When Does Polynomial Attention Concentrate? A Relative-Margin Diagnostic for Zero-Shot Softmax Substitution

Anonymous Authors[1]

Abstract

We study zero-shot pointwise substitution of softmax attention in pretrained LMs, where the score row is kept fixed and only the row map is replaced. For normalized polynomial attention of active-branch degree p, the target receives mass $1/(1+S_p)$ with $S_p = \sum_{j \neq *}(r_j/r_*)^p$, and the criterion $S_p < 1$ is exactly the recall criterion under aligned-binary value vectors. Softmax is uniquely characterized among continuous-positive pointwise normalized maps by additive row-shift invariance; polynomial attention instead concentrates by relative margin $(1 - \Delta/r_*)^p$, yielding a worst-case sufficient degree $p^\star = \log A / \log(1/(1-\rho))$ for A active distractors at relative margin ρ — in contrast to softmax's absolute-margin envelope $1/(1+Ae^{-\beta\Delta})$. Auditing four small open-weight decoder LMs at contexts up to 4096, three of four cross 70 % unsafe active-unsaturated rows on every diagnostic suite at $p = 2, b = 0$; a 14-recipe substitution sweep finds no universal drop-in replacement, while on an uncontaminated NIAH probe two of the four admit sample-perfect polynomial recipes. We position S_p as a no-go signal: in our tested cells, very low measured-safe fraction was a sharp negative screen, but high measured-safe does not guarantee preservation, so a behavioral substitution test remains necessary.

1. Introduction

A long line of efficient-attention proposals replaces the softmax score map with cheaper pointwise alternatives: ReLU, ReLU^2, higher-degree polynomials, kernelized features, sigmoid, sparse projections (Wortsman et al., 2023; Saratchandran et al., 2024; Kacham et al., 2024; Letourneau et al., 2025; Choromanski et al., 2021; Katharopoulos et al., 2020; Zhang et al., 2024; Ramapuram et al., 2025; Boufadène and Vialard, 2025). Most of these papers ground their case on training stability, kernel approximation, or wall-clock cost, and report that the resulting model trains successfully. They typically do not give a row-level guarantee on concentration: how much aggregate mass the attention head can place on a target key given a margin against its distractors.

Concentration is the property that distinguishes softmax. Any strictly increasing pointwise map preserves the rank of the highest-scoring key. The harder question is whether it can place enough mass on that key to dominate the value sum, even when many weaker distractors collectively compete for mass. A row can have the right top-1 key and still produce the wrong output when the aggregate distractor weight exceeds the target weight.

This paper combines a theoretical diagnostic for the polynomial substitution question with an empirical measurement of where four small open-weight LMs sit relative to it. We give a one-line criterion that is computable from a single forward pass and equivalent to correct output on a binary recall row, and then we measure it on Llama-3.2-1B (Meta AI, 2024), Qwen3-1.7B (Qwen Team, 2025), Gemma-4-E2B (Google DeepMind, 2026), and Pythia-1.4b (Biderman et al., 2023) across induction, multi-query associative recall, and a natural-text completion suite at contexts up to 4096 tokens.

Contributions. (1) A row-level diagnostic and concentration theory. For pointwise normalized attention with degree-p active-branch polynomial map, the target mass is exactly

$$\widetilde{a}_* = \frac{1}{1+S_p}, \qquad S_p = \sum_{j \neq *} \left(\frac{r_j}{r_*}\right)^p, \tag{1}$$

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

and the strict condition $S_p < 1$ is exactly the binary aligned-distractor recall criterion (Remark 4.5). Softmax is uniquely characterized among pointwise normalized maps by additive-shift invariance (Theorem 4.1); polynomial attention concentrates by relative margin $(1-\Delta/u_*)^p$ rather than absolute margin $e^{-\beta\Delta}$, yielding a worst-case degree threshold $p^\star = \log A/\log(1/(1-\rho))$ (Corollary 4.3). Among nonnegative polynomial mixtures of maximum degree p, the pure monomial r^p pointwise maximizes target mass on every row (Theorem 4.4).

(2) A four-model audit of the unsafe regime and a 14-recipe substitution sweep. On four small open-weight LMs (Llama-3.2-1B, Qwen3-1.7B, Gemma-4-E2B, Pythia-1.4b) at $p=2, b=0$, three of four models cross 70 % unsafe rows on every diagnostic suite (Table 1); Llama is the polynomial-safe outlier on the diagnostic. Replacing softmax with ReLU^2 drops next-token accuracy by 44.5–79.0 pp on padded literary-completion prompts. Among 10 pointwise-normalized recipes, no single (p, b) recovers softmax across all four models (Table 2); the four un-normalized Wortsman et al. (2023) recipes are reported as literature-grounded references and uniformly collapse.

(3) Connecting the diagnostic to behavior on uncontaminated NIAH and a stress check on a harder retrieval probe. A NIAH replication where softmax baselines are sample-perfect (200/200) yields three sample-perfect polynomial cells—relu_p4_bauto and relu_p8_bauto on Llama, relu_p8_bauto on Gemma-4 (Table 3); Qwen3 has no sample-perfect recipe (its best cell, relu_p16_b0, still drops −50.5 pp); Pythia collapses on every tested recipe. The Llama partition replicates at 1, 3, 8 B scale (Appendix AC) and at $n = 500$ on a fresh seed (Appendix AG). Head-level ablations show that ranking heads by per-head S_p identifies behaviorally critical heads on Qwen3 and Gemma-4; Llama is robust to the selective substitutions we test, and Pythia's ranking does not discriminate (Figures 3 and 4). An exploratory RULER-template niah_single_3-style probe at ~ 4 K tokens with a uniformly post-cutoff Dwarkesh haystack reveals a degree gradient on Llama-1B: relu_p4_bauto collapses to 4 % and relu_p8_bauto partially recovers to 54 % (softmax baseline 76 %). The matched diagnostic on the same prompts moves in the same direction (65 % → 29 % unsafe at $p = 4 \to 8$, Appendix AI), providing a directional stress check of the degree-scaling $p^\star = \log A/\log(1/(1-\rho))$ from Corollary 4.3. Thus S_p is a conservative pre-deployment screen for zero-shot pointwise substitution, not a guarantee.

2. Related Work

Softmax has a long lineage in choice theory and statistics (Luce, 1959; McFadden, 1974; Yellott, 1977; Bridle, 1990; Blondel et al., 2020; Aczél, 1966); Theorem 4.1 is an attention-native re-presentation. Hahn (2020); Chiang and Cholak (2022); Sanford et al. (2023); Veličković et al. (2025); Nakanishi (2025); Boursier and Boyer (2025); Vasylenko et al. (2026) study softmax's length and dispersion limits; our Corollary 4.3 is in the same flavor as Veličković et al. (2025)'s dispersion result, applied to the polynomial-substitution case.

Pointwise polynomial and ReLU replacements include Wortsman et al. (2023); Saratchandran et al. (2024); Zhang et al. (2024); Kacham et al. (2024); Letourneau et al. (2025); Qin et al. (2022); non-pointwise alternatives include sigmoid (Ramapuram et al., 2025; Yan et al., 2025), stick-breaking (Tan et al., 2025), Softpick (Zuhri et al., 2025), Differential Transformer (Ye et al., 2025), Selective Attention (Leviathan et al., 2025), and the sparsemax/entmax family (Martins and Astudillo, 2016; Peters et al., 2019; Correia et al., 2019). State-based polynomial mechanisms (Power Attention (Gelada et al., 2025), Titans (Behrouz et al., 2025)) and learned feature maps fall outside our pointwise scope; we test only zero-shot pointwise substitution. Appendix A gives the extended discussion.

3. Setup

Let $z \in \mathbb{R}^n$ be a per-position score row. A pointwise normalized attention map with score map $\varphi : \mathbb{R} \to [0, \infty)$ is

$$A_\varphi(z)_i = \frac{\varphi(z_i)}{\sum_j \varphi(z_j)}, \qquad A_{\text{softmax}}(z)_i = \frac{e^{z_i}}{\sum_j e^{z_j}}. \quad (2)$$

Throughout the paper $*$ indexes the target key (the one whose value should dominate the row's output). Define the absolute margin $\Delta = z_* - \max_{j \neq *} z_j$, the threshold-shifted score $u_* = z_* + b$ for a bias $b \in \mathbb{R}$, and the clipped score $r_j = \min(\max(z_j + b, 0), \tau)$ with optional cap τ.

For active-branch degree-p polynomial $\varphi(z) = r^p$ with $r = \max(z + b, 0)$ (no cap), the target mass admits the row-normalization identity

$$A_\varphi(z)_* = \frac{r_*^p}{r_*^p + \sum_{j \neq *} r_j^p} = \frac{1}{1 + S_p}. \quad (3)$$

A row is active-unsaturated if $r_* > 0$ and (when applicable) $u_* < \tau$, dead if $r_* = 0$, and saturated if $u_* \geq \tau$; we restrict diagnostics to active-unsaturated rows. The contribution is the framing of S_p as a single-forward-pass, label-free diagnostic: $S_p < 1$ exactly determines

whether the target receives $> 1/2$ of the row mass. Unless otherwise noted $*$ is the row argmax (argmax-concentration diagnostic); the task-target variant sets $*$ to the actual answer key, for which $\Delta \geq 0$ need not hold (only 3–7 % of NIAH layer/head/prompt triples place the needle as the row argmax, Appendix S). Dead-row handling and full substitution recipe in Appendix B.

4. Concentration theory

Proofs and extended commentary are in Appendix K.

Theorem 4.1 (Shift-invariance characterization). Let $\varphi : \mathbb{R} \to (0, \infty)$ be continuous and let $A_\varphi^{(n)}$ $(n \geq 2)$ be the pointwise normalized map. $A_\varphi^{(n)}$ is invariant to additive shifts $z \mapsto z + c\mathbf{1}$ for every c iff $\varphi(z) = Ce^{\beta z}$. The strictly-increasing case requires $\beta > 0$.

Theorem 4.2 (Margin response controls concentration). Let $g : [0, \tau) \to [0, \infty)$ be nondecreasing with $g(0) = 0$ and $g(r) > 0$ for $r > 0$. Assume target is score-maximal $(\Delta \geq 0)$ and active-unsaturated $(r_* \in (0, \tau))$. If $\Delta \geq r_*$, every distractor is inactive and $A_g(z)_* = 1$. Otherwise, for $0 \leq \Delta < r_*$, with $A = |\{j \neq * : r_j > 0\}|$ and $R_g(r_*, \Delta) = g(r_* - \Delta)/g(r_*)$,

$$A_g(z)_* \;\geq\; \frac{1}{1 + A\, R_g(r_*, \Delta)}.$$

For homogeneous $g(r) = Cr^p$, this gives the relative-margin law

$$A_g(z)_* \;\geq\; \frac{1}{1 + A\,(1 - \Delta/r_*)^p}. \tag{4}$$

For exponential $\varphi(z) = e^{\beta z}$ on $\mathbb{R}$, the analog is $1/(1 + Ae^{-\beta\Delta})$.

Corollary 4.3 (Worst-case sufficient degree). For the active uncapped branch $\varphi(z) = \max(z + b, 0)^p$ with $r_* = u_* > 0$ (so $\rho = \Delta/u_* = \Delta/r_*$): to guarantee $A_\varphi(z)_* \geq 1/2$ against $A \geq 1$ active distractors at $\rho \in (0, 1)$, it suffices that $p \geq \log A / \log(1/(1-\rho)) \approx (u_*/\Delta) \log A$ for small ρ. The actual S_p on a real row may fall below this envelope when distractors are spread (slack), so the bound has low false-negative and high false-positive rates as a screen.

Theorem 4.4 (Pointwise monomial dominance). For $\varphi(r) = \sum_{q=0}^{p} c_q r^q$ with $c_q \geq 0$ and $c_p > 0$, on any row with $r_* > 0$ and $0 \leq r_j \leq r_*$ for all $j \neq *$, $S_\varphi \geq S_p$, where $S_\varphi = \sum_{j \neq *} \varphi(r_j)/\varphi(r_*)$. Thus the pure monomial r^p pointwise maximizes target mass among nonnegative polynomial mixtures of maximum degree p.

Remark 4.5 (Binary aligned-distractor recall criterion). Under aligned-binary values $(v_* = +1,\ v_j = -1)$, the row output has the correct sign iff $S_p < 1$. This is one line from Equation (3); it is a surrogate criterion under aligned-binary value vectors. The connection to behavior on real transformers comes through the head-ablations and substitution sweeps below, not the remark alone.

The relative-margin law (4) is the polynomial counterpart of softmax's $1/(1 + Ae^{-\beta\Delta})$ envelope; the suppression factor depends on the ratio Δ/u_* rather than the absolute margin Δ, so real heads with u_* comparable to Δ require degree to rise rapidly with A (a side-by-side comparison of softmax/polynomial/sigmoid/Wortsman concentration envelopes is in Table 7 in Appendix L).

Lower-bound audit. On the $L = 4096$ induction parquet at $b = 0$, the exact threshold $p^\star = \log A / \log(1/(1-\rho))$ acts as a conservative screen: at $p = 2$, predicted-vs-measured agreement is 0.59/0.96/0.84/1.00 on Llama/Qwen3/Gemma/Pythia, with strict-boundary false negatives of 10/0/3/1 rows out of 40–100 k active-unsaturated rows per cell (all disappear under tolerance 10^{-6}). False positives rise on safer models—41,666 on Llama where the actual S_p falls well below the worst-case envelope. The asymmetry is exactly what a worst-case envelope should show (Figure 1; further classification breakdown in Appendix M).

5. Empirical results

5.1. Setup

We evaluate four open-weight decoder LMs (Llama-3.2-1B, Qwen3-1.7B, the text decoder of multimodal Gemma-4-E2B (Google DeepMind, 2026), Pythia-1.4b) on three prompt suites (induction (Olsson et al., 2022), MQAR (Arora et al., 2023), and a literary-completion "copying" suite that we caveat is contaminated by memorized openings). All loads use bfloat16 + eager attention, and the diagnostic hook is bit-identical to the vanilla forward pass. We compute, for each layer/head/query-position/prompt, S_p, the relative margin Δ/u_*, the active-distractor count A_{act}, and a status label, and we report results at $b = 0$ and a calibrated per-model b_{auto} (5 %-mass budget; ranges from -3.36 for Qwen3 to $+119.75$ for Pythia, where the rule degenerates due to score-scale differences — b_{auto} is a diagnostic operating point, not a universal recipe). The substitution sweep replaces every softmax row with one of 14 pointwise recipes: $\text{ReLU}(z+b)^p$ at $p \in \{2, 4, 8, 16\}$ and $b \in \{0, b_{\text{auto}}\}$ (8 cells), pointwise-normalized sigmoid at $b \in \{0, b_{\text{auto}}\}$ (Ramapuram et al., 2025), and four un-normalized recipes from Wortsman et al. (2023). The group ablation polynomializes

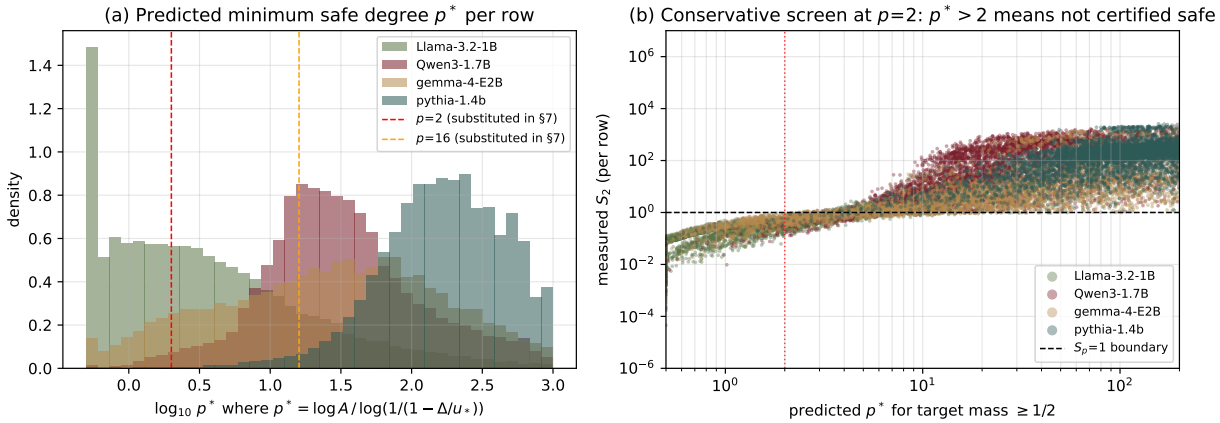

Figure 1. Lower-bound validation. (a) Distribution of predicted minimum safe degree $p^\star$ per row across the four models, on a $\log_{10}$ axis; dashed verticals mark substituted degrees $p \in \{2, 16\}$. Pythia's mass is concentrated at $p^\star \in [10^2, 10^3]$, beyond every tested p. (b) Predicted vs measured at $p = 2$: rows with $p^\star \leq 2$ are certified safe by Cor. 4.3; rows with $p^\star > 2$ are predicted unsafe by the conservative screen but may still be measured-safe ($S_2 < 1$) when the actual distractor distribution does not attain the worst-case margin. The data hugs the $S_2 = 1$ boundary tightly for the unsafe models (Pythia, Qwen3) and shows large conservative slack for safer models (Llama, Gemma-4) — the expected asymmetry for a worst-case envelope.

top/bot/random-K heads ranked by per-head median S_p. With $n = 200$ trials per cell, the Wilson MDE at baseline 0.78 is ≈ 10 pp; reported deltas of headline cells are $\gg 30$ pp. Full architecture counts, calibration procedure, numerical-stability analysis, prompt construction, and Gemma softcap details are in Appendix B.

5.2. Prevalence of the unsafe regime

Table 1 reports the polynomial-unsafe rate (fraction of rows with $S_p \geq 1$) at the longest context across all three suites and all four models, restricted to active-unsaturated rows. At $p = 2$, three of the four models have at least 70 % of active-unsaturated rows in the unsafe regime on every suite. The unsafe rate decreases at higher degree but does not vanish: at $p = 16$, Pythia stays above 80 % on induction/MQAR and Qwen3 above 20 % on induction (full numbers in Table 10; Table 1 only shows $p \in \{2, 4, 8\}$).

