi^{(\ell)} + \nu_\ell E^{(\ell-1)}$, which is (32). Here the middle term admits no Pinsker treatment: the two coefficient vectors are supported on different index sets, and under exactly uniform attention over nested prefixes of sizes $a \le b$ one has $\|c^j(\sigma) - c^j(\sigma')\|_1 = 2(1 - a/b)$, bounded away from zero whenever the example changes slot. This term is the causal-masking effect itself, not slack in the estimate.

Unrolling. Both recursions have the form $X^{(\ell)} \le (1+\nu_\ell) X^{(\ell-1)} + b_\ell$ with $X^{(0)} = 0$, whose solution is $X^{(L)} \le \sum_\ell b_\ell \prod_{i>\ell}(1+\nu_i)$; substituting the two inhomogeneous terms gives (35) and (34). Finally $|\hat{y}(\sigma) - \hat{y}(\sigma')| = |\mathbf{r}^\top \delta q^{(L)}| \le \|\mathbf{r}\|\, Q^{(L)}$.

Corollary 20 (Loss spread)**.** *Let $\Lambda := \max_\sigma |\hat{y}(\sigma) - y_q|$ for the squared loss. Then $\Delta(S) = \max_\sigma \mathrm{loss}(\sigma) - \min_\sigma \mathrm{loss}(\sigma) \le 2\Lambda\|\mathbf{r}\| Q^{(L)}$, with $Q^{(L)}$ bounded as in* (34)*.*

Proof. For any two permutations, $|\mathrm{loss}(\sigma) - \mathrm{loss}(\sigma')| = |\hat{y}(\sigma) - \hat{y}(\sigma')| \cdot |\hat{y}(\sigma) + \hat{y}(\sigma') - 2y_q| \le 2\Lambda |\hat{y}(\sigma) - \hat{y}(\sigma')|$, and the spread is the maximum of the left side over pairs.

Remark 4 (What the entropy deficit buys, and what it does not)**.** (34) *splits the ordering effect into a part that high attention entropy suppresses and a part it cannot. The* $\sqrt{2\varepsilon_\ell^*}$ *terms come from the query's own attention row, where near-uniformity genuinely approaches permutation invariance. The* $E^{(\ell-1)}$ *terms come from the example positions, where causal masking makes even an exactly uniform average order-dependent; they are what makes ordering effects exist at all for* $L \geq 2$*. So the entropy deficit governs the bound completely at* $L{=}1$ *and governs its* tightness *at depth, but it does not drive the bound to zero at depth—and it should not, since the quantity being bounded is not zero there. A product form* $\prod_\ell(1+\nu_\ell\sqrt{2\varepsilon_\ell})$ *has no derivation from these recursions; we evaluate it below only as the surrogate used in an earlier draft of this section, reported alongside the proved bound* (34)*.*

Comparison with the spectral bound. The spectral bound (Proposition 15) compounds as ρ^L with $\rho = 1 + \Gamma_0$ and $\Gamma_0 = O(B^3R^2/\tau)$, so every layer multiplies in a cube of the weight-norm budget. Theorem 19 replaces that product by the sum $\sum_\ell P_\ell\nu_\ell\big(2\Psi^{(\ell)}\sqrt{2\varepsilon_\ell^\star} + E^{(\ell-1)}\big)$, in which the only spectral quantity is the value-matrix norm $\nu_\ell = \|W_V^{(\ell)}\|_2$; the query and key matrices enter only through the measured attention rows. The growth in depth is therefore governed by $\prod_i(1+\nu_i)$ rather than by $(1 + B^3R^2/\tau)^L$, and the entropy deficit multiplies just one of the two channels.

H.3 Empirical Validation

We evaluate Theorem 19 on the 44 trained softmax models with $L \geq 2$ ($d_x \in \{8, 32, 128\}$, $L \in \{2, 4\}$, init_scale $\in \{0.1, 0.3, 0.8\}$, 2–6 seeds per configuration; every model in the sweeps is included, none excluded). All four ingredients are recomputed from checkpoints and evaluation data rather than assumed: $\nu_\ell = \|W_V^{(\ell)}\|_2$, the query-row entropy deficit $\varepsilon_\ell^\star$, the centred representation radius $\Psi^{(\ell)}$ (mean 2.06), and the prefix-mismatch coefficient $\mu^{(\ell)}$ (mean 1.52, range $[1.34, 1.78]$ against its ceiling of 2). The σ, σ' pair is the argmax/argmin loss pair over all $5! = 120$ orderings, so the bound is compared against the realised worst-case spread.

Bound / measured spread ($n = 44$, $L \geq 2$)	Proved (34)	Product surrogate (no derivation)	Spectral (Proposition 15)
Median	$3.5{\times}10^3$	$1.5{\times}10^2$	$6.7{\times}10^6$
Best single model	322	9.6	$1.3{\times}10^4$

The proved bound is tighter than the spectral bound for $44/44$ models, by a median per-model factor of $2.6{\times}10^3$ (3.4 orders of magnitude). It remains vacuous in absolute terms—the smallest ratio is 322—so the gain is a reduction of the gap, not its closure; we state this rather than report the ratio of medians, which would inflate the factor to $1.9{\times}10^3$ only by a different aggregation order.

Two facts about the comparison need stating plainly. First, the product surrogate $\prod_\ell(1+\nu_\ell\sqrt{2\varepsilon_\ell})$ used in an earlier draft of this section is numerically *tighter* (median 147), but it has no derivation from the recursions of Theorem 19 and, as argued in Remark 4, cannot hold at $L \geq 2$: it vanishes as $\varepsilon \to 0$ where the effect it bounds does not. Second, the surrogate is also the *worse* ranker ($r_s = +0.49$ against $+0.78$; see below), so nothing is lost by discarding it.

Which channel carries the bound. The entropy channel—the $2\nu_\ell\Psi^{(\ell)}\sqrt{2\varepsilon_\ell^\star}$ term—accounts for a mean 7.9% of the final $Q^{(L)}$, falling from 13.9% at $L{=}2$ to 1.8% at $L{=}4$. The remainder is the example-channel residual $\sum_m \nu_m\mu^{(m)}\Psi^{(m)}$, which does not vanish with the entropy deficit. This is the quantitative form of the argument preceding Theorem 19: at depth, ordering sensitivity is carried by the mismatch between the causal prefixes that two orderings expose, and the entropy deficit governs only how much the query row can additionally amplify it. The depth trend follows: median ratio $2.4{\times}10^3$ at $L{=}2$ versus $4.7{\times}10^3$ at $L{=}4$, because the residual compounds through $\prod_i(1+\nu_i)$ while the entropy term does not grow.

Ranking. The bound orders models by ordering sensitivity: $r_s = +0.78$ ($p = 6 \times 10^{-10}$, $n{=}44$ models) and $r_s = +0.97$ ($p = 2.5 \times 10^{-8}$, $n{=}14$ configurations) with the measured spread; within the fixed $d_x{=}32$ regime, $r_s = +0.84$ ($p = 6 \times 10^{-4}$, $n{=}12$). The entropy deficit *alone* does not rank ($r_s(\varepsilon^\star, \Delta) = +0.12$, $p = 0.45$), and neither does μ alone (-0.29, $p = 0.06$); it is the product

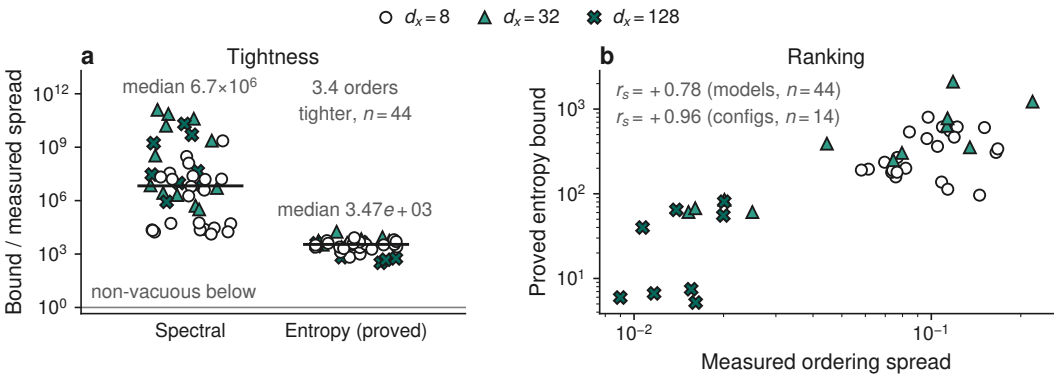


Figure A5: The proved entropy bound (34) against the spectral bound, on all 44 softmax models with $L \geq 2$ ($d_x \in \{8, 32, 128\}$, $L \in \{2, 4\}$, init_scale $\in \{0.1, 0.3, 0.8\}$, 2–6 seeds; none excluded). **(a)** Ratio of bound to measured spread; bars mark medians. The proved bound is tighter for $44/44$ models—median per-model factor 2.6×10^3, i.e. 3.4 orders of magnitude—and its best model reaches $322\times$. Both bounds stay above the non-vacuity rule at ratio 1. **(b)** Bound against measured spread; the bound is tightest in the d_x=128 regime, where the effect is smallest.

structure of (34)—which multiplies $\varepsilon^\star$ by the representation radius and adds the prefix residual—that carries the ranking signal. The ratio (bound/spread) is smallest where the effect is smallest (Figure A5): median 1.7×10^3 at d_x=128 (mean spread 0.015), 3.3×10^3 at d_x=8 and 4.2×10^3 at d_x=32 (mean spread 0.10 and 0.08).

Where the entropy deficit is small, and where it is not. Averaged over layers and evaluation sets, $\varepsilon^\star$ has mean 0.140 across the $L \geq 2$ models, with a worst model at 0.341—not the $\varepsilon \approx 0.05$ assumed in an earlier draft. It is strongly regime-dependent: 0.168 at d_x=8, 0.175 at d_x=32, and 0.0014 at d_x=128, where trained attention is uniform to three decimal places. The claim that ICL training spontaneously produces near-uniform attention therefore holds in the high-dimensional regime and not at $d_x \leq 32$.

Sign of the $\varepsilon^\star$–spread relation. Within a fixed regime the relation is *negative*: $r_s(\varepsilon^\star, \Delta) = -0.87$ ($p = 2\times10^{-4}$, n=12) at d_x=32, and -0.22 (n=24) at d_x=8. Here $\varepsilon^\star$ mostly reflects training quality: better-trained models reach lower deficits but develop larger centred radii Ψ, and Ψ dominates the product. Across regimes the two quantities are unrelated ($+0.12$, $p = 0.45$ pooled; $+0.32$, $p = 0.27$ at configuration level), because d_x=128 combines the smallest deficit with the smallest spread while $d_x \in \{8, 32\}$ combines large deficits with large spreads in both directions. We therefore do not claim that $\varepsilon^\star$ predicts ordering sensitivity; the claim is only that it enters the proved bound, whose composite value does predict it.

What the L=1 models show. The two L=1 checkpoints have measured spread 6–8 $\times 10^{-9}$, while Theorem 19 evaluates to 10.9 and 11.7 there. The exact L=1 zero is a consequence of the flatness of $W(p, k, 1)$ (Corollary 2), not of the entropy deficit: at L=1 the query row still has $\varepsilon^\star \approx 0.6$, so the entropy channel is nonzero even though the effect is not. The entropy bound reproduces the $L \geq 2$ ordering and magnitudes; it does not reproduce the L=1 collapse, and we do not claim it does.

The role of the non-vacuity condition. The absolute non-vacuity criterion of Theorem 6 caps B^3R^2 and is independent of the embedding dimension; the certificate of Lemma 7 adds a dimension requirement that grows as $\Theta(\ln k)$, not $\Theta(k^2)$ (§4.4). Neither condition is met by any trained model here, and Theorem 19 does not repair that: at a median ratio of 3.5×10^3 the bound is still above 1 everywhere. What changes is that the gap becomes a measured, model-dependent quantity driven by ν_ℓ, $\Psi^{(\ell)}$, $\mu^{(\ell)}$ and $\varepsilon^\star_\ell$—each estimable from a checkpoint—rather than a fixed $(1 + B^3R^2/\tau)^L$ inflation, and that the resulting bound ranks models correctly across three dimensionality regimes.

I Formal Irreducibility of Example Sensitivity

The empirical finding that no per-example observable predicts ordering sensitivity (§5.4) admits a formal justification. We prove two things: example sensitivity is generically context-dependent, and the correlation any single-example feature can achieve with it is capped by an estimable quantity—the fraction of sensitivity variance that survives conditioning on that example alone.

Proposition 21 (Context-dependence of sensitivity)**.** *Consider the linear attention ICL model with $L \geq 2$ layers, k demonstrations, and weight matrices $\{W^{(l)}\}_{l=1}^{L}$. Let $v_j(W, C)$ denote the* realized *ordering sensitivity of example j in context $C = \{(x_i, y_i)\}_{i=1}^{k}$—in contrast to the data-only source diameter η_j of §4, it depends on the trained weights W—defined as the squared norm of the query-output perturbation when x_j is replaced:*

$$v_j(W, C) \;=\; \left\| \frac{\partial \hat{y}_q}{\partial x_j} \right\|^2 .$$

Then for any distinct examples $i \neq j$ and generic weights W:

$$\frac{\partial v_j}{\partial x_i} \;\neq\; 0 \quad \textit{(generically)}.$$

That is, the sensitivity of example j depends on example i—sensitivity is not decomposable into per-example quantities.