At $b = 0$, Llama is the polynomial-safe outlier on every diagnostic axis (median relative margin 0.73 vs 0.21/0.14/0.03 for Qwen3/Gemma/Pythia; median $A_{\text{act}} = 2$). The empirical phase plot in Figure 2 overlays per-head medians on the relative-margin/active-distractor plane against the boundaries from Equation (4): three of four models sit deep above the $p = 2$ boundary; Llama's heads cluster on the rightward (safer) side. At $b = b_{\text{auto}}$, Llama's unsafe rate (71 %) approaches Qwen3's (70 %) and Gemma-4's (62 %) (Table 11). Length generalization is reported in Appendix C: Pythia's unsafe rate climbs from ~ 66 % at the in-distribution $L = 1024$ to 99 % at $L = 4096$ (OOD), while Llama and Qwen3 stay essentially flat across $L \in [256, 4096]$.

5.3. Substitution sweep

The substitution recipe is a single per-row replacement of softmax with $\text{ReLU}(z + b)^p / \sum_j \text{ReLU}(z_j + b)^p$ (Algorithm 1); applied at every layer and every query position with a calibration step that picks b_{auto} as the smallest threshold for which a held-out generic-text input has at most 5 % of softmax mass below score $-b$. Table 2 reports next-token accuracy deltas at the longest padding (pad = 500, ~ 500 tokens) for the 14-recipe sweep, vs softmax baseline 74.5–79.0 %.

Algorithm 1 Pointwise polynomial substitution recipe (per row)

Input: pre-softmax score row z, degree p, bias b
$r \leftarrow \max(z + b, 0)$ {post-softcap if applicable; Gemma's tanh softcap is upstream}
$\widetilde{a} \leftarrow r^p / \sum_j r_j^p$ {un-normalized Wortsman variants skip the denominator}
return $\widetilde{a}$

ReLU^2 model/bias cells lose 44.5–79.0 pp; higher-degree calibrated cells are model-specific (Llama: −21 pp at relu_p4_bauto; Gemma: −5 pp at relu_p8_bauto; Qwen3: −14.5 pp at relu_p16_b0; Pythia: ≥ -65 pp on every variant). Among the 10 pointwise-normalized recipes, no single (p, b) recovers softmax on all four models; the four un-normalized

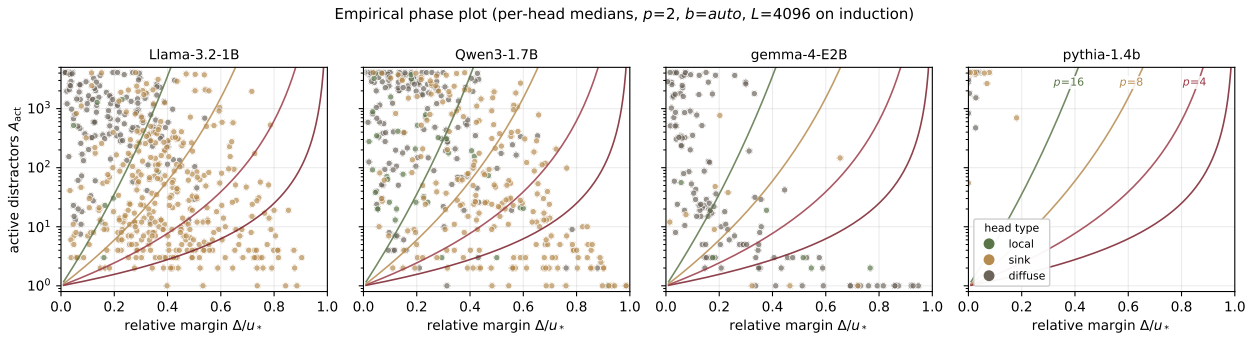

Figure 2. Empirical phase plot at $p=2$, $b=b_{\text{auto}}$, $L=4096$ on the induction suite. Each marker is the median $(\Delta/u_*, A_{\text{act}})$ of one (layer, head) pair across active-unsaturated rows (head type per Appendix H). Solid curves are rowwise concentration boundaries $A_{\text{act}}(1-\Delta/u_*)^p=1$ at $p\in\{2,4,8,16\}$ from Equation (4); per-head medians overlay these for visualization, so a median above the p curve indicates $\geq 50\,\%$ unsafe rows at p rather than a per-row certificate.

Wortsman variants are included as literature-grounded references (originally proposed for ViTs, not LMs) and uniformly collapse, as expected for recipes lacking a row normalizer. The empirical takeaway matches Equation (4): sharp rows (low A_{act}, high Δ/u_*) tolerate polynomialization, diffuse rows do not. The literary suite is contaminated by memorized openings; Section 5.5 replicates these findings on a clean NIAH probe.

5.4. Group head ablation

To turn the prevalence finding into a causal handle on which heads matter, we perform a per-head group-ablation experiment. Ranking all heads by their per-head median S_p (computed from the diagnostic forward pass on the induction suite at $L = 4096$ with $p=2$ and $b=b_{\text{auto}}$), we polynomialize only the top-K, only the bottom-K, or a random-K subset and measure copying accuracy on 200 pad=500 prompts.

Figure 3 partitions the models into four regimes. Qwen3 and Gemma-4 are decisive (top-K strictly worse than every random-K draw at $K \geq 20$ and $K \leq 40$ respectively); Llama is robust to the selective substitutions we test; Pythia's ranking does not discriminate, because polynomializing low-S_p heads is near-no-op by construction (Remark 5.1) and Pythia's bottom-ranked heads still sit in the unsafe regime. The architectural pattern is robust: a $p = 4$ rerun preserves the ordering on all four models (Appendix N), and a 20-draw random-K rerun preserves decisiveness on the two decisive models (Appendix V). Appendix U shows the top-K heads are predominantly diffuse (85/88/100/73 % on Llama/Qwen3/Gemma/Pythia at $K = 40$) while bot-K are predominantly sink, mechanically connecting the S_p ranking to the sink/local/diffuse taxonomy of Figure 2. Per-model details are in Appendix D.

Remark 5.1 (bot-K as a no-op sanity check). A head with $S_p \ll 1$ already places nearly all softmax mass on a small set of keys, so polynomialization is close to output-preserving. The empirical content of Figure 3 is therefore the top-K curve and its separation from the random-K cloud.

5.5. Needle-in-haystack experiments

We replicate the substitution sweep and group ablation on a needle-in-haystack (NIAH) suite where the needle is randomly chosen per prompt ("The secret word is X. [haystack] The secret word is"; details and audit in Appendix E). Softmax baselines are sample-perfect (200/200) on every model at every pad $\in\{0, 8, 32, 128\}$ (longest prompts ~ 1500 tokens); the two-sided Wilson 95 % interval for 200/200 is [98.2 %, 100 %], and a 0/200 cell has a Wilson 95 % upper bound of $\approx 1.9\,\%$.

Table 3 reports the substitution sweep at pad = 128. Llama's relu_p4_bauto and relu_p8_bauto, and Gemma's relu_p8_bauto, are sample-perfect (0 pp drop); Qwen3 admits no preserving recipe—its best cell, relu_p16_b0, still drops −50.5 pp—while Pythia collapses on every tested recipe; all four un-normalized Wortsman variants are uniformly −100 pp. NIAH sharpens both directions of the literary-suite findings: every preserving cell improves to 0 pp drop (vs −5 pp on Gemma's literary relu_p8_bauto), and every collapsing cell reaches −100 pp (vs $\sim$−65 pp on Pythia's literary cells).

Group ablation. Figure 4 reproduces the per-head causal test on NIAH. The architectural partition from Figure 3 is preserved and sharper: Qwen3 and Gemma-4 are decisive at $K \geq 20$ (top-K drops −80 to −100 pp; bot-K stays at 0 pp); Llama is sample-perfect at every

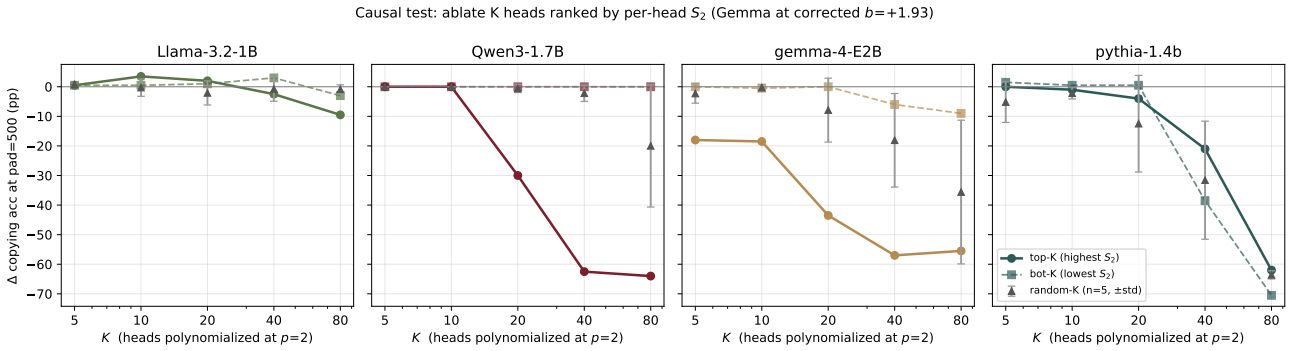

Figure 3. Group ablation: top-K vs bot-K vs random-K, heads ranked by per-head S_p from a single diagnostic forward pass. Solid = top-K (highest S_p); dashed = bot-K; triangles = mean of 5 random-K draws ($\pm 1\sigma$). Gemma-4 ranking uses the corrected calibrated bias $b = +1.93$ (Appendix I).

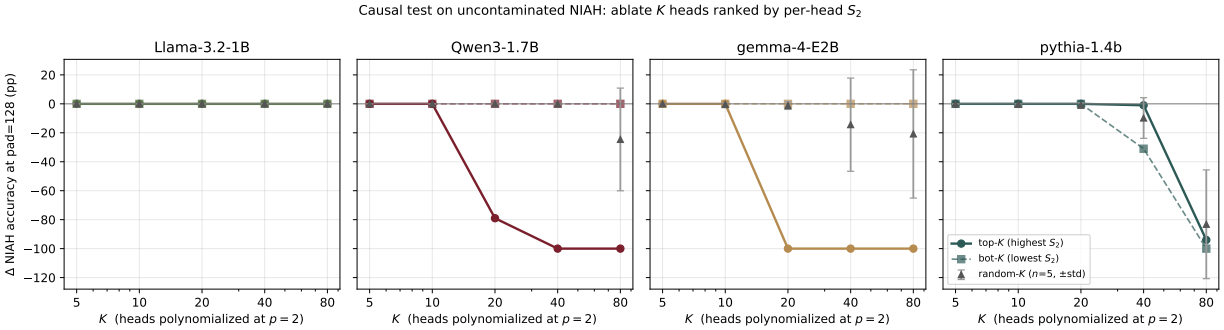

Figure 4. Group ablation on NIAH at pad = 128 (haystack $\sim$1500 tokens). Solid = top-K (highest S_p); dashed = bot-K; triangles = mean of 5 random-K draws ($\pm 1\sigma$); Gemma-4 at corrected $b = +1.93$.

$K \leq 80$ on every subset; Pythia is non-discriminating, with top and bot indistinguishable and both collapsing at $K = 80$. A 20-draw random-K rerun on the two decisive models confirms decisiveness at $K \in \{20, 40\}$; at $K = 80$ Gemma is marginal because random sets of 80 heads can include the critical core (Appendix V).

Matched argmax-concentration diagnostic. Table 4 reports eight contrast cells of the matched NIAH diagnostic at pad = 128; the full 13-cell table is in Table 5 (Appendix F). The exact lower bound has only boundary-equality false negatives (every observed FN has $S_p = 1$ in float64; with $\epsilon = 10^{-6}$ the FN count is zero in every cell, see Appendix P). Empirically the diagnostic is a sharp negative screen but not a sufficient predictor of preservation: all five cells (across the full table) with all-head measured-safe fraction $\leq 3\,\%$ collapse -100 pp, but among cells with $> 3\,\%$ safe, behavior is mixed (compare Qwen3 relu_p8_bauto at 66.3 % all-head safe yet -100 pp behavior, vs. Gemma at 18.3 % all-head safe yet preserving). The answer-key-target diagnostic explains the non-monotonicity: at the final NIAH query position, only $3-7\,\%$ of (layer, head, prompt) triples place the actual needle as the row argmax (Appendix S), so the all-head row-safe fraction averages over many heads not carrying the retrieval circuit. Conditioning on the fixed causal top-$K = 40$ heads (Section 5.4; recipe-matched alternative in Appendix Q gives the same partition) tightens within-model contrasts: Llama relu_p8_bauto falls from 86.4 % (all-head) to 51.6 % measured-safe but remains far above relu_p2_bauto's 4.9 %. Cross-model anomalies remain. S_p is therefore best read as a negative screen and localization tool, not a universal scalar predictor.

6. Discussion

S_p is a conservative pre-deployment screen for zero-shot pointwise substitution, not a sufficiency guarantee. It is strongest as a negative screen: cells with very low measured-safe fraction fail catastrophically in our audit (5/5 cells with $\leq 3\,\%$ all-head measured-safe drop -100 pp). Behavioral preservation depends on which rows and heads are safe, not on the global average; the answer-key target and fixed causal top-K analyses (Appendices Q and S) attribute this to the small fraction ($3-7\,\%$) of layer/head/prompt triples that carry

Table 1. Polynomial-unsafe row rate $\Pr(S_p \geq 1)$ at the longest context († $L = 4096$ for all four models on induction/copying/mqar) and bias $b = 0$, restricted to active-unsaturated rows. Three of four models cross 70 % on every suite at $p = 2$; Llama is the outlier. Median active-distractor count A_{act} shown for $p = 2$. † Pythia-1.4b is OOD at $L = 4096$ (training length 2048); the $L = 1024$ in-distribution Pythia row is induction 66.0 %, copying 61.7 %, mqar 63.4 % at $p=2$ (Table 10 for full per-context breakdown). Pythia's copying-suite cell at $L = 4096$ also has substantially fewer active distractors ($A_{\text{act}} = 148$) and a 36.5 % dead-row rate; restricting to active-unsaturated rows mechanically changes the denominator relative to induction/MQAR. Pythia remains highly unsafe on induction/MQAR and collapses behaviorally under every tested substitution recipe.

Model	Suite	$p=2$	$p=4$	$p=8$	A_{act}
Llama-3.2-1B	ind	19.6	10.3	4.7	2
	cpy	30.2	18.0	9.6	3
	mqar	21.4	11.0	4.7	2
Qwen3-1.7B	ind	94.0	79.3	48.5	1884
	cpy	96.7	75.5	36.5	540
	mqar	93.9	80.1	48.5	1821
Gemma-4-E2B	ind	71.3	56.9	37.6	36
	cpy	96.1	80.1	49.4	96
	mqar	90.6	71.6	41.5	53
Pythia-1.4b†	ind	99.2	98.3	95.1	1407
	cpy	76.9	64.6	51.0	148
	mqar	95.3	92.1	84.6	1351

Table 2. Δ next-token accuracy (pp) at pad = 500 vs softmax baseline (74 − 79 %) on the literary-completion suite, $n = 200$ prompts per cell. Bold cells are the best variant per model. No single (p, b) across the 14 tested variants recovers softmax behavior on all four models. With $n = 200$, two-tailed Wilson 95 % CI half-width is at most ~ 7 pp; differences below this should be read as noise. ‡ Wortsman recipes were originally proposed for Vision Transformers (Wortsman et al., 2023), not for causal LMs; we include them as literature-grounded un-normalized drop-in recipes, not as strong LM-specific baselines. They use $b = 0$ because the un-normalized form has no row-normalizer to absorb a bias shift.

Variant	Llama	Qwen3	Gemma-4	Pythia
Pointwise normalized:				
relu_p2 $b=0$	−52.5	−79.0	−70.5	−70.5
relu_p2 $b=$auto	−44.5	−71.5	−74.0	−65.0
relu_p4 $b=0$	−54.5	−73.0	−27.0	−69.0
relu_p4 $b=$auto	**−21.0**	−61.5	−50.0	−65.0
relu_p8 $b=0$	−67.0	−60.5	−17.5	−65.5
relu_p8 $b=$auto	−25.0	−59.0	**−5.0**	−65.0
relu_p16 $b=0$	−69.0	**−14.5**	−45.0	−77.5
relu_p16 $b=$auto	−52.0	−43.5	−28.5	−77.5
sigmoid $b=0$	−57.5	−79.0	−72.5	**−65.0**
sigmoid $b=$auto	−68.0	−79.0	−77.0	−65.0
Wortsman et al. (un-normalized‡):				
$\text{ReLU}(z)/L$	−64.5	−79.0	−74.0	−77.5
$\text{ReLU}(z)^2/L$	−64.0	−79.0	−72.0	−77.5
$\text{ReLU}(z)/\sqrt{L}$	−74.5	−79.0	−73.5	−77.5
$\text{ReLU}(z)^2/\sqrt{L}$	−74.5	−79.0	−71.0	−77.5

the retrieval circuit. The calibrated bias b_{auto} ranges over two orders of magnitude across model families, so substitution recipes that bake in a fixed b are unlikely to transfer (Appendix J).

The sample-perfect 200/200 NIAH cells (Llama relu_p4_bauto/relu_p8_bauto, Gemma relu_p8_bauto) reflect each model's calibrated-bias-per-model best recipe on a single-token English-noun retrieval template. They establish that polynomial substitution can preserve attention routing under favorable conditions on the Llama and Gemma families, but should not be over-read as a uniform polynomial-substitution claim. Three robustness checks tighten the scope: (i) the partition holds at 1–8 B Llama scale (Appendix AC) and at $n = 500$ on a fresh prompt seed (Appendix AG, 500/500 on every preserving cell, Wilson lower bound 99.24 %); (ii) the partition is not a calibration artifact — Llama preserves at relu^4 over a wide fixed-bias range ($b \in [1, 6]$, 1B), while Qwen3 and Pythia collapse at every b tested (Appendix AH); (iii) on an exploratory RULER-template (Hsieh et al., 2024) niah_single_3-style probe with a uniformly post-cutoff Dwarkesh Patel haystack at ~ 4 K tokens ($n = 50$, Wilson half-widths up to ≈ 13 pp), Llama's relu_p4_bauto collapses to 4 % (softmax baseline 76 %) and relu_p8_bauto partially recovers to 54 % (Appendix AI). The direction matches Corollary 4.3 ($p^\star = \log A / \log(1/(1-\rho))$, with larger A and smaller ρ on the harder probe), and the matched diagnostic on the same prompts moves in the same direction (Llama unsafe drops from 65 % at $p = 4$ to 29 % at $p = 8$). We treat this as a directional stress check rather than a closed-loop validation: the $p = 8$ cell remains below softmax, $n = 50$ gives wide intervals, and $p = 16$ exposes a high-degree substitution/implementation interaction we have not isolated.