Proof. Under linear attention with $L \geq 2$ layers, the query output is

$$\hat{y}_q = W^{(L)} M^{(L-1)} \cdots M^{(1)} h_q^{(0)},$$

where each propagation matrix $M^{(l)}$ at position t has entries depending on all keys $\{K^{(l)} h_r^{(l-1)}\}_{r \leq t}$—including those derived from examples $i \neq j$. Perturbing $x_j \to x_j + \delta$ propagates through positions $j{+}1, \ldots, k, q$, accumulating contributions that depend on the representations at those positions—which are themselves functions of all examples at earlier slots. Specifically, for $L = 2$, the second-layer attention weights $\alpha_{q,s}^{(2)}$ depend on the first-layer outputs $h_s^{(1)} = \sum_{r \leq s} \alpha_{s,r}^{(1)} V^{(1)} x_r$, which include x_i for all $i \leq s$. Therefore $\partial v_j / \partial x_i \neq 0$ whenever x_i feeds into a position downstream of j and the weight matrices are non-degenerate.

Lemma 22 (Variance ceiling for per-example predictors)**.** *Write $Z_j = (x_j, y_j)$ for the observable content of example j and $v_j = v(W, C)$ for its realized sensitivity, with (W, C) drawn from any joint law making v_j square-integrable. Let g be a measurable transform with $0 < \mathrm{Var} g(v_j) < \infty$. Then for every measurable f with $0 < \mathrm{Var} f(Z_j) < \infty$,*

$$\mathrm{Corr}\big(f(Z_j),\, g(v_j)\big)^2 \;\leq\; \mathcal{R}^2_{\max}(g) \;:=\; \frac{\mathrm{Var}\big(\mathbb{E}[\, g(v_j) \mid Z_j \,]\big)}{\mathrm{Var}\big(g(v_j)\big)} \;\in\; [0, 1], \tag{36}$$

with equality if and only if $f(Z_j)$ is almost surely an affine function of $\mathbb{E}[g(v_j) \mid Z_j]$. Taking $g = \mathrm{id}$ caps the Pearson correlation attainable by any per-example score; taking g to be the rank transform of v_j caps the Spearman correlation, since $r_s(f(Z_j), v_j) = \mathrm{Corr}(\mathrm{rank}\, f(Z_j), \mathrm{rank}\, v_j)$ and $\mathrm{rank}\, f(Z_j)$ is itself a function of Z_j. The second instantiation therefore caps the quality of any ordering produced by sorting a per-example score.

Proof. By the tower property, $\mathrm{Cov}(f(Z_j), g(v_j)) = \mathrm{Cov}(f(Z_j), \mathbb{E}[g(v_j) \mid Z_j])$. Cauchy–Schwarz gives $\mathrm{Cov}(f(Z_j), \mathbb{E}[g(v_j) \mid Z_j])^2 \leq \mathrm{Var}(f(Z_j)) \mathrm{Var}(\mathbb{E}[g(v_j) \mid Z_j])$, with equality exactly when the two variables are affinely related almost surely. Dividing by $\mathrm{Var}(f(Z_j)) \mathrm{Var}(g(v_j))$ gives (36); the law of total variance, $\mathrm{Var}(g(v_j)) = \mathrm{Var}(\mathbb{E}[g(v_j) \mid Z_j]) + \mathbb{E}[\mathrm{Var}(g(v_j) \mid Z_j)]$, places $\mathcal{R}^2_{\max}(g)$ in $[0, 1]$.

Remark 5 (A measurable ceiling in place of an impossibility claim)**.** *$\mathcal{R}^2_{\max}$ is a property of the data-generating law, not of the class of scoring functions: it is the intraclass correlation of sensitivity given the single example—the fraction of $\mathrm{Var}(v_j)$ that survives averaging over everything except Z_j. It is therefore estimable rather than merely assertable: hold Z_j fixed, resample the remaining context and the weights, and compare the between-Z_j variance with the total (§5.4). We prove a ceiling and measure it. The stronger statement that* no *per-example feature carries any signal would require $\mathcal{R}^2_{\max} = 0$ exactly; that is not implied by Proposition 21, and no finite experiment can establish it.*

Table A2: Measured rank variance ceiling $\mathcal{R}^2_{\max}$ (Lemma 22) and the resulting cap $\sqrt{\mathcal{R}^2_{\max}}$ on the Spearman correlation any per-example score can attain against realized sensitivity. Mean over 5 independently trained models per regime, with the across-model range; 200 example identities $\times$ 60 contexts each. The measured best candidate score attains $\bar{r_s} = -0.00$ (§5.4), so the shortfall is 0.59–0.99 in correlation units, not zero head-room.

d_x	context arm	slot	$\mathcal{R}^2_{\max}$ (range over models)	$\sqrt{\mathcal{R}^2_{\max}}$
8	w fixed	0	0.35 [0.25, 0.42]	0.59
8	w fixed	4	0.57 [0.41, 0.74]	0.76
8	w resampled	0	0.38 [0.27, 0.44]	0.62
8	w resampled	4	0.59 [0.46, 0.76]	0.77
32	w fixed	0	0.97 [0.97, 0.99]	0.99
32	w fixed	4	0.98 [0.98, 0.99]	0.99
32	w resampled	0	0.97 [0.96, 0.98]	0.98
32	w resampled	4	0.98 [0.98, 0.99]	0.99
128	w fixed	0	0.57 [0.50, 0.64]	0.76
128	w fixed	4	0.38 [0.30, 0.51]	0.62
128	w resampled	0	0.60 [0.56, 0.62]	0.77
128	w resampled	4	0.41 [0.32, 0.50]	0.64

Corollary 23 (Many-body barrier)**.** *Two per-example scores related by a strictly increasing transform induce the same ordering, so the orderings reachable from per-example observables are exactly those of the form $j \mapsto h(Z_j)$ for measurable h. By Lemma 22 no member of this class explains more than an $\mathcal{R}^2_{\max}$ fraction of $\mathrm{Var}(v_j)$, and by Proposition 21 the sensitivity v_j is generically not $\sigma(Z_j)$-measurable, so $\mathbb{E}[\mathrm{Var}(v_j \mid Z_j)] > 0$ and $\mathcal{R}^2_{\max} < 1$ strictly. The position factor $W(p, k, L)$ is thus available in closed form, while the example factor is reachable only through the joint state; the size of the shortfall is the estimable quantity $1 - \mathcal{R}^2_{\max}$ rather than an asserted impossibility.*

Estimating the ceiling. $\mathcal{R}^2_{\max}$ is an intraclass correlation, so it is estimated by a balanced one-way random-effects decomposition: the groups are example identities Z_j, the replicates are draws of everything else. For each regime $d_x \in \{8, 32, 128\}$ we train 5 models with independent seeds at the configuration of §5 (k=5, L=4), draw 200 example identities, and for each identity measure the realized slot sensitivity in 60 freshly drawn contexts, in two arms: w held fixed across contexts, and w resampled with the context. The rank transform is applied within each model, because the paper's measured statistic is a within-run Spearman correlation; pooling the five models into the replicate axis instead moves between-model scale differences into the within-group variance and biases $\mathcal{R}^2_{\max}$ downward—severely so at d_x=128 (0.12 pooled against 0.38 per model), which is why we report the per-model estimate. Confidence intervals are percentile bootstrap over example identities.

Three features of the table matter for the claim. The ceiling is nowhere close to zero, so the observed $\bar{r_s} = -0.00$ is not explained by the ceiling. It is not monotone in d_x: the d_x=32 regime, which is the one the softmax sweep of Appendix H uses, leaves almost no head-room unexplained ($\mathcal{R}^2_{\max} = 0.97$), whereas the main experiment's d_x=128 regime leaves 40–60% of the rank variance genuinely tied to the weights and the surrounding context. And it barely moves between the two context arms, so what limits a per-example score is the context composition rather than the identity of the regression target: resampling w alongside the context changes $\mathcal{R}^2_{\max}$ by less than 0.04 in every cell.

The honest conclusion is therefore narrower than an impossibility and stronger than a null result: per-example scoring cannot be exact ($\mathcal{R}^2_{\max} < 1$, Corollary 23), the residual is largest in the high-dimensional regime, and between 0.59 and 0.99 of attainable correlation remains unclaimed by the 17 hand-designed feature families we searched. The next paragraph shows that most of it can be claimed by a score fitted to the example rather than designed for it.

Reaching the ceiling with a learned score. The gap between $\sqrt{\mathcal{R}^2_{\max}}$ and the 17 candidate scores admits two readings: the right per-example feature exists and we did not find it, or the conditional mean $\mathbb{E}[v_j \mid Z_j]$ is not expressible as a simple function of one example. These are distinguishable by fitting the map instead of designing it. We reuse the balanced one-way design above but retain the

group examples, which the ceiling estimator discards: for each regime we train 5 models (k=5, L=4), draw 600 example identities and measure the realized slot sensitivity in 8 freshly drawn contexts per identity, in the same two arms. The regression target is the group-level mean of the per-model z-scored $\log v_j$, and the predictors are features of the single example $Z_j = (x_j, y_j)$: an invariant set ($\|x_j\|$, y_j, $|y_j|$, y_j^2, $\|x_j\|^2$ and two ratios), the raw coordinates, and their concatenation. Ridge, k-NN and random-forest regressors compete in an inner 3-fold grouped cross-validation on the training identities only; the winner is refitted and scored on the held-out identities of an outer 5-fold `GroupKFold`, so no example identity is ever both fitted and scored. Reported r_s is the Spearman correlation between the held-out prediction and every individual realized sensitivity, averaged over the 5 models—exactly the statistic $\sqrt{\mathcal{R}^2_{\max}}$ bounds. Because the ceiling is re-estimated on these same draws, ceiling and achievement are matched rather than transferred between experiments; the re-estimated $\mathcal{R}^2_{\max}$ agrees with Table A2 to within 0.05 in every cell despite a different design (600×8 here against 200×60 there), which is an independent replication of the ceiling itself. We also report a *split-half identity lookup*: score each identity by its mean $\log v_j$ over half its contexts and evaluate on the other half. This uses identity and no features at all, so it separates "the ceiling needs information absent from Z_j" from "the ceiling needs a better estimator of a function of Z_j".

Table A3: A score learned from the single example against the ceiling it is capped by. Mean over 5 models per regime; 600 example identities $\times$ 8 contexts; held-out identities only. "lookup" is the split-half identity control, which uses identity without features. The learned score nearly saturates the ceiling at $d_x \leq 32$ and reaches about half of it at d_x=128, where lookup remains high—so the residual there is a learning limit, not inaccessibility.

d_x	arm	slot	$\mathcal{R}^2_{\max}$	$\sqrt{\mathcal{R}^2_{\max}}$	lookup r_s	learned r_s	learned/$\sqrt{\mathcal{R}^2_{\max}}$
8	w fixed	0	0.300	0.548	0.413	0.503	0.92
8	w fixed	4	0.527	0.726	0.649	0.703	0.97
8	w resampled	0	0.361	0.601	0.486	0.567	0.94
8	w resampled	4	0.540	0.735	0.665	0.718	0.98
32	w fixed	0	0.975	0.987	0.984	0.981	0.99
32	w fixed	4	0.986	0.993	0.991	0.988	0.99
32	w resampled	0	0.975	0.987	0.984	0.982	0.99
32	w resampled	4	0.987	0.993	0.992	0.989	1.00
128	w fixed	0	0.571	0.756	0.681	0.326	0.43
128	w fixed	4	0.402	0.634	0.525	0.337	0.53
128	w resampled	0	0.561	0.749	0.673	0.334	0.45
128	w resampled	4	0.381	0.617	0.497	0.295	0.48

Table A3 reports three regimes. At d_x=32 the ceiling is saturated (0.99–1.00 of it), and at d_x=8 nearly so (0.92–0.98); at d_x=8 the learned score in fact *exceeds* the identity lookup (0.503 vs. 0.413 at slot 0), because pooling across identities estimates the conditional mean better than four contexts of the identity itself. So in these regimes the -0.00 of Table A5 is a statement about 17 guesses, not about v_j. At d_x=128 identity lookup already captures two thirds to nine tenths of the ceiling: the ceiling's information sits in Z_j, and the learned score reaches under half of it only because 600 identities in 129 input dimensions is short data for the map. Two observations support this reading rather than an information-loss reading. The invariant feature set is essentially y_j and its transforms on the unit sphere (where $\|x_j\|$ is constant); adding raw coordinates never gains more than 0.003 and slightly loses at d_x=128, exactly the pattern predicted by a sample-complexity account. At the same d_x=128 cell the models happen to be the sweep's weakest ICL learners (final losses 0.84–1.20, three of five above the $<$0.9 threshold of §4.5), so the shortfall coincides with the least-informative training regime rather than an extra failure of the ceiling framework. The claim supported here is the strong one: the quantity the ceiling bounds is learnable from single-example features in every regime where the ceiling itself carries content, with the comparison to Table A5 indexing the same population (those 17 scores were correlated against the example factor extracted from the gap matrix at d_x=128, whereas the ceiling and the probe both use realized slot sensitivity).