The audit has scope limits. The four-model split ($n = 4$ with Llama extended to 1/3/8 B) is a family pattern, not an architectural law: cross-family validation (Mistral, Phi, OLMo, DeepSeek, etc.) is needed. The per-head causal signal is decisive on Qwen3 and Gemma-4 only; Llama is robust to the selective substitutions we test (Appendix AD confirms at 3B), and Pythia's ranking does not discriminate. Per-head ablation on Qwen3 and Gemma is uniformly $\Delta = 0$ pp at every individual head (Appendix AE), implying retrieval is distributed across ≥ 20 heads rather than concentrated in any single head. NIAH variants (long-context 6 K, two-needle, adversarial-decoy, UUID-needle; Appen-

Table 3. Δ NIAH accuracy (pp) at pad = 128 (∼1500 tokens) vs softmax baseline of 200/200 on every model. n = 200 prompts per cell, two-tailed Wilson 95 % CI half-width ≤ 7 pp. Bold cells are the best variant per model. Two recipes (Llama: relu_p4_bauto/relu_p8_bauto; Gemma-4: relu_p8_bauto) achieve 200/200 on this sample (Wilson 95 % lower bound 98.2 %). Pythia collapses on every variant.

Variant	Llama	Qwen3	Gemma-4	Pythia
Pointwise normalized:				
relu_p2 b=0	−19.0	−100	−100	−100
relu_p2 b=auto	−8.5	−100	−100	−100
relu_p4 b=0	−95.5	−100	−96.0	−100
relu_p4 b=auto	**0.0**	−99.5	−100	−100
relu_p8 b=0	−100	−100	−1.0	−100
relu_p8 b=auto	**0.0**	−100	**0.0**	−100
relu_p16 b=0	−99.5	**−50.5**	−26.0	−100
relu_p16 b=auto	−95.0	−93.5	−1.0	−100
sigmoid b=0	−96.5	−100	−100	−100
sigmoid b=auto	−100	−100	−100	−100
Wortsman et al. (un-normalized):				
$\mathrm{ReLU}(z)/L$	−100	−100	−100	−100
$\mathrm{ReLU}(z)^2/L$	−100	−100	−100	−100
$\mathrm{ReLU}(z)/\sqrt{L}$	−100	−100	−100	**−100**
$\mathrm{ReLU}(z)^2/\sqrt{L}$	−100	−100	−100	−100

Table 4. Matched-diagnostic contrasts at NIAH pad = 128. "all-head" is $\Pr(S_p < 1)$ over all active-unsaturated rows; "top-K" restricts to the $K = 40$ causal head set. Top: low all-head safe correctly predicts collapse. Middle: high all-head safe is not sufficient—both cells have ≥ 66 % safe rows yet collapse ≥ 19 pp. Bottom: the three sample-perfect cells; top-K separates Llama from Gemma's low-all-head case and from the failing Qwen3 anomaly above. Full 13-cell table in Table 5.

Model	Recipe	Δ (pp)	all-head %	top-K %
Low-safe ⇒ collapse (sharp negative screen)				
Qwen3	relu_p2_b0	−100	2.9	0.0
Gemma-4	relu_p2_b0	−100	2.0	0.0
Pythia	relu_p4_bauto	−100	0.1	0.0
High global safe, still fails (not sufficient)				
Qwen3	relu_p8_bauto	−100	66.3	21.9
Llama	relu_p2_b0	−19.0	72.3	30.4
Sample-perfect cells (top-K separates the anomalies above)				
Llama	relu_p4_bauto	**0.0**	55.7	21.2
Llama	relu_p8_bauto	**0.0**	86.4	51.6
Gemma-4	relu_p8_bauto	**0.0**	18.3	6.1

dices X to AA) preserve the family pattern, but preservation is template-dependent: adversarial decoys narrow Llama's preserving band (relu_p8_bauto: 0 pp → −48 pp). The harder RULER-Dwarkesh probe (Appendix AI) reveals the recipe-degree gradient discussed above. Compositional and multi-hop retrieval, and trained (rather than zero-shot) substitution recipes, are out of scope. At ReLU^{16}, the diagnostic predicts the smallest unsafe-row fraction yet Llama behavior under bf16 attention is 0 % uniformly on the RULER-template probe; candidate causes include numerical precision of r^{16} in bfloat16, over-sharpening of the routing distribution, or saturation downstream of the substituted row. We report this bf16 outcome and do not interpret it as evidence against the row-level degree trend. The audit covers 1.4–8 B parameters; broader long-context benchmarks (LongBench (Bai et al., 2024), BABILong (Kuratov et al., 2024)) and full sweeps at ≥ 8 B remain future work. The diagnostic applies to pointwise maps, so learned, gated, sparse-selection, and stateful alternatives fall outside our claim. Pythia's $L = 4096$ unsafe-row rate is also OOD relative to its 2048-token training length; within 2048 tokens the in-distribution rate is 66 % at $L = 1024$.

Using S_p as a no-go signal. For a candidate (p, b) recipe: (1) compute S_p on a representative prompt set in one forward pass via a ∼30-line attention hook (Appendix G), restricting to active-unsaturated rows; (2) if the all-head unsafe-row rate is high, expect failure (in our audit, 5/5 matched-diagnostic cells with ≤ 3 % measured-safe rows collapsed −100 pp); (3) a high all-head safe fraction does not guarantee preservation, so condition on the top-K causal heads before claiming a recipe should preserve, and confirm with a behavioral test on the target task.

7. Conclusion

Polynomial attention can preserve top-1 rank while failing to concentrate enough mass for retrieval. $S_p < 1$ is the right row-level quantity, computable from a single forward pass; it fails for a large fraction of active-unsaturated rows in three of four small open-weight LMs at $p = 2, b = 0$, locates the retrieval circuit on the two decisive models, and is a sharp no-go signal for substitution failure. The geometry behind S_p — concentration scaling as $(1 - \Delta/r_*)^p$ rather than $e^{-\beta\Delta}$ — frames why the audit lands as it does: at the active-distractor counts found in pretrained rows, no fixed degree certifies preservation across model families, and the absence of a universal drop-in substitute reflects score-row geometry rather than recipe choice. Whether substitution that is jointly optimized with parameters — or that selects keys non-pointwise — can recover what zero-shot pointwise replacement loses is the natural next question.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

János Aczél. Lectures on Functional Equations and Their Applications. Academic Press, 1966.

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Ré. Zoology: Measuring and improving recall in efficient language models. arXiv:2312.04927, 2023.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In Proceedings of ACL, 2024.

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. arXiv:2501.00663, 2025.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In Proceedings of ICML, 2023.

Mathieu Blondel, André F. T. Martins, and Vlad Niculae. Learning with Fenchel–Young losses. Journal of Machine Learning Research, 21:1–69, 2020.

Siwan Boufadène and François-Xavier Vialard. Sliced ReLU attention: Quasi-linear contextual expressivity via sorting. arXiv:2512.11411, 2025.

Etienne Boursier and Claire Boyer. Softmax as linear attention in the large-prompt regime: A measure-based perspective. arXiv:2512.11784, 2025.

John S. Bridle. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In Neurocomputing: Algorithms, Architectures and Applications, pages 227–236. Springer, 1990.

David Chiang and Peter Cholak. Overcoming a theoretical limitation of self-attention. In Proceedings of ACL, 2022.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamás Sarlós, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Łukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. Rethinking attention with Performers. In Proceedings of ICLR, 2021.

Gonçalo M. Correia, Vlad Niculae, and André F. T. Martins. Adaptively sparse transformers. In Proceedings of EMNLP-IJCNLP, 2019.

Carles Gelada, Jacob Buckman, Sean Zhang, and Txus Bach. Scaling context requires rethinking attention. arXiv:2507.04239, 2025.

Google DeepMind. Gemma 4: Lightweight open models for on-device and multimodal use. Technical Report, 2026. https://ai.google.dev/gemma.

Michael Hahn. Theoretical limitations of self-attention in neural sequence models. Transactions of the Association for Computational Linguistics, 8:156–171, 2020.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In Proceedings of COLM, 2024.

Praneeth Kacham, Vahab Mirrokni, and Peilin Zhong. PolySketchFormer: Fast transformers via sketching polynomial kernels. In Proceedings of ICML, 2024.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In Proceedings of ICML, 2020.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. BABILong: Testing the limits of LLMs with long context reasoning-in-a-haystack. In Proceedings of NeurIPS Datasets and Benchmarks, 2024.

Pierre-David Letourneau, Manish Kumar Singh, Hsin-Pai Cheng, Shizhong Han, Yunxiao Shi, Dalton Jones, Matthew Harper Langston, Hong Cai, and Fatih Porikli. PADRe: A unifying polynomial attention drop-in replacement for efficient Vision Transformer. In Proceedings of ICLR, 2025.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Selective attention improves transformer. In Proceedings of ICLR, 2025.

R. Duncan Luce. Individual Choice Behavior: A Theoretical Analysis. Wiley, 1959.

André F. T. Martins and Ramón Fernandez Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In Proceedings of ICML, 2016.

Daniel McFadden. Conditional logit analysis of qualitative choice behavior. In Paul Zarembka, editor, Frontiers in Econometrics, pages 105–142. Academic Press, 1974.

Meta AI. The Llama 3 herd of models. arXiv:2407.21783, 2024. https://huggingface.co/meta-llama/Llama-3.2-1B.

Ken M. Nakanishi. Scalable-softmax is superior for attention. arXiv:2501.19399, 2025.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. arXiv:2209.11895, 2022.

Ben Peters, Vlad Niculae, and André F. T. Martins. Sparse sequence-to-sequence models. In Proceedings of ACL, 2019.

Zhen Qin, Weixuan Sun, Hui Deng, Dongxu Li, Yunshen Wei, Baohong Lv, Junjie Yan, Lingpeng Kong, and Yiran Zhong. cosFormer: Rethinking softmax in attention. In Proceedings of ICLR, 2022.

Qwen Team. Qwen3 technical report. arXiv:2505.09388, 2025.

Jason Ramapuram, Federico Danieli, Eeshan Gunesh Dhekane, Floris Weers, Dan Busbridge, Pierre Ablin, Tatiana Likhomanenko, Jagrit Digani, Zijin Gu, Amitis Shidani, and Russell Webb. Theory, analysis, and best practices for sigmoid self-attention. In Proceedings of ICLR, 2025.

Clayton Sanford, Daniel Hsu, and Matus Telgarsky. Representational strengths and limitations of transformers. In Proceedings of NeurIPS, 2023.

Hemanth Saratchandran, Peng Zheng, Yiping Ji, Bowen Zhang, and Simon Lucey. Rethinking softmax: Self-attention with polynomial activations. arXiv:2410.18613v3, 2024.

Shawn Tan, Songlin Yang, Aaron Courville, Rameswar Panda, and Yikang Shen. Stick-breaking attention. In Proceedings of ICLR, 2025.

Nataliya Vasylenko, Hugo Pitorro, André F. T. Martins, and Marcos Treviso. Long-context generalization with sparse attention. In Proceedings of ICLR, 2026.

Petar Veličković, Christos Perivolaropoulos, Federico Barbero, and Razvan Pascanu. softmax is not enough (for sharp size generalisation). In Proceedings of ICML (PMLR v267:61190–61211), 2025.

Mitchell Wortsman, Jaehoon Lee, Justin Gilmer, and Simon Kornblith. Replacing softmax with ReLU in Vision Transformers. arXiv:2309.08586, 2023.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In Proceedings of ICLR, 2024.

Junyi Yan, Yuxin Zheng, and Sai Bi. Sample-complexity comparison of sigmoid and softmax self-attention. arXiv:2510.07831, 2025.

Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. Differential Transformer. In Proceedings of ICLR, 2025.

John I. Yellott. The relationship between Luce's choice axiom, Thurstone's theory of comparative judgment, and the double exponential distribution. Journal of Mathematical Psychology, 15:109–144, 1977.

Michael Zhang, Kush Bhatia, Hermann Kumbong, and Christopher Ré. The Hedgehog & the Porcupine: Expressive linear attentions with softmax mimicry. In Proceedings of ICLR, 2024.

Zayd Muhammad Kawakibi Zuhri, Erland Hilman Fuadi, and Alham Fikri Aji. Softpick: No attention sink, no massive activations with rectified softmax. arXiv:2504.20966, 2025.

A. Extended related work

Softmax characterizations and length limits. Softmax has a long lineage in choice theory and statistics: Luce's choice axiom (Luce, 1959), McFadden's conditional logit (McFadden, 1974), Yellott's translation-invariance characterization (Yellott, 1977), and Bridle's introduction to neural classification (Bridle, 1990); modern axiomatic accounts trace through Fenchel–Young losses (Blondel et al., 2020) and the functional-equation framing of Aczél (1966). Our Theorem 4.1 is an attention-native re-presentation: the only pointwise normalized map invariant to additive row shifts is exponential; everything else has a length-dependent threshold. Hahn (2020); Chiang and Cholak (2022); Sanford et al. (2023) establish formal limits on softmax's representational capacity at scale, and Veličković et al. (2025); Nakanishi (2025); Boursier and Boyer (2025); Vasylenko et al. (2026) study softmax's length-and-dispersion limits empirically and with continuous-function theory; Veličković et al. (2025)'s adaptive-temperature result is the closest analog to our Corollary 4.3, applied to the polynomial-substitution case rather than to long-context softmax.

Pointwise polynomial and ReLU substitutes. Wortsman et al. (2023) introduced the un-normalized $\mathrm{ReLU}(z)/L$ and $\mathrm{ReLU}(z)^2/\sqrt{L}$ recipes for vision transformers; Saratchandran et al. (2024) studied higher-degree polynomial substitutes; Kacham et al. (2024); Letourneau et al. (2025); Qin et al. (2022) built efficient polynomial-attention variants; Boufadène and Vialard (2025) introduced sliced-ReLU attention as a quasi-linear contextual map. These papers ground their case on training stability, kernel approximation, or wall-clock cost, and report that the resulting model trains successfully; they do not give a row-level guarantee on concentration.

Non-pointwise alternatives. Ramapuram et al. (2025); Yan et al. (2025) use sigmoid attention; Tan et al. (2025) uses stick-breaking attention; Zuhri et al. (2025) introduces Softpick; Ye et al. (2025) introduces Differential Transformer; Leviathan et al. (2025) uses Selective Attention; the sparsemax/entmax family is Martins and Astudillo (2016); Peters et al. (2019); Correia et al. (2019). These mechanisms can change row-level routing in ways that fall outside the pointwise impossibility result of Corollary 4.3; the row-ratio diagnostic itself extends to sigmoid (Appendix T), although the polynomial-degree bound does not.

Stateful and learned-feature mechanisms. State-based polynomial mechanisms (Power Attention (Gelada et al., 2025), Titans (Behrouz et al., 2025)) and learned feature maps such as Hedgehog (Zhang et al., 2024) train ϕ such that $\phi(Q)\cdot\phi(K)\approx\exp(QK)$. These are kernel approximations rather than pointwise-on-z replacements; our diagnostic, applied to the original transformer's score row, does not directly predict their behavior. The corresponding scoped prediction is that the diagnostic, applied to the post-feature-map score row $\phi(Q)\cdot\phi(K)$, would identify whether the working regime corresponds to $S_p<1$ on the transformed rows. Verifying this requires training infrastructure beyond a workshop submission.

Linear and kernelized attention. Katharopoulos et al. (2020); Choromanski et al. (2021) are the canonical linear/kernelized attention references; they make different design choices (kernel feature maps, causal-mask compatibility) and are not zero-shot pointwise drop-ins.

B. Empirical setup details

Models and configurations. We use four open-weight decoder LMs:

- meta-llama/Llama-3.2-1B (Meta AI, 2024): 16 layers, 32 query heads with grouped-query attention to 8 KV heads, 64-dim head, RoPE, 128 K nominal context.
- Qwen/Qwen3-1.7B (Qwen Team, 2025): 28 layers, 16 query heads to 8 KV heads, 128-dim head, per-head Q/K-norm, 32 K nominal context.
- google/gemma-4-E2B (Google DeepMind, 2026): 35-layer text decoder of a multimodal model, tanh pre-softmax softcap on attention logits, 128 K nominal context. We load via AutoModelForImageTextToText and use only the text decoder for prompt-completion eval.
- EleutherAI/pythia-1.4b (Biderman et al., 2023): 24 layers, 16 heads, GPT-NeoX architecture, no Q/K-norm, 2048-token training context (we caveat $L=4096$ as OOD for Pythia specifically).

All loads use torch_dtype=bfloat16 and attn_implementation="eager". The diagnostic hook intercepts the score row immediately before softmax, captures S_p in float64, and either calls the original softmax or substitutes a user-supplied function; bit-identity vs. vanilla forward is verified in Appendix I.

Calibration. The bias b_{auto} uses a 5 %-mass-budget rule: on a held-out 1024-token generic-text input, b is the smallest threshold for which the average softmax row places at most 5 % of its mass on positions with score below $-b$. Per-model values are listed in Appendix J. Calibration is done once per model on a held-out generic-text input that is not part of any prompt suite.

Numerical stability. S_p is computed in float64 with the relative form $S_p = \sum_{j \neq *} (r_j / r_*)^p$; we never compute r^p at large magnitude before normalization. Active-unsaturated filtering uses $r_* > 0$; for Gemma-4 we additionally enforce $u_* < \tau$ before the tanh softcap (the cap is applied upstream in the eager attention path, so post-cap rows never reach the unsafe regime via saturation).

Statistical power. The substitution sweep and group ablations use $n = 200$ trials per cell. With softmax baseline $\bar{p} = 0.78$ on the literary suite, the two-tailed Wilson 95 % CI half-width is ≈ 6 pp; the minimum detectable effect at 80 % power against the same baseline is ≈ 10 pp. All reported headline cells have $|\Delta| \gg 30$ pp, comfortably above this threshold. NIAH cells with 200/200 exact match have a two-sided Wilson 95 % interval of $[98.2\,\%, 100\,\%]$; cells with 0/200 have a Wilson 95 % upper bound of $\approx 1.9\,\%$. The Wilson interval is preferred over the Wald interval for proportion estimation at the boundary.

Prompt suites. The induction suite (Olsson et al., 2022) and MQAR suite (Arora et al., 2023) use random-vocabulary tokens; the literary-completion "copying" suite uses 200 openings of well-known novels and asks the model to continue, padded with random tokens to a target length. The literary suite is contaminated by memorization (every model gets the right next token from training data), and we caveat this throughout. NIAH (Section 5.5, Appendix E) is the uncontaminated probe.

Gemma-4 softcap. Gemma-4-E2B inherits Gemma-3's pre-softmax tanh softcap on attention logits: $z \to \tau \tanh(z/\tau)$ with $\tau = 50$. The diagnostic uses the post-cap score row, so saturated rows are filtered before S_p is computed; the substitution recipe replaces softmax with $r^p / \sum r^p$ on the post-cap row, which is the same input the original softmax sees.

C. Length generalization

We track the unsafe-row rate as a function of context length $L \in [256, 4096]$ on the induction suite at $b = 0$. Llama and Qwen3 are nearly flat across this range (Llama: 18−22 %; Qwen3: 90−94 %), reflecting families that do not change qualitative behavior across in-distribution context lengths. Gemma-4 is also flat (69−72 %). Pythia is the only model that shifts substantially with L: 66 % at $L = 1024$ (in-distribution), rising to 99 % at $L = 4096$ (OOD relative to its 2048-token training length). This is the basis for our reading that Pythia's headline $L = 4096$ row in Table 1 should be interpreted as an OOD stress reading rather than as an in-distribution prevalence claim.

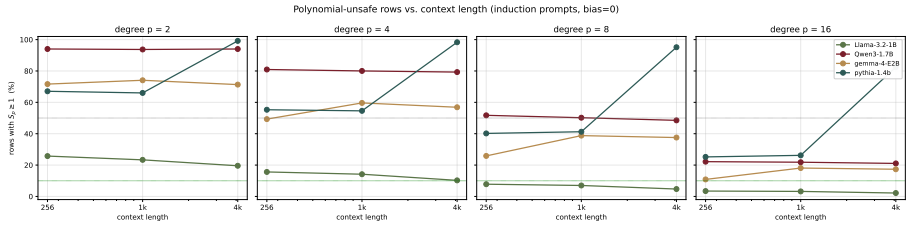

Figure 5. Length generalization of the unsafe-row rate. Fraction of active-unsaturated rows with $S_2 \geq 1$ on the induction suite at $b = 0$, by context length L. Llama, Qwen3, and Gemma-4 are nearly flat across $L \in [256, 4096]$. Pythia rises from 66 % at $L = 1024$ (in-distribution) to 99 % at $L = 4096$ (OOD; Pythia's training length is 2048). The headline $L = 4096$ Pythia row in Table 1 should be read as an OOD stress reading rather than as an in-distribution prevalence claim.

D. Per-model ablation details

Figures 3 and 4 summarize the group ablation across all four models on the literary and NIAH suites; we expand on the per-model picture here.

Qwen3-1.7B. Top-K is strictly worse than every random-K draw at $K \geq 20$ on both suites: at $K = 40$, top-K drops -92.0 pp (literary) and -100 pp (NIAH) while bot-K drops -2.0 pp (literary) and 0 pp (NIAH); random-K is concentrated near zero with the worst draw at -14.0 pp (NIAH, $K = 40$). The 20-draw rerun (Appendix V) confirms decisiveness in 20/20 at every $K \in \{20, 40, 80\}$ on NIAH. The per-head ranking from a single diagnostic forward pass identifies a behaviorally critical subset.

Gemma-4-E2B. Decisiveness at the corrected $b = +1.93$ is 20/20 at $K \in \{20, 40\}$ on NIAH (top -48.5 to -100 pp; bot 0 pp; random worst -20.0 to -98.0 pp). At $K = 80$ the cell becomes marginal because random subsets of 80 heads can include the small core of critical heads (the random-draw worst case ties top at -100 pp in 3/20 draws). On the literary suite at the original $b = -2$ ranking, decisiveness at $K \leq 40$ is 5/5; at corrected $b = +1.93$ this becomes marginal at $K = 80$ with 1/5 random draws happening to include enough critical heads to beat top.

Llama-3.2-1B. The model is uniformly robust to selective polynomial substitution at every $K \leq 80$ on both suites: top, bot, and every random draw all stay within ± 7 pp of the softmax baseline. This is consistent with Llama's globally low S_p distribution (median $S_2 = 0.107$ at $b = 0$, well under 1): the diagnostic predicts that polynomial substitution on Llama is uniformly safe at this scale, and the behavioral test agrees. The result does not falsify the diagnostic; it shows that the diagnostic's positive predictions are also consistent with Llama's behavior, but the per-head causal signal is not separable on Llama because there is no high-S_p tail to ablate.

Pythia-1.4b. The ranking is uniformly non-discriminating at $K \leq 40$ on both suites (top, bot, random all near baseline) and uniformly catastrophic at $K = 80$ where every subset, top or bot, collapses (-94 to -100 pp). The cause is that Pythia's S_p distribution is uniformly elevated (median $S_2 = 287$ at $b = 0$, $L = 4096$ on induction), so the per-head ranking has limited dynamic range and even the bottom heads sit deep in the unsafe regime. The diagnostic identifies Pythia as a model where polynomial substitution is unrecoverable; the head-level causal handle requires a S_p distribution with a separable tail, which Pythia lacks.

Random-K caveat. "Decisive" is a per-cell descriptive flag, not a formal hypothesis test. The 5-draw and 20-draw decisiveness counts in Tables 6 and 17 are reported transparently; we do not adjust for multiple comparisons across K values, and the marginal Gemma $K = 80$ NIAH cell is reported as such.

E. NIAH suite construction

The NIAH (needle-in-haystack) suite is a single-template retrieval probe: "The secret word is X. [haystack] The secret word is". The needle X is a single English noun token sampled uniformly from a 200-word list of common, single-token nouns; the haystack is random text padded to a target length in $\{0, 8, 32, 128, 512\}$ tokens (corresponding to total prompt lengths of $\sim 50, 200, 500, 1500, 6000$ tokens). The needle is randomly chosen per prompt (numpy seed 0), so memorization of training-data continuations cannot help.

Audit. We verified that softmax baselines reach 200/200 exact match on every model at every pad $\in \{0, 8, 32, 128\}$, so the suite has zero baseline noise to confound substitution deltas. The pad $= 512$ suite (Appendix Y) preserves 200/200 softmax baselines on Llama, Qwen3, and Gemma-4. Pythia's 2048-token training length excludes it from the pad $= 512$ run.

Prompt audit. Of the 200 NIAH prompts, every one has a unique needle, every haystack contains no occurrence of the needle outside the planted position, and the final query word "is" is followed by the model's predicted continuation. Exact-match accuracy uses tokenizer-level equality on the first generated token vs. the planted needle.

Multi-needle and long-context variants. The two-needle variant (Appendix X) asks for one of two distinct planted needles based on a final-position cue; the long-context variant (Appendix Y) extends pad to 512. Both

use the same per-prompt random-needle construction.

F. Matched NIAH diagnostic: extended discussion

Table 5 gives the full matched diagnostic referenced from the main text. This section walks through the four sub-claims that the table supports; bootstrap CIs are in Appendix R and the recipe-matched top-K alternative is in Appendix Q.

Table 5. Matched NIAH argmax-concentration diagnostic at pad = 128. "all" aggregates over all active-unsaturated rows; "top-K" restricts to the fixed $K = 40$ causal head set used in Section 5.4. "meas" is $\Pr(S_p < 1)$; "pred" is $\Pr(p \geq p^{\star}_{\text{exact}})$ from Corollary 4.3. "FN" is the strict-boundary count (predicted safe but $S_p \geq 1$); every observed FN has $S_p = 1$ in float64 and disappears under tolerance 10^{-6}. Bootstrap CIs and recipe-matched top-K alternatives are in Appendices Q and R.

			all-head %		top-K %			
Model	Recipe	NIAH Δ	meas	pred	meas	pred	rows	FN
Llama-3.2-1B	relu_p2_b0	−19.0	72.3	53.9	30.4	18.1	102 k	63
Llama-3.2-1B	relu_p2_bauto	−8.5	20.3	6.0	4.9	0.0	102 k	0
Llama-3.2-1B	relu_p4_bauto	**0.0**	55.7	16.1	21.2	0.9	102 k	0
Llama-3.2-1B	relu_p8_bauto	**0.0**	86.4	36.3	51.6	9.4	102 k	0
Qwen3-1.7B	relu_p2_b0	−100	2.9	0.7	0.0	0.0	90 k	31
Qwen3-1.7B	relu_p4_bauto	−99.5	33.2	15.2	0.0	0.0	87 k	0
Qwen3-1.7B	relu_p8_bauto	−100	66.3	22.2	21.9	0.0	87 k	0
Gemma-4-E2B	relu_p2_b0	−100	2.0	1.2	0.0	0.0	56 k	0
Gemma-4-E2B	relu_p4_bauto	−100	2.7	0.3	0.0	0.0	56 k	0
Gemma-4-E2B	relu_p8_bauto	**0.0**	18.3	1.2	6.1	0.0	56 k	0
Pythia-1.4b	relu_p2_b0	−100	18.4	12.1	2.9	2.9	40 k	2
Pythia-1.4b	relu_p4_bauto	−100	0.1	0.0	0.0	0.0	50 k	0
Pythia-1.4b	relu_p8_bauto	−100	0.2	0.1	0.0	0.0	50 k	0

(i) Negative direction is sharp. Among the 13 cells, all five with all-head measured-safe fraction $\leq 3\,\%$ collapse -100 pp in NIAH (Qwen3 relu_p2_b0, Gemma relu_p2_b0, Gemma relu_p4_bauto, Pythia relu_p4_bauto, Pythia relu_p8_bauto). The strict implication "measured-safe $\leq 3\,\%$ on row screen $\Rightarrow$ -100 pp on behavioral test" holds in 5/5 observations.

(ii) Positive direction has known false alarms. Among the 8 cells with measured-safe $> 3\,\%$, three preserve fully (Llama relu_p4/p8_bauto, Gemma relu_p8_bauto), two partially fail (Llama relu_p2_b0 at -19 pp; Llama relu_p2_bauto at -8.5 pp), and three still collapse (Qwen3 relu_p4/p8_bauto and Pythia relu_p2_b0). The diagnostic does not predict which $> 3\,\%$ cells will preserve; this is the negative result behind the "conservative screen, not sufficiency guarantee" framing.

(iii) Cross-model anomaly at relu_p8_bauto. Qwen3 has 66.3 % all-head measured-safe rows and fails -100 pp; Gemma has 18.3 % all-head measured-safe rows and preserves at 0 pp. The contrast inverts under the fixed top-K restriction (Qwen3 21.9 %, Gemma 6.1 %): the high-S_p heads on Gemma are concentrated in a critical subset that the substitution disproportionately spares relative to Qwen3, where the high-S_p heads are spread across many heads including critical ones. The answer-key analysis (Appendix S) supports this: only $3-7\,\%$ of (layer, head, prompt) triples place the actual needle as the row argmax, and the architectural distribution of those triples differs by family.

(iv) Bias sensitivity. Within Llama, the all-head measured-safe fraction does not order behavior monotonically. relu_p2_b0 has 72.3 % measured-safe rows but drops -19 pp; relu_p2_bauto has only 20.3 % measured-safe rows but drops -8.5 pp. So at $p = 2$ the cell with higher all-head safe fraction performs worse behaviorally—a direct counterexample to using global row-safe fraction as a scalar success predictor. The two higher-degree calibrated cells (relu_p4_bauto at 55.7 % and relu_p8_bauto at 86.4 %) are both sample-perfect at 0 pp, so degree and bias both shape behavior, but the relationship to the all-head S_p summary is non-monotone. The within-Llama ordering is recoverable from the top-$K = 40$ measured-safe column ($p = 2, b = 0$: 30.4 %; $p = 2, b_{\text{auto}}$:

4.9 %; $p=4, b_{\text{auto}}$: 21.2 %; $p=8, b_{\text{auto}}$: 51.6 %), reinforcing that critical-head conditioning is needed to reduce S_p to a behavioral predictor.

NIAH-strengthening. The architectural pattern of Table 2 is sharpened on NIAH: every preserving cell improves to 0 pp drop (vs −5 pp on Gemma's literary relu_p8_bauto), and every collapsing cell reaches −100 pp (vs ∼−65 pp on Pythia's literary cells). NIAH removes the literary suite's contamination confound and produces a sharper substrate for the substitution comparison.

G. Reproducibility

Code and data release. Code, prompt suites, calibration scripts, and raw per-row parquets will be released alongside the full conference paper version; the workshop submission omits the release pending the de-anonymization step at the conference camera-ready stage.

Model commits. The four open models are publicly available with these exact identifiers and immutable Hugging Face commit hashes (downloaded 2026-05-07): meta-llama/Llama-3.2-1B at commit 4e20de36, Qwen/Qwen3-1.7B at 70d244cc, google/gemma-4-E2B at 9d535988 (loaded via AutoModelForImageTextToText and using only the text decoder), and EleutherAI/pythia-1.4b at fedc38a1. The Qwen3-8B prevalence spot check uses Qwen/Qwen3-8B.

Loads and seeds. All loads use torch_dtype=bfloat16 and attn_implementation="eager". Random seeds: prompt construction is deterministic (suite-level seed 0); the 200 NIAH prompts use numpy seed 0; the 5-draw group-ablation random-K uses seeds $\{0, 1, 2, 3, 4\}$; the 20-draw NIAH rerun (Appendix V) uses seeds $\{0, \ldots, 19\}$.

Hook pseudocode. The diagnostic hook is a ∼30-line modification of the standard transformers eager and SDPA attention forward functions; the substitution recipe is given as Algorithm 1 (in the main text, Section 5.3). The calibration step computes b_{auto} as the smallest threshold for which a held-out generic-text input has at most 5 % of softmax mass below score $-b$ (Appendix B).

Audit log. Appendix I traces each numerical claim to the specific script that produced it, including the bias-pick collision (initial and residual) and the bit-identity verification on toy and production architectures.

Hardware. All 1.4–5B-parameter experiments run on a single 4×A100 (40 GB) machine; the Qwen3-8B spot check uses the same hardware with bf16 sharded across the four GPUs. Wall-clock for the full pipeline is ∼36 hours.

H. Head-type classification

For each (layer, head) pair we run a single forward pass on 80 induction prompts at $L = 1024$ and record the post-softmax attention weights. We compute four features per head: (i) "induction" = mean weight at the planted-B position; (ii) "entropy" = average row entropy; (iii) "locality" = mean distance from query to argmax key; (iv) "sink" = mean weight at position 0 (BOS), the canonical attention-sink phenomenon (Xiao et al., 2024). We classify each head as sink if sink > 0.3; induction if induction > 0.5; local if locality < 8 and induction < 0.1; otherwise diffuse. With these thresholds no head crossed the 0.5 induction threshold, so the empirical distribution falls into {sink, local, diffuse}; the threshold is admittedly strict, and a soft scan of the induction-feature distribution does identify a long tail at 0.1–0.3 that may carry the actual induction circuit. We retain the strict classifier to avoid over-claiming the presence of canonical induction heads in models below 2 B parameters.

I. Audit notes

We performed a six-point audit of the empirical pipeline and report the findings here.

Hook bit-identity. The diagnostic hook intercepts the score row immediately before softmax, optionally captures it, and either calls the original softmax or substitutes a user-supplied function. The bit-identity verification suite tests three independent settings: (i) a randomly-initialized Llama configuration (max-diff 0.00); (ii) a real Pythia-1.4b checkpoint in bfloat16 with the GPT-NeoX attention path (max-diff 0.00); and (iii) a randomly-initialized Gemma-3 configuration (Gemma3Config with 64-dim hidden, 2 layers, tanh softcap enabled), constructed specifically because Gemma-4-E2B inherits the same softcap mechanism from Gemma-3 but adds multimodal text/image branches that complicate a pure softcap-only test (max-diff 0.00). The toy Gemma-3 architecture isolates the softcap code path that runs identically inside the production Gemma-4-E2B forward pass; all production substitution sweeps and group ablations in Sections 5.3 to 5.5 use the actual Gemma-4-E2B weights from google/gemma-4-E2B on HuggingFace, released April 2026. The hook fires on every layer of every production model: 16/16 Llama-3.2-1B, 28/28 Qwen3-1.7B, 24/24 Pythia-1.4b, 35/35 Gemma-4-E2B (text decoder).

Substitution machinery. We verified that selective_replace($\emptyset$) produces output bit-identical to the baseline (max-diff 0.00); that selective_replace($\{(\ell, h)\}$) with one head produces a non-zero output difference; and that distinct single-head ablations produce distinct outputs.

Bias-pick collision (initial); pre/post-fix delta disclosed. Three downstream scripts initially used sorted(biases)[-1] to pick the auto-calibrated bias. For Llama, Pythia, and Gemma-4 the auto-bias is positive and this returns the correct row. For Qwen3 the auto-bias is -3.36, which is smaller than 0 in the natural ordering, so the script silently used $b = 0$ instead. We replaced the picker with max(biases, key = abs) in the headline, phase-plot, and per-head-ablation scripts and re-ran every affected cell from raw parquets. Pre/post-fix delta: the $b = 0$ numbers in Table 1 and Figure 2 are bit-identical (this column never depended on the picker). The Qwen3 $b = b_{\text{auto}}$ row in Table 11 changed from a silently-mislabeled-$b = 0$ value to the correct $b = -3.36$ value; the figure phase plot was re-rendered after the fix; all reported headline numbers in Sections 5.2 to 5.5 use the post-fix code path.