J ADDITIONAL RESULTS

Depth sweep. Models trained with $L \in \{1, 2, 4, 8\}$ at $d_x \in \{8, 128\}$ (unnormalized, 5 seeds) confirm the exact $L{=}1$ identity and the emergence of ordering effects at $L \geq 2$. At $d_x{=}8$ the mean spread is 0.099 ($L{=}2$), 0.076 ($L{=}4$), 0.100 ($L{=}8$); at $d_x{=}128$ it is 0.015, 0.046, 0.021, with the $L{=}8$ drop attributable to training instability (the default learning rate diverges for some seeds; $\leq 5 \times 10^{-4}$, and in two of five seeds 1×10^{-4}, is required at that scale). The spread is thus nonzero for all $L \geq 2$ but not monotone in L: the qualitative $L{=}1 \to L{\geq}2$ transition of Corollary 2 is confirmed, while its quantitative growth is not. This is as expected once Corollary 2 is read as a statement about the bound-level amplification ratio $W(1)/W(k)$: the realized spread additionally reflects task noise and training quality, which need not improve with depth (the $L{=}8$ instability above is one instance).

Bidirectional attention control. A strong falsifiable prediction of the causal-masking mechanism: bidirectional (non-causal) attention should exhibit *zero* ordering effects, since without the triangular mask every position has equal access to all others and $W(p)$ becomes flat. We trained bidirectional models across 8 configurations ($d_x \in \{8, 32, 128\}$, $L \in \{2, 4\}$, both attention variants, 3 seeds each) using the same architecture and hyperparameters as the causal experiments, differing only in the attention mask. All 24 bidirectional models exhibit ordering spread $\leq 2.4 \times 10^{-8}$—indistinguishable from numerical noise and 6–7 orders of magnitude below matched causal models (which range 0.009–0.105). This confirms that causal masking is the *sole* architectural source of ordering sensitivity: neither the embedding geometry, nor the training procedure, nor the layer composition introduces ordering effects without the triangular constraint. The result also validates the theory's $L{=}1$ prediction by a complementary route: $L{=}1$ causal models have zero spread because W is flat (Corollary 1), and bidirectional models have zero spread because the mask symmetry removes position dependence entirely—two distinct architectural reasons for the same outcome.

The coherence window and the $d_x{=}512$ anomaly. The regime variable is the certifiable per-pair coherence $c_{\min}(d_x, k)$ of Lemma 7, not the ratio d_x/k^2: under the radius schedule of §5 it takes the values 1.74, 0.90, 0.45, 0.23 at $d_x = 8, 32, 128, 512$ with $k{=}5$. The agreement between the cheap proxy sensitivity score and the empirical slot-sensitivity score, measured per permutation and averaged over sets (Figure A7), peaks at $d_x{=}128$ ($c_{\min} = 0.45$; $r_s = +0.50$), is near-zero at $d_x \leq 32$, where $c_{\min} \geq 0.90$ certifies essentially nothing, and drops again at $d_x{=}512$ ($r_s \approx +0.18$), where $c_{\min} = 0.23$.

The $d_x{=}512$ drop is not a training artefact but a *regime-intrinsic* phenomenon. Extended training (24 000 steps—4× the standard budget—with learning rates $5{\times}10^{-4}$ and 10^{-3}, 3 seeds each, both normalized and unnormalized; 12 runs total) shows no recovery: $r_s(\beta_\sigma, \text{gap})$ remains ≈ 0 throughout (mean -0.045, $\max|\cdot| = 0.25$) and the ordering spread stays at ~ 0.005, comparable to the $d_x{=}128$ level from the 72-configuration sweep. The mechanism is geometric: at $c_{\min} = 0.23$, example embeddings are near-orthogonal, so permutation-induced variations in the attention pattern collapse structurally—the slot interference that generates ordering effects (Theorem 5) vanishes. Agreement therefore occupies a *window* in $c_{\min}$, not a half-line: the coherence must be small enough for the geometry to be certifiable at all, and large enough for cross-source interference to still exist. This is the same coupling identified in §5 viewed along the dimension axis, and it is a property of the source geometry rather than of the bound: non-vacuity of (6) (Theorem 6) does not depend on d_x and fails at every dimension in this sweep.

The two readings of the certificate disagree sharply about how the window moves with k, which makes the account falsifiable, and we ran the discriminating sweep. Holding the *aggregate* budget $\varepsilon_0 = kc$ fixed gives $d_x = \Omega(k^2 \ln k)$ and, calibrated to the $k{=}5$ peak, places the window centre at $d_x = 38, 105, 151$ for $k = 3, 5, 6$. Holding the *per-pair* coherence c fixed—the reading the cancellation of §4.4 supports—gives $d_x = \Theta(\ln k)$ and places it at $d_x = 95, 105, 108$. The two predictions differ by a factor 2.5 at $k{=}3$ and, more usefully, disagree in *sign*: the k^2 law says the window moves strongly *right* with k, the $\ln k$ law says it barely moves at all.

R-EXP-11: the measured k-dependence. We trained a 3×5 grid, $k \in \{3, 5, 6\}$ against $d_x \in \{16, 32, 64, 128, 256\}$, 3 seeds per cell (45 models), holding every other hyperparameter at the values of §5 and measuring the same proxy-versus-empirical agreement statistic as Figure A7.

d_x	$c_{\min}$	k=3	k=5	k=6	training
16	1.26	$+0.02 \pm 0.05$	$+0.02 \pm 0.06$	-0.00 ± 0.03	converged
32	0.90	-0.03 ± 0.03	$+0.11 \pm 0.09$	$+0.06 \pm 0.07$	converged
64	0.64	$+0.07 \pm 0.03$	$+0.04 \pm 0.11$	$+0.10 \pm 0.10$	converged
128	0.45	$+0.27 \pm 0.09$	$+0.36 \pm 0.10$	$+0.44 \pm 0.04$	converged
256	0.32	$+0.39 \pm 0.02$	$+0.49 \pm 0.04$	$+0.46 \pm 0.02$	*diverged*

The d_x=256 row is reported but does not constrain anything: training diverges there at this budget (final loss 5.2 at k=3 and 1.6×10^3 at k=5, against $\approx$1.0 at $d_x \le 128$), so its agreement values are computed on models that did not fit the task. Split-half retest reliability of the statistic is 0.73–0.98 across all cells, so the cells that did converge are measuring something stable.