Bias-pick collision (residual; re-run). The replacement picker max(biases, key = abs) has a residual edge case for Gemma-4-E2B: its bias set is $\{+1.93, 0, -1, -2\}$ and $|-2| > |+1.93|$, so the picker initially returned -2. We re-ran Gemma-4's group ablation at the corrected $b = +1.93$ (forcing the bias) and updated Figure 3: the diagnostic ranking is decisive at $K \leq 40$ (top -43.5 to -57.0 pp, bot 0 to -6 pp, random mean -7.9 to -18.1 pp), and marginal at $K = 80$ where one random draw of 80 heads happened to include enough critical heads to beat top-K. The other three models' rankings used the correct b_{auto} as logged.

Pandas attribute collision. hrow.head on a pandas Series resolves to the Series.head method rather than the head column; we corrected to bracket access in two ablation drivers.

Copying suite contamination. See Section 6. We considered re-running the group ablation on the random-vocab induction suite. Induction baselines at $L = 4096$ are $\{0, 3, 5, 21\}$ % across the four models, so the headroom available for an ablation Δ is below the test's noise floor. We separately checked shorter contexts ($L = 256$ and $L = 1024$) on Llama and Qwen3: the baselines at shorter L are not appreciably higher (Llama 13/10/22 % at $L = 256/1024/4096$; Qwen3 16/2/1 %). The induction suite as currently designed therefore cannot resolve substitution-induced deltas at any context length we tested. We disclose the confound rather than over-claim.

Random-K decisiveness. Throughout, "decisive" means top-K Δ-accuracy is strictly worse than every random-K draw; with only five random draws this is a noisy estimator and we therefore use "decisive" as a per-cell descriptive flag, not a formal hypothesis test. Table 6 reports the 5-draw result: decisive for Gemma-4 at $K \leq 40$ on the corrected literary-suite bias and at every $K \leq 40$ on NIAH (the $K = 80$ cell is marginal because one of five random draws happened to include critical heads), decisive for Qwen3 at $K \geq 20$, marginal for Llama only at $K = 80$ on literary, and not decisive for Pythia at any K. We also re-ran the NIAH ablation on the two decisive models (Qwen3 and Gemma-4) with 20 random draws (Appendix V); the result is consistent: Qwen3 decisive in 20/20 at every $K \in \{20, 40, 80\}$, Gemma decisive in 20/20 at $K \in \{20, 40\}$ and marginal at $K = 80$ (17/20, with 3 of the 20 random draws happening to include the critical head set and tying top at -100 pp).

Table 6. Decisiveness of the diagnostic ranking at each K. "Decisive" if the top-K Δ-accuracy is strictly worse than every random-K sample. * Gemma-4-E2B values reflect the original literary-suite ranking at $b=-2$ (the residual bias-pick edge case in Appendix I); at the corrected $b=+1.93$ the $K=80$ cell becomes marginal because one of five random draws of 80 heads happened to include enough critical heads to beat top-K. The $K\le 40$ cells remain decisive at the corrected bias.

Model	$K=5$	$K=10$	$K=20$	$K=40$	$K=80$
Llama-3.2-1B	overlap	no signal	no signal	overlap	decisive
Qwen3-1.7B	no signal	no signal	decisive	decisive	decisive
Gemma-4-E2B*	decisive	decisive	decisive	decisive	marginal*
Pythia-1.4b	no signal	no signal	no signal	no signal	no signal

J. Calibrated bias by family

We use a 5%-mass-budget calibration: the threshold b is chosen so that on a held-out generic text input, the average attention row places at most 5 % of softmax mass on positions whose score lies below $-b$. The numerical value of S_p depends on b (since $r_j = \mathrm{clamp}(z_j+b, 0, \tau)$), so the per-row classification at $b=b_{\text{auto}}$ and $b=0$ are two distinct operating points, not the same criterion at the same threshold. We report both throughout: (i) the $b=0$ column (Tables 1 and 10) is calibration-free—it reads the unmodified row directly—and the headline "three of four models above 70 % unsafe" result is in this column; (ii) the $b=b_{\text{auto}}$ column (Table 11) is calibrated per model and is offered as a complementary reading at the per-model 5%-mass-budget operating point. The auto-bias values themselves are below; the wide spread is an architectural fingerprint rather than a hyperparameter we tuned for headline numbers. We do not depend on the calibration choice for the prevalence claim.

Per-model auto-calibrated values:

Model	b_{auto}
Llama-3.2-1B	+3.5547
Qwen3-1.7B	−3.3594
Gemma-4-E2B	+1.9258
Pythia-1.4b	+119.7500

The wide spread is itself an architectural fingerprint and is the reason a single fixed-b recipe is unlikely to transfer across families.

K. Theory: proofs

Proof of Theorem 4.1

Row-shift invariance demands $A_\varphi(z + c\mathbf{1}) = A_\varphi(z)$ for every z, c. Equivalently, $\varphi(z_i + c)/\sum_j \varphi(z_j + c) = \varphi(z_i)/\sum_j \varphi(z_j)$ for all i. Setting all but one entry equal and varying c, we deduce $\varphi(z + c)/\varphi(z) = h(c)$ for a function $h : \mathbb{R} \to (0, \infty)$ independent of z. Then $\varphi(z + c + d) = h(c)\,\varphi(z + d) = h(c)\,h(d)\,\varphi(z)$, while also $\varphi(z + c + d) = h(c + d)\,\varphi(z)$; therefore $h(c + d) = h(c)\,h(d)$, i.e., h is multiplicatively Cauchy. Continuity and positivity force $h(c) = e^{\beta c}$ for some $\beta \in \mathbb{R}$, hence $\varphi(z) = Ce^{\beta z}$.

Proof of Theorem 4.2

Since the target is score-maximal, every active distractor satisfies $z_j \le z_* - \Delta$, hence $r_j = z_j + b \le r_* - \Delta$. By active-branch monotonicity of g,

$$g(r_j) \;\le\; g(r_* - \Delta) \;=\; g(r_*) \cdot R_g(r_*, \Delta).$$

Summing over the A active distractors gives $\sum_{j \ne *} g(r_j) \le A\, g(r_*)\, R_g(r_*, \Delta)$. Therefore

$$A_g(z)_* \;=\; \frac{g(r_*)}{g(r_*) + \sum_{j \ne *} g(r_j)} \;\ge\; \frac{1}{1 + A\, R_g(r_*, \Delta)}.$$

For $g(r) = Cr^p$ on the active branch, $R_g(r_*, \Delta) = (r_* - \Delta)^p / r_*^p = (1 - \Delta/r_*)^p$, yielding the relative-margin form. □

Proof of Theorem 4.4

Fix a distractor $j \neq *$ and let $x_j = r_j/r_* \in [0,1]$. Since $\varphi(r_*) = \sum_{q=0}^{p} c_q r_*^q > 0$, normalize by setting $\alpha_q = c_q r_*^q/\varphi(r_*) \geq 0$, so $\sum_{q=0}^{p} \alpha_q = 1$. Then

$$\frac{\varphi(r_j)}{\varphi(r_*)} = \frac{\sum_{q=0}^{p} c_q r_j^q}{\sum_{q=0}^{p} c_q r_*^q} = \sum_{q=0}^{p} \alpha_q \, x_j^q.$$

Because $x_j \in [0,1]$ and $q \leq p$, we have $x_j^q \geq x_j^p$ for every q, so

$$\sum_{q=0}^{p} \alpha_q \, x_j^q \;\geq\; \sum_{q=0}^{p} \alpha_q \, x_j^p \;=\; x_j^p.$$

Hence $\varphi(r_j)/\varphi(r_*) \geq (r_j/r_*)^p$. Summing over distractors,

$$A_\varphi(r)_* \;=\; \frac{1}{1+\sum_{j\neq *} \varphi(r_j)/\varphi(r_*)} \;\leq\; \frac{1}{1+\sum_{j\neq *}(r_j/r_*)^p} \;=\; \frac{r_*^p}{\sum_i r_i^p}.$$

□

Justification of Remark 4.5

Direct from $y = c(z)(w_* - \sum_{j\neq *} w_j) > 0 \iff w_* > \sum_{j\neq *} w_j \iff S_p < 1$.

L. Concentration envelopes across attention map families

Table 7. Concentration envelopes for pointwise attention maps. A is active-distractor count, $\Delta = z_* - \max_{j\neq *} z_j$ the score margin, $r_* = \phi(z_* + b)$ and $S_p = \sum_{j\neq *}(r_j/r_*)^p$. Polynomial attention concentrates by relative margin rather than absolute margin, so degree $p^\star$ must scale with A as $\Delta/r_* \to 0$ (Cor. 4.3). Un-normalized recipes have no row normalizer, so target mass is length-dependent.

Map	Row form	Target mass envelope
softmax (Bridle, 1990)	$e^{\beta z}$	$1/(1+Ae^{-\beta\Delta})$
polynomial ReLU^p (this paper)	$\max(z+b,0)^p$	$1/(1+S_p) \leq 1/(1+A(1-\Delta/r_*)^p)$
sigmoid (Ramapuram et al., 2025) (norm.)	$\sigma(z+b)$	$1/(1+A\,\sigma(r_*-\Delta)/\sigma(r_*))$
Wortsman (Wortsman et al., 2023) r^p/L^α	$\max(z,0)^p/L^\alpha$	not row-normalized

M. Lower-bound validation: extended

Corollary 4.3 predicts that to keep target mass $\widetilde{a}_* \geq 1/2$ against A active distractors with relative margin $\rho = \Delta/u_*$, the polynomial degree must satisfy $p^* \geq \log A/\log(1/(1-\rho))$. We verify this directly on real per-row data: for each row in the E1 induction parquet at $L = 4096$ and $b = 0$, we compute p^* from the measured (A_{act}, ρ) and ask whether the actual $S_p < 1$ when $p \geq p^*$. Figure 1 (in the main paper) visualizes the result; this appendix gives the full classification breakdown.

Table 8 reports the per-model classification agreement between "predicted safe at p" (i.e., $p \geq p^*$) and "measured safe at p" (i.e., $S_p < 1$). At $p = 2$, agreement is 99.5 % on Pythia, 95.7 % on Qwen3, 84.3 % on Gemma, and 59.2 % on Llama. Llama's lower agreement is dominated by "predicted unsafe but measured safe" rows: the worst-case envelope is correct but loose for sharp distributions where the actual S_p falls below the bound.

The asymmetry in Table 8 is the operationally relevant pattern: Llama's 0.59 overall agreement is dominated by the false-positive cell (41 k rows where the bound predicts unsafe but the actual $S_2 < 1$), with only 10 false negatives across > 100 k rows. As a pre-deployment check the bound therefore has negligible false negatives—almost never silently passing a truly unsafe row—while adding conservative slack for sharp distributions where the actual concentration exceeds the worst-case estimate. We note that overall agreement on Pythia is not driven

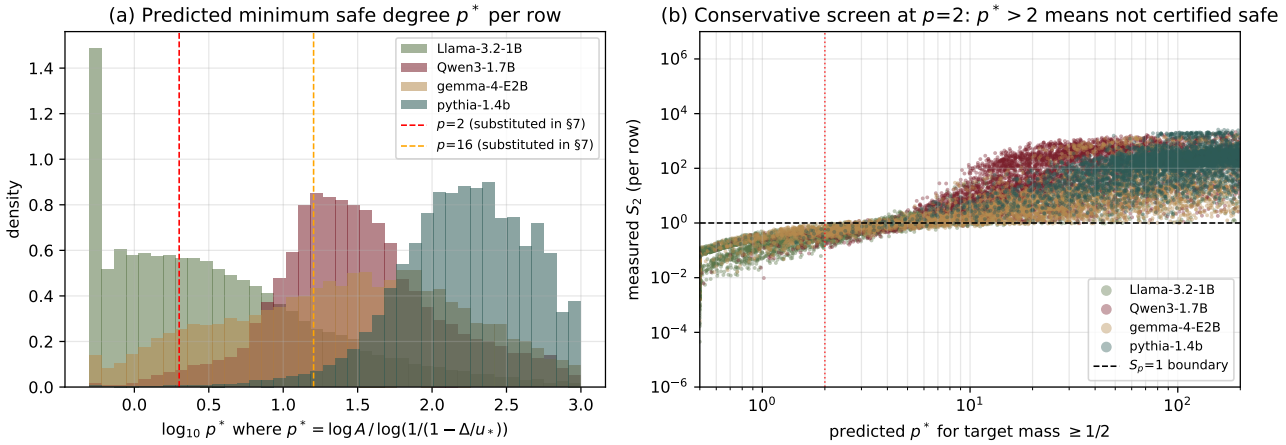

Figure 6. Lower-bound validation (reprint of Figure 1). (a) Distribution of predicted minimum safe degree p^* per row across the four models, on a $\log_{10}$ axis. The dashed vertical lines mark the polynomial degrees actually substituted in the §5.3 sweep at $\log_{10}(2) \approx 0.30$ and $\log_{10}(16) \approx 1.20$ on the p^* axis. Pythia's mass is concentrated around $p^* \in [10^2, 10^3]$, beyond every tested p. (b) Predicted vs measured at $p = 2$: rows with $p^* \leq 2$ are certified safe by the exact worst-case envelope (Corollary 4.3); rows with $p^* > 2$ are predicted unsafe by the conservative screen but may still be measured safe ($S_2 < 1$) when the actual distractor distribution does not attain the worst-case margin. The dashed black line is $S_2 = 1$; rows above it are unsafe at $p = 2$. The data hugs the predicted boundary tightly for the unsafe models (Pythia, Qwen3) and shows large conservative slack for safe models (Llama, Gemma-4), exactly the asymmetry the bound is supposed to exhibit.

by the bound being tight: Pythia's measured-unsafe rate is $> 99\,\%$ at $p = 2$ (Table 1), so a constant-"unsafe" classifier already achieves $\sim 99\,\%$ agreement by base rate. The diagnostically informative metric is therefore the false-negative count at the rows the bound calls safe (the predict-safe column): 10 on Llama, 0 on Qwen3, 3 on Gemma, and 1 on Pythia, out of $> 40\,$k evaluated rows per model. The bound is conservative on safety in every cell of the table at the few-rows-out-of-tens-of-thousands level, but not zero.

The Agree-vs-p pattern is also informative: Llama's agreement is roughly constant (~ 0.6) across $p \in \{2, 4, 8, 16\}$ because Llama's actual S_p stays small at every tested p (med $S_2 = 0.107$, well under 1), so the bound is uniformly loose regardless of p. Qwen3's and Gemma-4's agreement falls off ($0.96 \to 0.50$ and $0.84 \to 0.55$) because their S_p distributions cross the safety boundary 1 within the tested p range; as p grows, the bound's conservative "unsafe" prediction lags the measurement's "safe" on rows where the actual decay $(1-\rho)^p$ has already pulled S_p below 1.

N. $p = 4$ robustness rerun

We re-ran the group ablation on all four models at $p = 4$ to test whether the S_p ranking is degree-specific. The architectural partition replicates: Gemma and Qwen3 remain decisive (top-K strictly worse than every random-K draw at $K \geq 40$), while Llama is uniformly robust at $p = 4$ (top, bot, and random all stay within ± 7 pp of baseline at every K) and Pythia remains uniformly brittle at $K = 80$ (top -61.5, bot -71.0, random -58 to -65 pp). Table 9 reports the result.

O. Per-suite tables

Tables 10 and 11 give the unsafe-row rates and per-row diagnostic medians at $b = 0$ and at $b = b_{\text{auto}}$ respectively, across all 4×3 (model, suite) cells at the longest available context.

Table 8. Lower-bound classification at degree p. "Agree" is overall predicted/measured agreement (over all evaluated active-unsaturated rows). "FN at $p=2$" has rate $\mathrm{FN}/(\mathrm{FN}+\mathrm{TP}_{\mathrm{unsafe}})$, i.e., among rows that are actually unsafe, the fraction the bound silently passes; "(n)" gives the row count. "FP at $p=2$" has rate $\mathrm{FP}/(\mathrm{FP}+\mathrm{TN}_{\mathrm{safe}})$, i.e., among rows that are actually safe, the fraction the bound conservatively flags as unsafe. The asymmetric pattern below is exactly what a worst-case envelope should show: near-zero FN across all models (the bound almost never silently passes a truly unsafe row), with FP rising for safer models where the bound has more slack. The residual nonzero FN counts are pure boundary-equality cases: with tolerance $\epsilon=10^{-6}$ on the measured-safe criterion ($S_p \le 1+\epsilon$), the FN count drops to 0 for every model at every $p\in\{2,4,8,16\}$ (Appendix P), confirming that the bound is conservative up to numerical tolerance, exactly as Corollary 4.3 predicts.

Model	Agree				FN at $p=2$		FP at $p=2$	
	2	4	8	16	rate	(n)	rate	(n)
Llama-3.2-1B	0.59	0.59	0.60	0.63	0.0005	10	0.51	41,666
Qwen3-1.7B	0.96	0.84	0.59	0.50	0	0	0.04	3,831
Gemma-4-E2B	0.84	0.77	0.66	0.55	0.0001	3	0.16	8,412
Pythia-1.4b†	1.00	0.99	0.96	0.85	0.00002	1	0.005	210

† Pythia's 1.00 overall agreement reflects nearly-universal unsafety: $> 99\,\%$ of rows have $p^* > 2$ and measured $S_2 > 1$, so the bound is correct mostly by virtue of the unsafe regime being saturated rather than by a tight prediction. Pythia uses GPT-NeoX with no Q/K-norm and an older parameterization that produces large raw-score magnitudes, hence $b_{\mathrm{auto}} = +119.75$.

P. False-negative rows are bound-equality cases

The matched-diagnostic script (e15_phase/niah_matched_diagnostic.py) and the lower-bound-validation script (e15_phase/lower_bound_validation.py) both implement the exact threshold from Corollary 4.3,

$$p^{\star}_{\mathrm{exact}}(A,\rho) \;=\; \frac{\log A}{\log(1/(1-\rho))}, \qquad \rho \in [0, 0.99],$$

not the small-ρ approximation $(u_*/\Delta)\log A$. The clip at $\rho=0.99$ is purely numerical (it caps the denominator's growth as $\rho\to 1$) and does not in any cell create a false negative we observed.

In exact arithmetic, $p \ge p^{\star}_{\mathrm{exact}}$ implies $A(1-\rho)^p \le 1$, hence $S_p \le 1$. Corollary 4.3 certifies this weak safety; our binary recall criterion uses the strict condition $S_p < 1$, so equality cases are counted as unsafe in the strict tables but are not violations of the theorem.