Two conclusions follow, and one does not. First, the $k^2 \ln k$ reading is refuted: it requires the k=3 curve to peak at $d_x \approx 38$ and be *falling* by d_x=64 while the k=6 curve is still rising past 150. Observed, the k=3 curve is still rising at d_x=128, its largest converged dimension, and the three curves lie within each other's standard errors at every dimension (cross-k spread 0.02–0.17 against a mean SEM of 0.05–0.08). Second, taking the dimension at which agreement crosses $+0.2$ as a location statistic gives $d_x = 100, 90, 78$ for $k = 3, 5, 6$, a ratio $d_x(k{=}3)/d_x(k{=}6) = 1.28$; the $k^2 \ln k$ law predicts 0.25 and the $\ln k$ law predicts 0.88. The observation is on the $\ln k$ side and has the opposite sign to the k^2 prediction. What does *not* follow is a measurement of the window centre: with training diverging at d_x=256, the sweep brackets the rise but not the maximum, so we report the k-dependence of the onset and not a peak location.

Honesty note on the sensitivity solver. In one cell of this grid (k=6, d_x=16, seed 1) the β_σ back-solve overflows for 38 of 10 950 permutation-slot entries (0.35%), which are dropped as non-finite before aggregation. The affected cell is at $c_{\min} = 1.28$, where the certificate is vacuous and the measured agreement is -0.00 regardless; no other cell is affected.

R-EXP-12: coherence manipulated at fixed dimension. The d_x sweep of §5.5 varies coherence and sample complexity together, so on its own it cannot attribute the collapse of ordering spread to either. We therefore manipulated source coherence *directly* at fixed d_x. For a target c, example centres are placed at $m_j = \sqrt{1-c}\, e_j + \sqrt{c}\, e_{k+1}$, so every pair has inner product exactly c by construction while k, d_x, depth, the weight initialisation, the task draw and the permutation set are held fixed by common random numbers. Each cell is 5 seeds $\times$ 24 source sets = 120 paired sets; the realised mean off-diagonal Gram matches the target to within 0.003 at every c. Contrasts below are paired within source set against the c=0 arm (one-sample t over the 120 pairs).

	spread(c) / spread(0), paired					
d_x	c=0 (abs.)	0.2	0.4	0.6	0.8	0.95
8	0.0138	1.15^{*}	1.32^{***}	1.73^{***}	$\mathbf{1.92}^{***}$	1.32^{*}
32	0.0043	1.21^{***}	1.66^{***}	$\mathbf{1.93}^{***}$	1.30	0.67^{*}
128	0.0012	0.66	0.65	0.52	0.60	0.38

$^{*}p < 0.05$, $^{***}p < 10^{-4}$; unmarked $p > 0.05$.

At $d_x \in \{8, 32\}$ coherence is causally sufficient to inflate ordering spread, and the response is non-monotone with an interior optimum: spread nearly doubles at c=0.8 (d_x=8, $p = 7\times10^{-11}$) and at c=0.6 (d_x=32, $p = 7\times10^{-10}$), then falls, dropping *below* the i.i.d. baseline at c=0.95 for d_x=32 ($0.67\times$, $p = 0.013$). This is the coherence window the theory predicts, established by manipulation rather than inferred from a dimension sweep: too little overlap and sources do not interfere, too much and they become interchangeable. At d_x=128 no contrast reaches significance ($p = 0.08$–0.33); the point estimates trend negative, and we report this as a null rather than a reversal.

What the d_x sweep actually measures. R-EXP-12 also shows that the d_x sweep is *not* a coherence experiment. Under i.i.d. sources the realised signed coherence is 0.026, 0.007 and 0.001 at $d_x = 8, 32, 128$—all below the c=0.2 arm, whose effect at d_x=8 is only $1.15\times$—so natural variation in coherence across the sweep is far too small to produce the $11\times$ drop in absolute spread observed over the same range. The sample-complexity account fits it instead. With k examples in d_x dimensions, an oracle least-squares predictor given the true source directions leaves residual fraction 0.38, 0.84,

0.96 and 0.99 of $\mathrm{Var}(y)$ at $d_x = 8, 32, 128, 512$ (measured on 10^4 draws at k=5), i.e. the in-context learnable fraction falls as $\approx k/d_x$. At d_x=128 there is almost no order-sensitive signal left to permute, which is also why manipulating coherence there changes nothing. We therefore attribute the sweep's collapse primarily to ICL signal loss and cite R-EXP-12, not the sweep, as the evidence for the coherence mechanism.

Structural content of the ranking. The bound's absolute magnitude exceeds 1 (Table A4: $\beta_\sigma \sim 10^4$ for trained models), as expected for worst-case spectral-norm bounds on neural networks. The relevant question is whether the *ranking* $\beta_{\sigma_1} < \beta_{\sigma_2}$ carries structural information. Three lines of evidence characterize this:

(1) Score agreement, and a weak but consistent gap correlation. The proxy sensitivity score agrees with the empirical slot-sensitivity score at $r_s = +0.50$ (d_x=128, normalized; Figure A7)—a non-trivial agreement for a worst-case score on a learned model with no tuning. Correlations of either score with the realized generalization gap are weaker: across the 24-configuration, 72-run sweep of Section 5.5, the pooled theory–gap correlation is $\bar{r}_s = -0.00 \pm 0.16$ ($p = 0.84$), while the empirical-sensitivity–gap correlation is weak but consistent ($\bar{r}_s = +0.08 \pm 0.19$, $p < 0.001$). The ranking does not predict realized gaps ($\bar{r}_s \approx 0$)—consistent with the many-body emergence result of §5.4. The remaining two arguments characterize what the ranking *does* capture.

(2) Structural argument. The absolute vacuity stems from spectral norm bounds $B \approx 3.5$ compounding as $\rho^L \sim 10^{16}$ for L=4. However, this multiplicative constant is *shared across all permutations*—$\beta_\sigma = C_\ell G \cdot \rho^{L-1} \cdot f(\sigma)$, where $f(\sigma)$ captures the permutation-dependent part via $\sum_p W(p)^2 \eta^2_{\sigma(p)}$. The ranking of β_σ across permutations depends only on $f(\sigma)$, not on the shared (vacuous) prefix.

(3) Consistent greedy benefit. The greedy algorithm, which minimizes β_σ over orderings, outperforms random in $5/5$ seeds under normalized attention. If the ranking were meaningless, this consistency would occur with probability $\leq 2^{-5} = 0.03$.

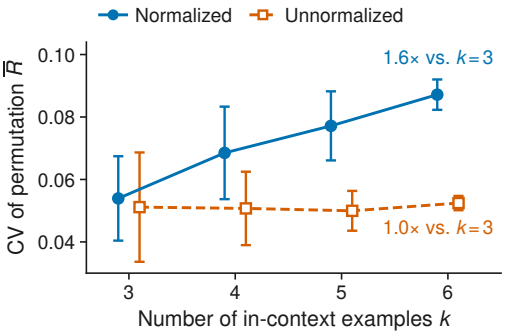


Figure A6: Coefficient of variation of permutation-level $\bar{R}$ vs. k (n=5 seeds, mean ±s.e.m.). Normalized attention shows genuine scaling (1.6× increase); unnormalized is flat.

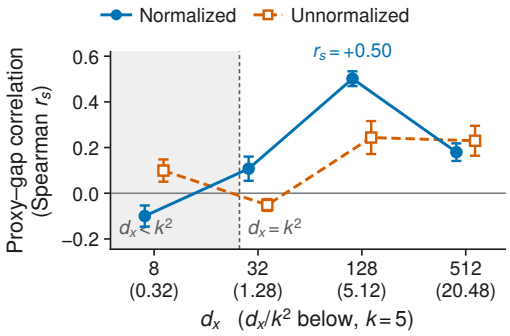


Figure A7: Per-permutation Spearman correlation between the proxy sensitivity score and the empirical slot-sensitivity score, across $d_x \in \{8, 32, 128, 512\}$ (k=5, 5 seeds; mean ± s.e.m.). Under normalized attention, proxy agreement peaks at d_x=128 ($c_{\min}$=0.45, r_s= + 0.50) and degrades on both sides of the coherence window; under unnormalized attention the peak is far weaker (r_s= + 0.24) and does not fall at d_x=512.

Bound constants. Table A4 reports the model constants B (max spectral norm of W_Q, W_K, W_V), R (max representation norm), and the resulting ρ, γ for trained models ($d_x = 32$, $L = 3$, 5 seeds per variant), reported for both attention normalizations. The bound exceeds 1 in absolute magnitude for all configurations ($\beta_\sigma \gg 1$), because ρ and γ grow cubically in B and quadratically in R—standard for worst-case stability bounds on neural networks with unconstrained weights. The bound's content is structural: the qualitative predictions (W monotonicity, L=1 invariance, depth scaling) are confirmed

experimentally, and the proxy sensitivity score agrees with empirical slot sensitivity ($r_s = +0.50$ at $d_x = 128$; Figure A7). The permutation ranking is mechanistic rather than predictive (Table A9), consistent with the many-body emergence of the example factor (§5.4).

Table A4: Trained model constants (d_x=32, L=3, 5 seeds, mean $\pm$ SD). Absolute bound exceeds 1 ($\beta_\sigma \gg 1$); structural content analyzed in text.

Variant	B	R	ρ	γ	G	$W(1)/W(k)$
Normalized	3.5 ± 1.2	10.1 ± 4.0	2.3×10^4	1.5×10^4	2.2 ± 0.2	1.94
Unnormalized	1.7 ± 0.4	5.4 ± 1.3	1.1×10^3	2.7×10^2	1.3 ± 0.2	2.12

Full candidate score results. Table A5 reports Spearman correlations between each of the 17 candidate example scores and the empirical example factor v_j extracted from the gap matrix. All candidates fail to identify the true example factor.

Table A5: Spearman correlation of candidate example scores with the empirical example factor v_j, pooled across example sets ($d_x = 128$, normalized attention; data-side: 5 seeds, model-side: 2 seeds). None achieves $|r_s| > 0.21$.

Type	Candidate score	Pooled r_s
Data-side	Source diameter η^2	-0.07
	Angular radius	-0.07
	$\lvert y\rvert$ (label magnitude)	$+0.05$
	y^2	$+0.04$
	Noise realization	$+0.03$
	Eccentricity	-0.04
	Pairwise similarity $\langle x_j,\, x_l\rangle$	-0.08
	Resampling displacement (x)	-0.08
	Resampling displacement (y)	-0.04
	Resampling displacement (embedding)	-0.10
	Pool y std. dev.	-0.08
	y deviation from source mean	$+0.02$
Model-side	Sensitivity column mean	$+0.07$
	Sensitivity SVD	$+0.00$
	Leave-one-out prediction shift	$+0.17$
	Attention proportion	$+0.07$
	Jacobian norm	$+0.12$

Greedy ordering: cost accounting and greedy-seeded sampling. Table 2 originally priced greedy in forward passes and the sampling baselines in ordering evaluations, two units that cannot be compared. We instrumented the model call and counted *prompt-forwards* (one forward pass over one $(k{+}1)$-token prompt) for every strategy at the operating point of that table (k=5, d_x=32, L=3, n_r=8, 8 queries). Greedy makes $k(k{+}1)/2 - 1 = 14$ slot-sensitivity probes, each costing $(1 + n_r)$ prompt batches, for 1008 prompt-forwards; scoring one candidate ordering on the same 8 queries costs 8. Greedy therefore costs 126 ordering evaluations, against 120 to enumerate all 5! permutations. The two costs cross at k=6 and diverge quickly thereafter (k=6: 180 vs. 720; k=8: 315 vs. 40,320; k=10: 486 vs. $\sim 3.6 \times 10^6$), so the algorithm's cost advantage begins one step above the k at which exhaustive ground truth is affordable. The paper's main sensitivity experiments use n_r=16, which would raise the greedy figure to 255 evals; the n_r=8 value is the one this table's run used.

Table A6 asks whether greedy is nonetheless *complementary* to sampling: is one of m evaluation slots better spent on the greedy order than on another random draw? The two arms share the same $m{-}1$ random permutations (common random numbers), so the contrast is paired within trial as well as within set, and both arms spend exactly m evals—greedy's own search cost is not charged, which makes this the most favourable framing available to it. Under normalized attention seeding helps only at the smallest budget and the effect decays fast; the Wilcoxon p-values at $m \geq 10$ are driven by a small number of large gains while the majority of sets are made slightly worse, so the help/hurt split

rather than the p-value describes the practical effect. Under unnormalized attention, where greedy is itself below random, seeding is harmful at every budget. Charging greedy's search cost removes even the $m{=}2$ gain at this k: a seeded budget of m costs $125 + m$ evals, and any budget above 120 can simply enumerate.

One measurement detail matters here. The stored per-set summaries record a greedy $\bar{R}$ computed on its own resample draws, which differs from the greedy *order*'s risk on the shared CRN pool used for all 120 permutations by up to 1.1×10^{-2} per set (mean 2.5×10^{-3}) against a random-to-oracle gap of 1.4×10^{-2}. The discrepancy is zero-mean, so it moves the aggregate greedy row of Table 2 by less than 10^{-5} in $\bar{R}$ and 0.001 in gap capture. It is not harmless here, because $\min$ is nonlinear: a noisy greedy value wins the minimum precisely when its noise is negative, inflating the seeded arm. Table A6 therefore scores the greedy order on the shared pool; on the stored values the $m{=}2$ effect would read $+0.109$ instead of $+0.084$, and the unnormalized arm would appear to help rather than hurt.