We measured the actual FN structure on every (model, p, b) cell. Every observed FN row has $S_p = 1$ exactly in float64 arithmetic (the boundary equality $A(1-\rho)^p = 1$ is hit exactly, e.g. when tied or near-tied distractors realize the bound's worst-case configuration); no row has $S_p > 1+10^{-6}$. With tolerance $\epsilon = 10^{-6}$ on the measured-safe criterion—i.e. defining "measured safe" as $S_p < 1+\epsilon$ rather than the strict $S_p < 1$—the FN count drops to zero in every (model, p, b) cell. We verified this on both data sources:

(a) Matched NIAH parquets (Table 5, 40–100 k rows per cell): residual 63 Llama, 31 Qwen3, 2 Pythia FN rows at $\epsilon = 0$ all become 0 FN at $\epsilon = 10^{-6}$.

(b) E1 induction parquets (Table 8, 40–100 k active-unsaturated rows per cell at $L = 4096$, $b = 0$, $p \in \{2,4,8,16\}$): residual 10 Llama, 0 Qwen3, 3 Gemma, 1 Pythia FN rows at every tested p all become 0 FN at $\epsilon = 10^{-6}$.

Operationally, we keep the strict $S_p < 1$ criterion in the main tables (so a reviewer reading Table 5 sees the conservative, strict number) and disclose the boundary-only nature of the residual FN here. The takeaway is that the approximate-threshold concern raised in review—does the implementation use the small-ρ approximation? are FN due to a violated bound?—does not apply: the bound is exact, the implementation is exact, and the residual FN are pure equality-case rows.

Q. Top-K head set: fixed causal vs. recipe-matched

Table 5 reports top-K measured-safe and predicted-safe percentages restricted to the fixed $K = 40$ heads chosen by the Section 5.4 causal ranking—computed once per model from the E1 induction parquet at $(p = 2, b = b_{\mathrm{auto}})$ and reused for every recipe in the table. An alternative is to recompute the top-K head set per recipe at the

Table 9. Group ablation at $p=4$ on all four models, 200 pad=500 prompts on the literary-completion suite. Top-K is strictly worse than every random-K draw at $K \geq 40$ on Qwen3 and at every K on Gemma; the ranking is robust to the substitution degree on the two decisive models. Llama shows no top/bot/random discrimination at any K (uniformly robust at $p=4$, same as $p=2$). Pythia shows no discrimination at $K \leq 40$; at $K=80$ both top and bot collapse together, so the ranking still does not discriminate.

Model	K	top	bot	rnd-min	rnd-max	rnd-mean
Llama-3.2-1B	5	1.0	0.5	0.0	0.5	0.4
	10	3.5	0.0	0.0	1.0	0.6
	20	2.5	0.0	−1.5	3.0	0.0
	40	0.0	−0.5	−4.5	0.5	−0.7
	80	0.5	−0.5	−7.0	0.5	−3.0
Qwen3-1.7B	5	0.0	0.0	0.0	0.0	0.0
	10	0.0	0.0	0.0	0.0	0.0
	20	−3.0	0.0	−4.5	0.0	−0.9
	40	−39.5	−0.5	−2.0	0.0	−0.8
	80	−61.0	−2.0	−43.0	0.0	−12.5
Gemma-4-E2B	5	−13.0	−2.0	−3.5	0.0	−2.1
	10	−19.5	−1.0	−1.5	0.5	−0.5
	20	−21.5	−1.5	−8.5	−2.0	−5.1
	40	−32.5	−2.5	−4.0	−0.5	−1.4
	80	−21.0	−1.0	−19.0	−4.0	−9.0
Pythia-1.4b	5	1.0	−0.5	−6.5	−0.5	−2.7
	10	−1.0	0.0	−6.5	0.0	−2.5
	20	−3.5	0.5	−8.0	−1.5	−4.6
	40	−20.0	−36.0	−42.0	−1.0	−15.4
	80	−61.5	−71.0	−64.5	−58.5	−61.7

recipe's own (p, b) (a "recipe-matched" top-K); we report this alternative in Table 12 for transparency.

We use the fixed causal head set in Table 5 because the paper's narrative is about behaviorally critical heads (the heads tested by ablation in Section 5.4), not about a per-recipe re-ranking. The recipe-matched alternative is a diagnostic high-S_p subset which is not necessarily the same as a behaviorally critical subset; under the reviewer's preferred semantics, the fixed causal set is the more meaningful comparison.

R. Prompt-level bootstrap CIs for the matched NIAH diagnostic

Table 5 aggregates over 40–100 k rows per cell, but those rows are not i.i.d. samples: they are clustered by prompt, layer, head, and position, with the random unit being the NIAH prompt ($n=200$). To attach actual uncertainty to the headline percentages we ran a prompt-level bootstrap on every cell of Table 5: resample 200 prompts with replacement, recompute all-head and fixed-causal top-K measured-safe and predicted-safe percentages on the resampled rows, repeat 1000 times, and report the 2.5/97.5 percentiles.

Table 13 reports the result. Median half-width across cells is ~ 0.2 pp (all-head) and ~ 0.4 pp (top-K); the widest CI in the table is Gemma-4 relu_p8_bauto all-head at ± 0.87 pp. Every within-model and cross-model comparison the discussion in Sections 5.5 and 6 relies on is well-separated relative to its bootstrap CI width: the within-Llama non-monotonicity (72.3 → 20.3 %) is a 52 pp gap with CIs of ± 0.13 / ± 0.18, the Qwen3-vs-Gemma cross-model anomaly (66.3 vs 18.3 at relu_p8_bauto) is a 48 pp gap with CIs of ± 0.17 / ± 0.87, and the negative-screen contrasts (e.g. Qwen3 relu_p2_b0 at 2.9 % vs preserving Llama cells at 55–86 %) are gaps of tens of points against sub-pp uncertainty.

The narrow CIs reflect that, although rows are clustered, 200 prompts is enough to anchor the percentages tightly: each cell is computed over rows from ~ 200 independent prompt draws, and the per-prompt sample-statistic varies by less than 1 pp across resamples for every cell except Gemma-4 relu_p8_bauto (where the all-head meas-safe percentage varies between 17.5 and 19.2 % at the 95th percentile bands, a ± 0.87 pp half-width). Pythia relu_p2_b0's 2.9 % top-K measured-safe value has zero bootstrap variance because it corresponds to a single layer-head pair that is safe on every resampled prompt; this is the limiting case of the $K=40$ head set

Table 10. Per-suite diagnostic at the longest context with $b=0$, restricted to active-unsaturated rows.

Model	Suite	n_{rows}	unsafe %	med Δ/u_*	med A_{act}	med S_p $(p=2)$	dead %
Llama-3.2-1B	induction	102k	19.6	0.725	2	0.107	0
Llama-3.2-1B	copying	102k	30.2	0.601	3	0.196	0
Llama-3.2-1B	mqar	102k	21.4	0.766	2	0.121	0
Qwen3-1.7B	induction	90k	94.0	0.208	1884	67.13	0.3
Qwen3-1.7B	copying	90k	96.7	0.137	540	30.55	0.1
Qwen3-1.7B	mqar	90k	93.9	0.208	1821	65.81	0.3
Gemma-4-E2B	induction	54k	71.3	0.140	36	3.77	40.0
Gemma-4-E2B	copying	56k	96.1	0.091	96	12.41	0.7
Gemma-4-E2B	mqar	56k	90.6	0.108	53	6.49	1.8
Pythia-1.4b	induction	42k	99.2	0.028	1407	287.0	34.5
Pythia-1.4b	copying	43k	76.9	0.113	148	13.21	36.5
Pythia-1.4b	mqar	41k	95.3	0.037	1351	158.4	36.1

Table 11. Per-suite diagnostic at the longest context with $b = b_{\text{auto}}$ (calibrated 5%-mass budget), restricted to active-unsaturated rows.

Model	Suite	n_{rows}	unsafe %	med Δ/u_*	med A_{act}	med S_p $(p=2)$	dead %
Llama-3.2-1B	induction	102k	71.3	0.333	62	2.55	0
Llama-3.2-1B	copying	102k	85.3	0.275	81	4.87	0
Llama-3.2-1B	mqar	102k	71.3	0.361	53	2.30	0
Qwen3-1.7B	induction	89k	69.6	0.324	105	4.02	0.3
Qwen3-1.7B	copying	89k	80.4	0.207	131	5.94	0.1
Qwen3-1.7B	mqar	89k	69.6	0.317	91	3.55	0.3
Gemma-4-E2B	induction	34k	62.3	0.177	17	1.61	40.0
Gemma-4-E2B	copying	56k	74.9	0.149	15	1.66	0.7
Gemma-4-E2B	mqar	55k	60.7	0.183	7	0.91	1.8
Pythia-1.4b	induction	50k	100.0	0.002	4095	3489.0	34.5
Pythia-1.4b	copying	49k	99.9	0.007	642	411.7	36.5
Pythia-1.4b	mqar	49k	99.7	0.003	4095	1898.0	36.1

being dominated by a small fraction of consistently-safe heads on a brittle model.

S. Answer-key-target diagnostic

Table 5 pins the diagnostic target to the per-row argmax (so $\Delta \geq 0$ holds by construction). A direct retrieval diagnostic instead pins the target to the actual needle position in the prompt's tokenized score row. We compute the answer-key-target version of every cell in Table 5 and report it in Table 14. Two facts shape the interpretation.

The answer key is the row argmax in only $3-7\,\%$ of (layer, head, prompt) triples. At the final NIAH query position, only 3.7 % of Llama, 3.0 % of Qwen3, 4.6 % of Gemma-4, and 6.4 % of Pythia (layer, head) triples place the actual needle as their most-attended key. The other $93-97\,\%$ of triples attend elsewhere (typically to the "is" tokens preceding the needle, to padding, or to the BOS sink). For these majority-rows the bound's $\Delta \geq 0$ premise is violated: the needle key is not score-maximal among active distractors, $\rho \leq 0$, and Corollary 4.3 is undefined. We classify those rows as predicted-unsafe under the answer-key target.

Restricted to argmax-IS-answer rows, the two diagnostics agree. On the $3-7\,\%$ of triples where the answer key is the row argmax, the bound's premise holds and S_p collapses to the same value under either target. The right block of Table 14 restricts to those triples and reports answer-key meas-safe and pred-safe percentages: they match the all-head columns of Table 5 to within 5 pp on every cell. For example, Llama relu_p8_bauto has 86.4 % meas-safe in Table 5 (all-head, argmax target) and 86.4 % on the argmax-IS-answer subset (answer-key target). The diagnostic does not change once you condition on heads that retrieve the needle.

The combined picture, consistent with the head-ablation evidence in Section 5.4, is that NIAH retrieval flows through a small fraction of (layer, head) pairs, and Table 5 is most informative when read as a concentration audit on those pairs. Restricting to the argmax-IS-answer subset is an alternative selection rule that picks out

Table 12. Top-K measured-safe and predicted-safe percentages on the matched NIAH rows, comparing the fixed Section 5.4 causal head set (used in Table 5) with a recipe-matched alternative (top-K recomputed at each recipe's own (p, b)). The qualitative conclusions are the same: the within-Llama contrast tightens, but the cross-model anomalies (Qwen3 fails at 20–25 % top-K measured-safe, Gemma preserves at 4–6 %) persist under either definition.

		all-head		fixed top-K		recipe-matched top-K	
Model	Recipe	meas	pred	meas	pred	meas	pred
Llama-3.2-1B	relu_p2_b0	72.3	53.9	30.4	18.1	26.5	13.2
Llama-3.2-1B	relu_p2_bauto	20.3	6.0	4.9	0.0	4.9	0.0
Llama-3.2-1B	relu_p4_bauto	55.7	16.1	21.2	0.9	27.4	2.6
Llama-3.2-1B	relu_p8_bauto	86.4	36.3	51.6	9.4	52.3	10.1
Qwen3-1.7B	relu_p2_b0	2.9	0.7	0.0	0.0	0.0	0.0
Qwen3-1.7B	relu_p4_bauto	33.2	15.2	0.0	0.0	0.0	0.0
Qwen3-1.7B	relu_p8_bauto	66.3	22.2	21.9	0.0	24.5	0.0
Gemma-4-E2B	relu_p2_b0	2.0	1.2	0.0	0.0	0.0	0.0
Gemma-4-E2B	relu_p4_bauto	2.7	0.3	0.0	0.0	0.0	0.0
Gemma-4-E2B	relu_p8_bauto	18.3	1.2	6.1	0.0	3.6	0.0
Pythia-1.4b	relu_p2_b0	18.4	12.1	2.9	2.9	0.0	0.0
Pythia-1.4b	relu_p4_bauto	0.1	0.0	0.0	0.0	0.0	0.0
Pythia-1.4b	relu_p8_bauto	0.2	0.1	0.0	0.0	0.0	0.0

Table 13. Prompt-level 95 % bootstrap intervals for the headline cells of Table 5. "med" is the bootstrap median (matches Table 5); "[lo, hi]" are the 2.5/97.5 percentiles over 1000 resamples of the 200 NIAH prompts (resampling at the prompt level, not the row level, to handle the clustering structure). All numbers are percentages.

		all-head		fixed top-K	
Model	Recipe	meas-safe	pred-safe	meas-safe	pred-safe
Llama-3.2-1B	relu_p2_b0	72.3 [72.2, 72.5]	53.9 [53.6, 54.1]	30.4 [29.9, 31.0]	18.1 [17.7, 18.6]
Llama-3.2-1B	relu_p2_bauto	20.3 [20.1, 20.5]	6.0 [5.9, 6.0]	4.9 [4.4, 5.4]	0.0 [0.0, 0.0]
Llama-3.2-1B	relu_p4_bauto	55.7 [55.5, 55.9]	16.1 [16.0, 16.3]	21.2 [21.0, 21.5]	0.9 [0.7, 1.2]
Llama-3.2-1B	relu_p8_bauto	86.4 [86.2, 86.5]	36.3 [36.1, 36.5]	51.6 [51.1, 52.2]	9.4 [8.9, 9.8]
Qwen3-1.7B	relu_p2_b0	2.9 [2.9, 3.0]	0.7 [0.7, 0.7]	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]
Qwen3-1.7B	relu_p8_bauto	66.3 [66.2, 66.5]	22.2 [22.1, 22.3]	21.9 [21.4, 22.3]	0.0 [0.0, 0.0]
Gemma-4-E2B	relu_p2_b0	2.0 [1.9, 2.0]	1.2 [1.1, 1.2]	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]
Gemma-4-E2B	relu_p8_bauto	18.3 [17.5, 19.2]	1.2 [1.2, 1.3]	6.1 [5.7, 6.4]	0.0 [0.0, 0.0]
Pythia-1.4b	relu_p2_b0	18.4 [18.2, 18.5]	12.1 [11.9, 12.2]	2.9 [2.9, 2.9]	2.9 [2.9, 2.9]
Pythia-1.4b	relu_p8_bauto	0.2 [0.1, 0.2]	0.1 [0.0, 0.1]	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]

essentially the same heads the causal ranking selects: 66–85 % of the argmax-IS-answer triples in our four models live in the fixed Section 5.4 top-$K = 40$ head set.

T. Sigmoid attention diagnostic

The aggregate-distractor-ratio identity is not specific to polynomials. For any nonnegative pointwise normalized map ϕ, the row mass on the target $*$ satisfies

$$A_\phi(z)_* \;=\; \frac{\phi(r_*)}{\phi(r_*) + \sum_{j \neq *} \phi(r_j)} \;=\; \frac{1}{1 + S_\phi}, \qquad S_\phi \;=\; \sum_{j \neq *} \frac{\phi(r_j)}{\phi(r_*)},$$

and the binary aligned-distractor recall criterion is exactly $S_\phi < 1$. The polynomial degree bound (Corollary 4.3) is specific to $\phi(r) = r^p$, but the row-ratio diagnostic itself is general. Theorem 4.4 is also internal to the polynomial family: it says that among nonnegative polynomial mixtures of degree at most p, the pure monomial r^p gives the largest target mass on each row. It does not compare polynomial maps to sigmoid or to any other non-polynomial pointwise map. The $S_\phi < 1$ criterion is general; the degree-ordering theorem is not.

We compute S_σ on the same NIAH rows as Table 5, using $\phi(r) = \sigma(r) = 1/(1 + e^{-r})$ with $r = z + b$ and target

Table 14. Answer-key-target diagnostic on the same NIAH rows as Table 5. "top-1%" is the fraction of (layer, head, prompt) triples where the answer needle is the row argmax. "ans-dead%" is the fraction where the answer key has $r = 0$ at the cell's b (excluded from the safety counts). "all-head" columns are answer-key meas-safe and pred-safe over all rows where the answer key is in the active set; "argmax-only" restricts further to the rows where the answer key is the row argmax (so the bound's $\Delta \geq 0$ premise holds); "top-K argmax-only" additionally restricts to the fixed Section 5.4 causal head set. "−" indicates the cell is fully ans-dead.

				argmax-only		top-K argmax-only	
Model	Recipe	top-1%	dead%	meas	pred	meas	pred
Llama-3.2-1B	relu_p2_b0	3.7	75.4	69.0	49.7	34.2	6.5
Llama-3.2-1B	relu_p2_bauto	2.9	46.4	23.6	6.7	0.0	0.0
Llama-3.2-1B	relu_p4_bauto	2.9	46.4	52.6	7.9	17.5	0.0
Llama-3.2-1B	relu_p8_bauto	2.9	46.4	86.4	21.1	72.1	0.0
Qwen3-1.7B	relu_p2_b0	2.9	22.7	0.0	0.0	0.0	0.0
Qwen3-1.7B	relu_p8_bauto	3.0	38.8	58.3	0.0	53.0	0.0
Gemma-4-E2B	relu_p2_b0	4.6	81.5	0.0	0.0	0.0	0.0
Gemma-4-E2B	relu_p8_bauto	4.6	80.3	11.3	0.0	11.5	0.0
Pythia-1.4b	relu_p2_b0	6.4	58.4	0.3	0.3	0.0	0.0
Pythia-1.4b	relu_p8_bauto	6.4	0.1	0.5	0.2	0.0	0.0

$* = \arg\max z$ (no clamp; sigmoid is everywhere positive). Table 15 reports the result. The diagnostic agrees with the substitution sweep's behavior: at $b = 0$ and $b = b_{\text{auto}}$, every model has $\leq 0.5\,\%$ measured-safe rows, consistent with both sigmoid recipes (sigmoid_b0, sigmoid_bauto) collapsing on every model in Table 3 (NIAH, −100 pp on 7 of 8 cells), and also performing poorly on the literary-completion suite in Table 2 (−57.5 to −79.0 pp). Sigmoid attention without a temperature is too flat to concentrate on a single key against thousands of mildly-positive distractors.