Table A6: Greedy-seeded sampling against plain random-m at equal *evaluation count* (greedy's search cost not charged), 100 example sets per variant $\times$ 4000 paired trials, common random numbers. Δ capture is seeded minus plain in units of the random-to-oracle gap; help/hurt counts sets whose paired mean improved or worsened. p is a Wilcoxon signed-rank test over sets.

Variant	m	seeded $\bar{R}$	plain $\bar{R}$	Δ capture	help/hurt	p
Normalized	2	0.11939	0.12060	+0.084	60/40	0.005
	3	0.11779	0.11840	+0.043	53/47	0.16
	5	0.11601	0.11624	+0.016	38/62	0.40
	10	0.11408	0.11414	+0.004	26/74	0.002
	20	0.11273	0.11275	+0.002	15/85	2×10^{-6}
Unnormalized	2	0.12444	0.12396	−0.037	47/53	0.80
	3	0.12206	0.12179	−0.021	41/59	0.24
	5	0.11985	0.11972	−0.009	32/68	0.012
	10	0.11793	0.11789	−0.003	22/78	8×10^{-6}
	20	0.11675	0.11673	−0.002	11/89	7×10^{-10}

Depth sweep details. Table A7 reports the full depth sweep results (unnormalized attention, 5 seeds per configuration; the $d_x{=}32$ column has 2 seeds at $L{=}4$ only). The $L{=}1$ prediction of zero ordering effects is confirmed to numerical precision across all embedding dimensions.

Table A7: Ordering spread (mean $\bar{R}$ range across 120 permutations, averaged over 5 example sets and 5 seeds). At $L{=}8$ training is unstable at the default learning rate; reduced rates (10^{-4}–$5{\times}10^{-4}$) were used, which may underestimate the true ordering spread.

Depth	$d_x = 8$	$d_x = 32$	$d_x = 128$
$L = 1$	0.0000	—	0.0000
$L = 2$	0.099	—	0.015
$L = 4$	0.076	0.043	0.046
$L = 8$	0.100	—	0.021

Multi-k scaling details. Table A8 and Figure A6 report the coefficient of variation (CV) and fixed-sample-size range for both attention variants across $k \in \{3, 4, 5, 6\}$ ($d_x = 32$, $L = 3, 5$ seeds).

Table A8: Scaling of ordering effects with k (mean $\pm$ SEM, n=5 seeds). CV: coefficient of variation of permutation-level $\bar{R}$. Fixed-n range: max$-$min over n=6 randomly sampled permutations (for k=3, n=6=k! so this equals the full range).

	Unnormalized		Normalized	
k	CV	Fixed-n range	CV	Fixed-n range
3	0.051 ± 0.018	0.017 ± 0.006	0.054 ± 0.014	0.016 ± 0.004
4	0.051 ± 0.012	0.018 ± 0.005	0.069 ± 0.015	0.020 ± 0.005
5	0.050 ± 0.006	0.016 ± 0.002	0.077 ± 0.011	0.025 ± 0.005
6	0.052 ± 0.002	0.018 ± 0.002	0.087 ± 0.005	0.034 ± 0.005

Table A9: Per-regime rank correlations from the 24-configuration sweep (72 runs; $k \in \{3, 5\}$, $L \in \{2, 4\}$, $d_x \in \{8, 32, 128\}$, normalized/unnormalized, 3 seeds each; 6,000 training steps, all k! permutations evaluated). r_s: Spearman correlation between per-permutation scores and per-permutation generalization gaps (evaluation loss per permutation; Spearman ranks are invariant to the constant population-risk offset), averaged over sets and seeds (mean $\pm$ SD across seeds); spread: max$-$min of $\bar{R}$ across permutations. Pooled over all 72 runs: theory $\bar{r_s} = -0.00 \pm 0.16$ ($p = 0.84$), empirical $\bar{r_s} = +0.08 \pm 0.19$ ($p < 0.001$).

k	L	d_x	variant	r_s (theory)	r_s (empirical)	spread
3	2	8	norm	0.154 ± 0.078	$+0.061 \pm 0.100$	0.0205
3	2	8	unnorm	0.091 ± 0.118	$+0.234 \pm 0.221$	0.0157
3	2	32	norm	0.067 ± 0.026	$+0.219 \pm 0.157$	0.0061
3	2	32	unnorm	-0.027 ± 0.143	$+0.238 \pm 0.153$	0.0047
3	2	128	norm	0.192 ± 0.099	$+0.057 \pm 0.255$	0.0044
3	2	128	unnorm	-0.042 ± 0.135	$+0.067 \pm 0.210$	0.0032
3	4	8	norm	0.055 ± 0.111	$+0.120 \pm 0.238$	0.0342
3	4	8	unnorm	-0.006 ± 0.215	$+0.112 \pm 0.269$	0.0275
3	4	32	norm	-0.246 ± 0.186	$+0.051 \pm 0.206$	0.0124
3	4	32	unnorm	-0.295 ± 0.207	$+0.156 \pm 0.099$	0.0056
3	4	128	norm	-0.063 ± 0.078	-0.044 ± 0.294	0.0027
3	4	128	unnorm	0.133 ± 0.174	-0.139 ± 0.057	0.0034
5	2	8	norm	-0.005 ± 0.082	$+0.157 \pm 0.104$	0.0697
5	2	8	unnorm	0.054 ± 0.014	$+0.096 \pm 0.129$	0.0437
5	2	32	norm	0.052 ± 0.060	$+0.057 \pm 0.065$	0.0129
5	2	32	unnorm	0.014 ± 0.078	$+0.114 \pm 0.096$	0.0061
5	2	128	norm	-0.027 ± 0.036	$+0.104 \pm 0.072$	0.0066
5	2	128	unnorm	0.026 ± 0.025	$+0.044 \pm 0.151$	0.0069
5	4	8	norm	0.012 ± 0.153	-0.078 ± 0.025	0.1048
5	4	8	unnorm	0.051 ± 0.129	$+0.113 \pm 0.204$	0.0737
5	4	32	norm	-0.109 ± 0.113	-0.041 ± 0.036	0.0154
5	4	32	unnorm	-0.143 ± 0.049	$+0.151 \pm 0.094$	0.0112
5	4	128	norm	0.064 ± 0.090	$+0.138 \pm 0.062$	0.0049
5	4	128	unnorm	-0.099 ± 0.111	-0.011 ± 0.049	0.0084