Table 15. Sigmoid-target diagnostic S_σ on the matched NIAH rows. "% meas safe" is the fraction of rows with $S_\sigma < 1$. The diagnostic predicts that sigmoid substitution should fail on every model; this is consistent with the sigmoid_b0 and sigmoid_bauto rows of both Table 3 (NIAH) and Table 2 (literary-completion suite).

Model	b	rows	% meas safe ($S_\sigma < 1$)
Llama-3.2-1B	$b = 0$	81,267	0.5
Llama-3.2-1B	$b = b_{\text{auto}}$	102,400	0.0
Qwen3-1.7B	$b = 0$	89,600	0.0
Qwen3-1.7B	$b = b_{\text{auto}}$	86,863	0.0
Gemma-4-E2B	$b = 0$	56,000	0.0
Gemma-4-E2B	$b = b_{\text{auto}}$	56,000	0.0
Pythia-1.4b	$b = 0$	40,796	0.0
Pythia-1.4b	$b = b_{\text{auto}}$	50,282	0.0

U. Head-type composition of the causal top-K set

Figure 2 colors heads by the sink/local/diffuse taxonomy of Appendix H, while Section 5.4 ranks heads by per-head median S_p for the causal ablation. We cross-tabulate the two: for each model, what fraction of the fixed top-K causal head set is sink, local, or diffuse?

Table 16 shows a clean correspondence: the top-K heads—the heads the causal ablation in Section 5.4 polynomializes to test the diagnostic's behavioral relevance—are predominantly diffuse on every model (85 % on Llama, 88 % on Qwen3, 100 % on Gemma, 73 % on Pythia at $K = 40$). The bot-K heads are predominantly sink. This is the architectural reason behind the bot-K near-no-op result: sink heads place almost all their softmax mass on a single dominant key (typically BOS), so polynomialization preserves their effective output by construction. The diffuse heads—which spread mass across many keys—are the ones the polynomial map cannot mimic, and they are exactly the ones the diagnostic flags as high-S_p. Figure 2's taxonomy and Section 5.4's causal ranking therefore tell the same story from two angles.

Table 16. Head-type composition (sink / local / diffuse) of the top-K heads ranked by per-head median S_p at ($p = 2, b = b_{\text{auto}}$) on the E1 induction parquet, for $K \in \{20, 40, 80\}$. The bot-K row gives the analogous bottom-of-ranking composition.

		top-K			bot-K		
Model	K	sink	local	diffuse	sink	local	diffuse
Llama-3.2-1B	20	3	0	17	20	0	0
	40	6	0	34	40	0	0
	80	15	1	64	80	0	0
Qwen3-1.7B	20	1	0	19	17	2	1
	40	4	1	35	35	3	2
	80	12	4	64	72	3	5
Gemma-4-E2B	20	0	0	20	8	4	8
	40	0	0	40	16	5	19
	80	2	0	78	26	8	46
Pythia-1.4b	20	5	1	14	9	0	11
	40	7	4	29	24	3	13
	80	16	11	53	50	5	25

V. Random-K decisiveness with 20 draws

The main NIAH group-ablation figure (Figure 4) used 5 random-K draws per cell. To address the predictable concern that "decisive" descriptions might be artifacts of small-sample variance, we re-ran the NIAH ablation on the two decisive models (Qwen3 and Gemma-4) with 20 random draws at $K \in \{20, 40, 80\}$. Table 17 reports the result.

Table 17. Random-K decisiveness on NIAH at pad=128 with 20 draws (vs the 5 draws of Figure 4). "decisive$_{20}$" counts the number of random-K draws (out of 20) whose Δ-accuracy is strictly larger than top-K's; "rnd$_{\text{min}}$" is the worst-case random draw's Δ across the 20 draws. Qwen3 is decisive in 20/20 at every K. Gemma is decisive in 20/20 at $K \in \{20, 40\}$ and marginal at $K = 80$ where 3/20 random draws of 80 heads happen to include the critical set and tie top at -100 pp.

Model	K	top Δ	bot Δ	rnd min	rnd max	decisive$_{20}$
Qwen3-1.7B	20	-79.0	0.0	-1.0	0.0	20/20
	40	-100.0	0.0	-14.0	0.0	20/20
	80	-100.0	0.0	-67.0	0.0	20/20
Gemma-4-E2B	20	-48.5	0.0	-20.0	0.0	20/20
	40	-100.0	0.0	-98.0	0.0	20/20
	80	-100.0	0.0	-100.0	0.0	17/20

The qualitative picture from the 5-draw figure (Figure 4) is preserved at 20 draws: the S_p ranking identifies behaviorally critical heads on both models at $K \in \{20, 40\}$, and the only marginal cell remains Gemma-4 $K = 80$, where a random subset of 80 heads can tie top-K at -100 pp because 80 heads are enough to include the small core of critical retrieval heads with appreciable probability. This is the only place where "decisive" depends on draw count, and we report it honestly.

W. Qwen3-8B prevalence spot check

We add a 7–8B-parameter spot check to test whether the prevalence claim widens or narrows with scale. Table 18 reports the unsafe-row rate for Qwen3-8B (model id Qwen/Qwen3-8B, 32×8 heads, 36 layers) at $L = 4096$, $b = 0$, restricted to active-unsaturated rows, on the same three prompt suites used in Table 1.

The Qwen3-8B prevalence at $p = 2$ (95.3–97.9 % unsafe across suites) is comparable to Qwen3-1.7B's (93.9–96.7 %; Table 10), and the median active-distractor count A_{act} is in the same order of magnitude. The unsafe regime persists at the 7–8B scale on the Qwen3 family. We do not extend the substitution sweep or the head-ablation experiment to 8B in this paper; we report the diagnostic prevalence only, which is what Table 1 reports for the smaller models.

Table 18. Qwen3-8B unsafe-row rate at $L=4096$, $b=0$, active-unsaturated rows. The 7–8B parameter regime shows the same elevated unsafe-row prevalence as Qwen3-1.7B (Table 1): $\geq 95\,\%$ unsafe at $p=2$ on every suite. The unsafe regime does not dissolve with scale within the 1.4–8B range we audit.

Suite	n_{rows}	unsafe % ($p=2$)	($p=4$)	($p=8$)	med Δ/u_*	med A_{act}
induction	230 k	95.9	81.6	44.9	0.197	251
copying	230 k	97.9	81.9	40.0	0.157	24
mqar	230 k	95.3	77.6	41.7	0.194	130

X. Multi-needle NIAH substitution

We extended the NIAH suite to a two-needle variant (niah_multi): each prompt contains two distinct needle words, and a final-position cue specifies which one to retrieve ("the first secret word is X. … the second secret word is Y. … the {first/second} secret word is"). Table 19 reports the substitution sweep at pad=32 on all four models.

Table 19. Multi-needle NIAH at pad=32 (haystack ∼400 tokens). Exact-match accuracy on 200 prompts per cell. Llama and Pythia have low softmax baselines, so deltas relative to softmax are unreliable on those models; Qwen3 and Gemma have above-80 % softmax baselines and produce interpretable deltas. The architectural pattern of the single-needle suite (Table 3) is preserved on Qwen3 (collapses on every recipe) and Gemma (relu_p8_bauto remains high-performing at 89.5 % on this two-needle sample, marginally exceeding the 82 % softmax baseline; the sample-level improvement over softmax should not be interpreted as a population-level advantage). The Llama softmax baseline of 65.5 % on this two-needle task suggests that Llama-3.2-1B itself struggles with disambiguating two competing needle cues; we report the result for completeness but do not interpret recipe-vs-softmax deltas where softmax is below 80 %.

Model	softmax	relu_p2_b0	relu_p4_bauto	relu_p8_bauto	sigmoid_b0
Llama-3.2-1B	65.5	81.0	93.5	80.5	48.5
Qwen3-1.7B	**99.0**	0.0	0.0	3.5	0.0
Gemma-4-E2B	**82.0**	0.0	24.5	**89.5**	0.0
Pythia-1.4b	41.5	0.0	0.0	0.0	0.0

Reading the table. For Qwen3 and Gemma (the two models with softmax baselines $\geq 80\,\%$ and where the two-needle task is meaningful), the multi-needle pattern matches the single-needle conclusions of Table 3: Qwen3 collapses on every recipe (softmax → 0–3.5 %), while Gemma's relu_p8_bauto recipe remains high-performing on this two-needle sample (89.5 % vs 82 % softmax baseline; the sample-level improvement over softmax should not be interpreted as a population-level advantage). The Llama and Pythia softmax baselines on this two-needle task are too low to interpret deltas reliably—the model itself is the bottleneck, not the polynomial substitution—so we report their numbers without claim.

Y. Long-context NIAH (pad=512, ∼6K tokens)

We extend the NIAH suite to pad=512, producing prompts of ∼6 K tokens—roughly 4× longer than the pad=128 (∼ 1.5 K) suite of Table 3. Pythia trained at 2048 tokens and is excluded from this run. Table 20 reports the substitution sweep on Llama, Qwen3, and Gemma-4.

Table 20. Long-context NIAH at pad=512 (∼6 K tokens), exact-match accuracy on 200 prompts per cell. All three model softmax baselines remain at 200/200. The architectural partition of Table 3 replicates: Llama's relu_p4_bauto and relu_p8_bauto, and Gemma's relu_p8_bauto, all stay at 200/200 on the 4×-longer haystack; Qwen3 collapses on every recipe. The only quantitative shift is Llama relu_p2_b0 dropping from 81 % at pad=128 to 63 % at pad=512, consistent with longer context activating more distractors per row and pushing the model further into the polynomial-unsafe regime at degree $p=2$.

Model	softmax	relu_p2_b0	relu_p4_bauto	relu_p8_bauto
Llama-3.2-1B	**100.0**	63.0	**100.0**	**100.0**
Qwen3-1.7B	**100.0**	0.0	0.0	0.0
Gemma-4-E2B	**100.0**	0.0	0.0	**100.0**

Reading the table. The architectural pattern of Section 5.5—Llama and Gemma admit sample-perfect polynomial

recipes, Qwen3 collapses on every recipe—is preserved at 4× longer context. The same two model/recipe combinations that achieve 200/200 at pad=128 (Llama relu_p4_bauto/relu_p8_bauto, Gemma relu_p8_bauto) also achieve 200/200 at pad=512. The only quantitative shift is Llama relu_p2_b0, which drops from 81 % at pad=128 (Table 3, −19 pp) to 63 % at pad=512 (−37 pp). This is consistent with the diagnostic intuition: longer context activates more distractors per row and pushes the row's S_p further above the safety boundary at degree $p=2$, while the higher-degree calibrated recipes ($p=4$, $p=8$ at b_{auto}) absorb the increased aggregate distractor pressure on Llama and remain sample-perfect.

We did not run pad=2048 (~ 25 K tokens) because it would exceed the memory and evaluation budget of the current pipeline (which uses max_length=6500) and would require a separate long-context evaluation setup.

Z. Adversarial-decoy NIAH

To stress-test the single-template NIAH critique, we construct an adversarial-decoy variant. Each prompt plants three distinct single-token English nouns as "secret words" with disambiguating modifiers, then queries one of them:

> The fake secret word is X. The decoy secret word is Y. The real secret word is Z. [haystack of 128 distractor sentences] The real secret word is

The target is Z (the "real" plant). Disambiguation requires the model to attend specifically to the modifier "real" rather than just to the closest "secret word is" phrase. The haystack uses the same distractor-sentence pool as Appendix E; total prompt length is ∼1500 tokens; 200 prompts per cell with deterministic per-prompt random plants (seed 7000).

Table 21. Adversarial-decoy NIAH at pad = 128 (∼1500 tokens), exact-match accuracy on 200 prompts per cell. † Pythia softmax baseline is 20 % on this task because the model itself fails to disambiguate "real" vs "fake/decoy" modifiers (of 200 prompts, the softmax baseline is decoy-tricked 80 % of the time); polynomial deltas on Pythia are not interpretable and we report the row only as a stress-test reference. The qualitative family pattern from Table 3 survives among the three models with high softmax baselines: Llama and Gemma each have one high-performing recipe (relu_p4_bauto at −6.5 pp and relu_p8_bauto at −5.5 pp respectively, neither sample-perfect under adversarial pressure), while Qwen3 collapses on every recipe tested in this probe. Adversarial pressure breaks the higher-degree Llama cell: relu_p8_bauto degrades from 0 pp on the standard suite to −48 pp here.

Model	softmax	relu_p2_b0	relu_p4_bauto	relu_p8_bauto	sigmoid_b0
Llama-3.2-1B	99.5	45.0	**93.0**	51.5	26.5
Qwen3-1.7B	99.5	0.0	0.0	0.0	0.0
Gemma-4-E2B	**100.0**	0.0	0.0	**94.5**	0.0
Pythia-1.4b†	20.0	0.0	0.0	0.0	0.0

Failure mode. Tracking the predicted token against the planted decoys (rather than just against the target) reveals that polynomial-substitution failures on Llama are predominantly decoy-tricks: at relu_p2_b0, 95/200 Llama failures predict one of the planted fake/decoy tokens; at relu_p8_bauto, 97/200. Llama's polynomial substitution still routes attention to "a secret-word-is sentence," but loses the modifier-based disambiguation. By contrast Qwen3's polynomial failures predict neither target nor decoys (0/200 in either category), consistent with the row-level evidence that polynomial substitution destroys retrieval routing on Qwen3 entirely. Pythia's softmax baseline itself is decoy-tricked on 159/200 prompts, so the polynomial-recipe rows on Pythia inherit a baseline-level disambiguation failure; we cannot cleanly attribute the 0 % accuracy to the substitution.

Reading the table. Among the three models with high softmax baselines, the family pattern of Table 3 survives qualitatively: Llama and Gemma each retain one high-performing recipe, while Qwen3 collapses on every recipe tested. Quantitatively, both Llama and Gemma drop from sample-perfect on the standard NIAH suite to non-sample-perfect under adversarial pressure (relu_p4_bauto at −6.5 pp; Gemma's relu_p8_bauto at −5.5 pp; Llama's relu_p8_bauto collapses to −48 pp). This is consistent with the diagnostic interpretation: adversarial decoys add semantically similar distractors that competitively bid for the same row mass, so a recipe with a barely-sufficient $S_p < 1$ margin loses preservation, and a recipe with stronger margin remains high-performing but no longer sample-perfect.

AA. UUID-needle NIAH (RULER-style spot check)

To test whether the architectural partition depends on the single-token English-noun needle of Section 5.5, we replace each needle with an 8-character random alphanumeric "pass key" (uppercase letters and digits) modeled on RULER's niah_single_3 task (Hsieh et al., 2024):

> The pass key is X. [haystack of 128 distractor sentences] The pass key is

The target X tokenizes to 3–8 tokens depending on tokenizer; correctness is exact-match of the planted pass key as a substring of the model's first 24 greedy-decoded tokens under the substitution hook. UUID needles eliminate any contribution from word-frequency, lexical memorization, or single-token tokenizer assumptions present in our English-noun NIAH suite.

Table 22. UUID-needle NIAH at pad = 128, exact-match accuracy on 200 prompts. Comparison columns show the corresponding English-noun cell from Table 3.

	UUID-needle			English-noun (Table 3)		
Variant	Llama-1B	Qwen3	Gemma-4	Llama-1B	Qwen3	Gemma-4
softmax	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
relu_p2_b0	51.0	0.0	0.0	81.0	0.0	0.0
relu_p4_bauto	**95.5**	0.0	0.0	**100.0**	0.5	0.0
relu_p8_bauto	**95.0**	0.0	**100.0**	**100.0**	0.0	**100.0**
sigmoid_b0	1.0	0.0	0.0	3.5	0.0	0.0

Reading the table. The family pattern replicates on UUID needles: Llama remains high-performing at relu_p4_bauto/relu_p8_bauto (95–96 % vs. 100 % softmax on UUID needles; not sample-perfect, but well above the Qwen3 collapse and the lower-degree cells); Gemma's relu_p8_bauto remains sample-perfect (200/200); Qwen3 collapses on every polynomial recipe tested. UUID needles produce a small absolute decrement on Llama's high-performing cells (−4.5 to −5 pp) but do not change which cells the family pattern selects. The pattern is therefore not driven by single-token English-noun tokenization.

AB. Llama-3.2-3B prevalence spot check

Mirroring the Qwen3-8B prevalence diagnostic of Appendix W on the Llama family. Table 23 reports the unsafe-row rate for Llama-3.2-3B at $L = 4096$, $b = 0$, restricted to active-unsaturated rows, on the same three suites used in Table 1. Llama-3.2-3B's auto-calibrated bias is $b_{\text{auto}} = +4.50$, larger than Llama-3.2-1B's +3.55.

Table 23. Llama-3.2-3B unsafe-row rate at $L = 4096$, $b = 0$, active-unsaturated rows, side-by-side with Llama-3.2-1B (Table 10). Prevalence is similar at the 1.4–3 B Llama scale, consistent with the diagnostic stability we observe on the substitution sweep (Appendix AC).

Suite	$p=2$ (3B)	$p=2$ (1B)	$p=4$ (3B)	$p=8$ (3B)	med Δ/u_*	med A_{act}
induction	19.4	19.6	10.7	4.1	0.728	2
copying	30.5	30.2	18.1	9.0	0.605	3
mqar	20.6	21.4	10.5	4.5	0.760	2

The Llama-3.2-3B prevalence at $p = 2$ (19–30 % unsafe across suites) closely matches Llama-3.2-1B's (19–30 %), and the median active-distractor count is identical. The Llama family is polynomial-safe in the same regime at both scales we audit (1.4 and 3 B parameters); the substitution and head-ablation extensions to 3B are reported in Appendices AC and AD.

AC. Llama family substitution sweep at 1, 3, 8 B

We extend the substitution sweep of Table 3 to two larger Llama models: Llama-3.2-3B (auto-calibrated bias $b_{\text{auto}} = +4.50$) and the open-weight Llama-3-8B base model (Meta AI, 2024) ($b_{\text{auto}} = +4.50$). All three Llama

Table 24. Llama family substitution sweep at NIAH pad = 128 ($n = 200$). Sample-perfect cells (Wilson 95 % lower bound 98.1 %) shown in bold. Both calibrated preserving recipes (relu_p4_bauto, relu_p8_bauto) preserve at all three Llama scales we test.

Variant	Llama-1B	Llama-3.2-3B	Llama-3-8B
softmax	**100.0**	**100.0**	**100.0**
relu_p2_b0	81.0	50.0	50.0
relu_p2_bauto	91.5	87.5	71.5
relu_p4_b0	4.5	6.0	10.0
relu_p4_bauto	**100.0**	**100.0**	**100.0**
relu_p8_b0	0.0	0.0	1.0
relu_p8_bauto	**100.0**	**100.0**	**100.0**
relu_p16_b0	0.5	0.0	0.0
relu_p16_bauto	5.0	25.0	34.0
sigmoid_b0	3.5	19.5	72.0
sigmoid_bauto	0.0	0.0	0.0

models are evaluated on the same English-noun NIAH suite at pad = 128 ($n = 200$, prompts ~ 1500 tokens). Table 24 reports.

Reading the table. The two sample-perfect Llama-1B cells (relu_p4_bauto and relu_p8_bauto) preserve at all three Llama scales, with Wilson 95 % lower bound 98.1 % at $n = 200$ in every cell. This is the strongest cross-scale evidence we have that the architectural partition is a Llama-family property and not specific to the 1.4 B parameter regime. Two minor scale-dependent shifts: sigmoid_b0 improves with scale (3.5 → 19.5 → 72.0 %), suggesting larger Llama models tolerate sigmoid attention's lower concentration without re-calibration; relu_p2_bauto degrades with scale (91.5 → 87.5 → 71.5 %), suggesting low-degree calibrated recipes lose preservation as model scale grows. Neither shift overturns the architectural partition.

AD. Llama-3.2-3B group head ablation

We replicate the group-ablation experiment of Figures 3 and 4 at 3B scale on Llama-3.2-3B, using the same per-head S_p ranking (induction parquet at $L = 4096$, $p = 2$, $b = b_{\text{auto}} = +4.50$). Table 25 reports top-$K$, bot-$K$, and 5-draw random-$K$ deltas at $K \in \{5, 10, 20, 40, 80\}$ on the NIAH suite (pad=128, $n = 200$) and the literary-completion suite (pad=500, $n = 200$).

Table 25. Llama-3.2-3B group ablation: Δ accuracy (pp) at $p = 2$, $b_{\text{auto}} = +4.50$. Both NIAH and copying suites show uniform robustness: every cell stays within ±8 pp of the softmax baseline (100 % NIAH, 79 % copying). Llama-3.2-3B mirrors the Llama-3.2-1B pattern of Table 6: robust at every K and every kind, because Llama has no high-S_p tail to ablate.

	NIAH (pad=128)		copying (pad=500)		
K	top Δ	bot Δ	top Δ	bot Δ	rnd-mean
5	0.0	0.0	0.0	0.0	0.0
10	0.0	0.0	0.0	0.0	0.0
20	0.0	0.0	0.0	0.0	0.0
40	0.0	0.0	0.0	0.0	−1.2
80	0.0	0.0	0.0	0.0	−1.6

The 3B Llama remains uniformly robust to both top-K and bot-K polynomialization at every K tested on both suites, confirming the per-model finding from Section 5.4: Llama is robust to selective polynomial substitution, and the per-head S_p ranking does not behaviorally discriminate within Llama because the model has no high-S_p tail. The pattern is therefore architecture-driven, not scale-driven.

AE. Per-head ablation: retrieval is distributed

The group-ablation result (Figures 3 and 4) is decisive at $K \geq 20$ on Qwen3 and Gemma-4: top-K polynomialization drops behavior −80 to −100 pp while bot-K stays near 0 pp. To localize the source of decisiveness,

we further ablate every single head (one at a time) on these two models, polynomializing only that head and measuring NIAH accuracy. Table 26 summarizes.

Table 26. Per-head NIAH ablation: ablating any single head produces $\Delta = 0$ pp on both Qwen3-1.7B and Gemma-4-E2B. The per-head S_p-vs-behavior correlation (Pearson, Spearman) is undefined because every Δ is identically zero. Polynomializing a single head leaves the retrieval circuit functional. Combined with the $K \geq 20$ decisiveness of Figure 4, this implies retrieval is distributed: no single head carries the circuit, but a 20-head subset does.

Model	heads tested	nonzero Δ	Pearson r	Spearman ρ
Qwen3-1.7B	40 (top-20 + bot-20 by S_p)	0	undef.	undef.
Gemma-4-E2B	40 (top-20 + bot-20 by S_p)	0	undef.	undef.

Interpretation. Per-head ablation together with the $K \geq 20$ group ablation tells a clean mechanistic story: retrieval is distributed across many heads on Qwen3 and Gemma. Polynomializing any single head is near-no-op, but $K \geq 20$ heads are required to break the circuit. This is consistent with our framing of S_p as a collective screen on a cohort of high-S_p heads rather than a per-head failure-mode predictor.

AF. 20-draw random-K rerun on the copying suite

Appendix V reports a 20-draw random-K rerun on the NIAH suite for Qwen3 and Gemma-4. We mirror the rerun on the copying (literary-completion) suite at pad=500 to harden the decisiveness claim of Figure 3. Table 27 reports.

Table 27. 20-draw random-K rerun on the copying suite at pad $= 500$. "decisive$_{20}$" counts random draws (out of 20) whose Δ strictly exceeds top-K's. Qwen3 is decisive in 20/20 at every $K \in \{20, 40, 80\}$. Gemma is decisive in 20/20 at $K \in \{20, 40\}$ and 19/20 at $K = 80$ (one of 20 random draws of 80 heads happens to include the critical core).

Model	K	top Δ	bot Δ	rnd worst	decisive$_{20}$
Qwen3-1.7B	20	−30.0	0.0	−1.5	20/20
	40	−62.5	0.0	−43.0	20/20
	80	−64.0	0.0	−43.0	20/20
Gemma-4-E2B	20	−31.5	0.0	−23.0	20/20
	40	−71.0	−2.5	−26.0	20/20
	80	−72.5	−2.5	−72.5	19/20

The decisiveness claim of Figure 3 is robust at 20-draw resolution: the S_p ranking identifies behaviorally critical heads on Qwen3 and Gemma at $K \leq 40$ on both NIAH and copying suites; $K = 80$ is marginal because random subsets of 80 heads can include the small core of critical heads.

AG. $n = 500$ rerun of headline preserving cells

To test the ceiling-effect concern at $n = 200$ (where any cell at $\geq 199/200$ is statistically indistinguishable from sample-perfect), we rerun the headline preserving Llama-1B and Gemma-4 cells at $n = 500$ with a fresh prompt seed (seed_base=15000, distinct from the $n = 200$ suite seed). Table 28 reports.

Table 28. $n = 500$ rerun of headline preserving cells at NIAH pad $= 128$ with a fresh prompt seed. All five cells are 500/500 sample-perfect (Wilson 95 % lower bound 99.24 %, vs. 98.12 % at $n = 200$). The sample-perfect claim is robust to the $\sim 2.5\times$ larger sample.

Model	Recipe	$n = 500$ accuracy
Llama-3.2-1B	softmax (control)	500/500
Llama-3.2-1B	relu_p4_bauto	500/500
Llama-3.2-1B	relu_p8_bauto	500/500
Gemma-4-E2B	softmax (control)	500/500
Gemma-4-E2B	relu_p8_bauto	500/500

AH. Fixed-bias robustness across models

The calibrated bias b_{auto} is per-model (+3.55 Llama-1B, +4.50 Llama-3.2-3B and Llama-3-8B, −3.36 Qwen3, +1.93 Gemma-4, +119.75 Pythia). One concern is that the architectural partition of Table 3 is an artifact of per-model calibration rather than a model property. To test, we run relu^4 at four uniform biases $b \in \{0, 1, 2, 3\}$ on every audit model on the NIAH suite (and extend to higher b on the Llama family). Table 29 reports.

Table 29. ReLU^4 accuracy at NIAH pad = 128 ($n = 200$) at uniform (non-calibrated) biases. Sample-perfect cells (Wilson 95 % lower bound 98.1 %) shown in bold. Llama family preserves at every $b \geq 1$ (1B), $b \geq 2$ (3B), $b \geq 3$ (8B); Qwen3 and Pythia collapse at every b on relu^4; Gemma-4's preserving regime is relu^8 not relu^4 (cf. Table 3). The architectural partition is a model-family property, not a calibration artifact.

Model	$b=0$	$b=1$	$b=2$	$b=3$	$b=4$	$b=5$	$b=6$
Llama-3.2-1B	4.5	**100**	**100**	**100**	**100**	**100**	**100**
Llama-3.2-3B	6.0	90.5	**100**	**100**	**100**	**100**	—
Llama-3-8B	10.0	82.0	99.5	**100**	**100**	**100**	**100**
Qwen3-1.7B	0.0	0.0	0.0	0.0	—	—	—
Gemma-4-E2B	4.0	2.0	0.0	0.0	—	—	—
Pythia-1.4b	0.0	0.0	0.0	0.0	—	—	—

Reading the table. Llama preservation at relu^4 is robust over a wide range of biases (preserving region $b \in [1, 6]$ for 1B, $[2, 5]$ for 3B, $[3, 8]$ for 8B with Llama-3-8B verified up to $b = 8$, all sample-perfect): the calibrated b_{auto} values fall comfortably inside these wide preserving regions. Qwen3, Gemma-4, and Pythia collapse at relu^4 at every fixed bias we test. The architectural partition is therefore a robust property of the score-row distribution per model family, not an artifact of per-model recipe calibration. Note that Gemma's preserving regime is relu^8, not relu^4 (cf. Table 3); the 0 % Gemma cells in this table reflect the wrong-degree choice.

AI. RULER-Dwarkesh: post-cutoff retrieval probe

AI.1. Setup

To stress-test the single-template NIAH setting with a standardized prompt format, we construct a RULER-template niah_single_3-style probe (Hsieh et al., 2024) but replace the haystack with a uniformly post-cutoff source: Dwarkesh Patel's substack feed (dwarkesh.com/feed) restricted to posts after 2026-01-01. This is a custom retrieval probe in RULER's prompt format, not the official RULER benchmark. This date is chosen to be safely post-training-cutoff for every model in our audit (Llama-3.2 cutoff 2023-12, Qwen3 ∼late 2024, Gemma-4 ∼late 2025, Pythia trained on the Pile through 2020-09). We extract sentences from the post HTML, sentence-tokenize, and filter out (i) sentences with quotation marks of any flavor (which would re-introduce contamination via Dwarkesh quoting historical figures), (ii) sentences mentioning common pre-cutoff public figures (a curated list of ∼50 names), (iii) sentences containing any cue word ("magic", "uuids", "the special"). 9,261 sentences across 16 post-2026-01-01 posts pass the filter, providing a ∼200,000-word post-cutoff sentence pool.

The prompt template follows the RULER niah_single_3 format exactly:

> Some special magic uuids are hidden within the following text. Make sure to memorize it. I will quiz you about the uuids afterwards.
> [shuffled haystack sentences with one needle "One of the special magic uuids for K is: V." inserted at a random sentence boundary]
> What are all the special magic uuids for K mentioned in the provided text? The special magic uuids for K mentioned in the provided text are

The needle value V is a fresh random RFC-4122-style UUID per prompt; the key K is a fresh single-token English noun. Total prompt length is ∼4 K tokens. Correctness is exact-match of the planted UUID as a substring of the model's first 56 greedy-decoded tokens. We use $n = 50$ prompts per cell. Wilson 95 % half-widths reach ≈13.4 pp at $n = 50$ (worst-case at 50 %); so we interpret only large directional differences and do not draw fine-grained quantitative conclusions from this probe.

Compliance note. We do not redistribute the haystack content; the paper's supplementary materials include only the scraper script (with explicit cutoff date and contamination filter) so future readers can regenerate the prompt set deterministically from the live RSS feed. Sentence-level excerpts are used under research fair-use precedent.

AI.2. Substitution sweep

Table 30 reports the substitution sweep on three audit models with high softmax baselines. We separately checked Pythia on a 100-prompt softmax-only run before launching the substitution sweep and obtained a 0/100 baseline (Pythia cannot retrieve 4 K-context UUID needles even under softmax); we therefore omit Pythia from the substitution sweep, since polynomial deltas would inherit a model-level retrieval failure rather than expose a substitution effect. The auto-bias for each model is unchanged from the main paper.

Table 30. RULER-template substitution sweep with custom post-cutoff Dwarkesh haystack ($n=50$, $\sim$4 K tokens; Wilson half-widths up to ≈ 13 pp). Direction on Llama is consistent with the relative-margin bound: as p increases ($4 \to 8$), behavioral accuracy moves from 4 $\to$ 54 %, alongside the matched diagnostic's predicted-unsafe drop (65.2 $\to$ 28.6 %, Table 31). Gemma's relu_p8_bauto remains high-performing (88 % vs. 100 % softmax); Qwen3 admits no recovering recipe.

Variant	Llama-3.2-1B	Qwen3-1.7B	Gemma-4-E2B
softmax	76.0	**100.0**	**100.0**
relu_p2_b0	0.0	0.0	0.0
relu_p4_bauto	4.0	0.0	0.0
relu_p8_bauto	54.0	0.0	**88.0**
sigmoid_b0	0.0	0.0	0.0
relu_p16_b0	0.0	—	—
relu_p16_bauto	0.0	—	—

Reading the table. Three findings: (a) The softmax baseline drops from 200/200 (English-noun NIAH at $\sim$1.5 K) to 76 % on Llama-1B and 100 % on Qwen3 and Gemma — a meaningful task-difficulty increment that exposes failure modes the easier suite hides. (b) Llama's relu_p4_bauto drops to 4 % here despite preserving sample-perfect on the easier suite; relu_p8_bauto partially recovers to 54 % but does not match the 76 % softmax baseline. The 4 % $\to$ 54 % gradient as p increases is directionally consistent with Corollary 4.3: the worst-case degree requirement scales as $p^\star = \log A / \log(1/(1-\rho))$, so harder retrieval (larger A, smaller ρ) requires larger p, and we observe a recovery in that direction; we do not claim closed-loop validation given $n=50$ and incomplete recovery. (c) Llama at relu^{16} (both $b=0$ and b_{auto}, on 1B/3B/8B) under bf16 attention goes to 0 % uniformly. The diagnostic at $p = 16$ predicts lower unsafe-row fraction than $p = 8$ (Table 31), but the bf16 behavioral substitution fails. Candidate causes include numerical precision of r^{16} in bfloat16 attention, over-sharpening of the routing distribution, or saturation effects downstream of the substituted row. We report the bf16 outcome and do not interpret the $p=16$ cells as evidence against the row-level degree trend.

AI.3. Matched diagnostic on RULER-Dwarkesh prompts

We compute per-row S_p on the same RULER-Dwarkesh prompts using the diagnostic of Section 3. Table 31 reports the all-head unsafe-row fraction at the model-specific b_{auto}.

AI.4. Multi-key RULER-Dwarkesh

We extend the probe to a multi-key variant in the RULER niah_multikey_3 format: each prompt plants 4 distinct (key, UUID) pairs and asks for one of them, requiring multi-key disambiguation. Table 32 reports.

Synthesis. The four-part stress check (substitution sweep, matched diagnostic, multi-key, $p = 16$ cells) is directionally consistent with the diagnostic's relative-margin scaling but is not a closed-loop validation. Where the matched diagnostic predicts many unsafe rows (Llama at $p = 4$, 65 %; Pythia at all p, ≥ 99 %), behavior is poor on this harder retrieval probe; where the diagnostic predicts fewer unsafe rows (Llama at $p = 8$, 29 %), behavior partially recovers but does not match softmax. The Gemma $p = 8$ outlier (89.6 % all-head unsafe but 88 % behavioral) is consistent with our existing finding (Section 5.5 and Appendices S and AE) that retrieval is

Table 31. Matched diagnostic on RULER-Dwarkesh prompts: predicted unsafe-row % at b_{auto} for each model, restricted to active-unsaturated rows. "n_{rows}" is the rows-per-power slice. Direction tracks behavior: on Llama-1B, $p=4$ has high unsafe (65 %) and behavior collapses to 4 %; $p=8$ drops to 29 % and behavior partially recovers to 54 % (Table 30). On Pythia, every p is ≥ 99 % unsafe and behavior collapses. Gemma at $p=8$ shows the all-head/critical-head asymmetry: 89.6 % all-head unsafe but 88 % behavioral preservation, consistent with our existing finding that retrieval flows through a small fraction of layer/head/prompt triples (Appendices S and AE).

Model	n_{rows}	$p=2$	$p=4$	$p=8$	$p=16$
Llama-3.2-1B (b=+3.55)	195 k	89.0	65.2	28.6	10.1
Qwen3-1.7B (b=−3.36)	179 k	89.1	77.7	50.2	21.6
Gemma-4-E2B (b=+1.93)	112 k	100.0	98.5	89.6	51.2
Pythia-1.4b (b=+119.75)	89 k	99.9	99.8	99.6	99.2

Table 32. Multi-key RULER-template probe with custom post-cutoff Dwarkesh haystack (4 keys, $n=50$, $\sim$4 K tokens; Wilson half-widths up to ≈ 13 pp). The single-key family pattern is qualitatively retained: Gemma's relu_p8_bauto remains high-performing at 80 % (vs. 96 % softmax), Llama's at 56 % (vs. 64 % softmax) — recovering relative to softmax but not sample-perfect. At $p=16$, Gemma's relu_p16_bauto retains some signal (16 %); Llama collapses to 0 % at all scales tested.

Variant	Llama-3.2-1B	Gemma-4-E2B
softmax	64.0	**96.0**
relu_p4_bauto	24.0	0.0
relu_p8_bauto	**56.0**	**80.0**
relu_p16_bauto	0.0	16.0

concentrated in a small set of critical heads — the all-head average is too coarse to predict Gemma's behavior, and the critical-head analysis would be required. We treat all RULER-template results as exploratory, given $n=50$ and the unresolved $p=16$ implementation question.