The Level Trap: Row Geometry and Head Compatibility in Attention Substitution

Sanny Kim [1]

Abstract

We study zero-shot row-normalizer substitution in frozen transformers and show that preserving score rank or softmax-era retrieval-head salience is not enough: the substitute must preserve the row-wise geometry and head outputs that downstream layers consume. Our primary row statistic, S_p, is a cheap row-level negative screen: from captured score rows it exactly checks whether a nominated key keeps majority mass under degree-p polynomial attention, but in real transformers it is not a sufficiency guarantee for output correctness. Fixed-bias polynomial attention exposes a level trap: additive row shifts, which softmax ignores, can change normalized polynomial mass; in a fixed 14-checkpoint NIAH audit, canonical zero-bias $P_{4,0}$ falls below 30% exact match on 12 of 14 core NIAH cells. Head compatibility is also operator-specific: on Mistral-7B-Instruct, top-ranked retrieval or answer-logit heads provide little rescue, while one layer-0 protected set rescues the narrow diagnostic but transfers negatively to NoLiMa and MMLU-STEM. We introduce ROWAN-(p, τ), a diagnostic row-centered, finite-support polynomial family, and use frozen ROWAN-$(4, 5)$ to separate the row-centering repair from the remaining finite-support boundary. Across NIAH transfer/no-harm, PG-19/WikiText, NoLiMa, MMLU-STEM, sparse controls, and temperature-scaled softmax, the resulting audit protocol is to report row-shift behavior, mass concentration, operator-specific protected-head compatibility, omitted-tail evidence, and full-support temperature baselines.

[1]Independent Researcher. Correspondence to: Sanny Kim <sannykim.24@gmail.com>.

Camera-ready version. ICML 2026 Workshop.

1. Introduction

Changing attention is a central route to longer-context and more efficient language models. Long-context systems modify inference-time attention, select KV-cache entries, or train sparse mechanisms end-to-end (Han et al., 2024; Li et al., 2024; Yuan et al., 2025), while efficient-attention alternatives use linear, kernelized, sigmoid, ReLU, and polynomial maps (Katharopoulos et al., 2020; Choromanski et al., 2021; Ramapuram et al., 2025; Wortsman et al., 2023; Saratchandran et al., 2026). We study the sharper zero-shot case: the pretrained weights are frozen and only the row-wise attention normalizer is swapped. Replacing softmax attention after training looks like a local operator swap, but pretrained transformers use softmax as a row-wise normalization that produces downstream attention outputs. We ask what a substitute must preserve when the rest of the network was trained to consume softmax attention weights and their value-weighted sums. Normalized ReLU-polynomial attention is a useful stress test: it is monotone and easy to analyze, yet its fixed-bias form reveals a row-level level-trap effect that retrieval-head rankings (Wu et al., 2024; Michalak and Abreu, 2025; Zhang et al., 2025) do not predict.

The thesis is simple: a zero-shot substitute can preserve score order and still break the row-wise attention distribution and head outputs that the pretrained network expects.

One row already shows the failure mode. For a target key and four distractors, take $z = (2, 0, 0, 0, 0)$ and substitute $P_{4,0}$. Softmax assigns the target mass $e^2/(e^2 + 4) \simeq 0.65$, and adding 10 to every score leaves this mass exactly unchanged. The fixed-bias polynomial map instead moves from target mass 1.00 on z to $12^4/(12^4 + 4 \cdot 10^4) \simeq 0.34$ after adding 10 to every coordinate. The score order and raw gap are preserved; only the absolute row level changed. This is the level trap.

This example is deliberately row-level. Here rank preservation means preserving the ordering of atten-

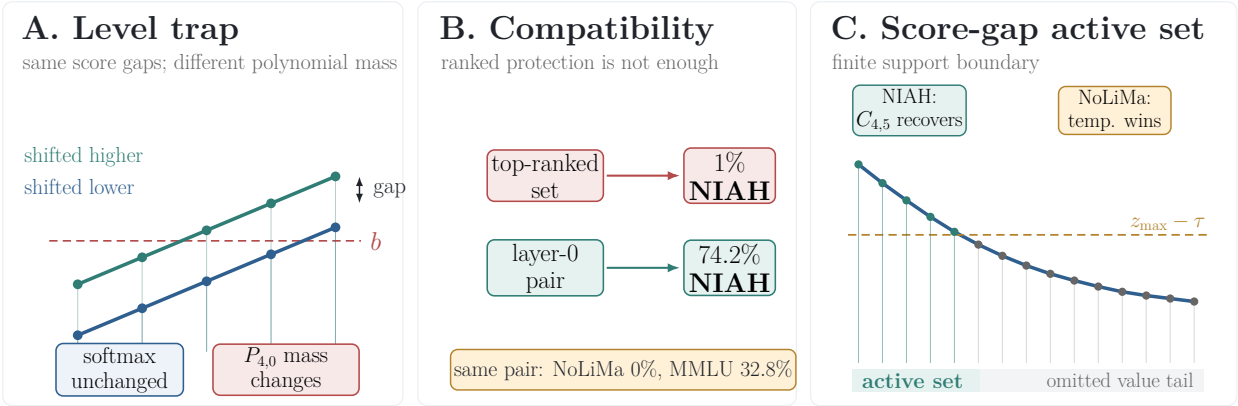

Figure 1. Mechanism overview, read left to right. A: fixed-bias polynomial attention is level-sensitive even when softmax is unchanged by an additive row shift. B: retrieval-head and answer-logit rankings are salience measures, whereas substitution requires an operator-specific protected head set; the Mistral pair rescues the narrow diagnostic but transfers negatively to NoLiMa/MMLU. C: ROWAN removes the level dependence with a score-gap active set, and its finite support predicts the temperature-scaled softmax boundary.

tion scores within a row. Even when a strictly increasing pointwise map preserves the top key, many weaker distractors can collectively receive enough attention mass to change the head output. We therefore separate rank preservation from concentration: a zero-shot substitute must keep enough mass on task-relevant support, beyond preserving score ordering.

The paper is organized as a sequence of audit checks (Figure 1). Each section turns a plausible substitution heuristic into a measurable design constraint: fixed-bias polynomial attention depends on absolute row offset, even when polynomial maps can retrieve after tuning; retrieval-head rankings do not determine substitution compatibility; ROWAN repairs the row-shift violation while mapping the finite-support operating envelope against temperature-scaled softmax and NoLiMa. The audit contract is explicit: a future zero-shot substitute should report row-wise invariances, concentration rather than rank alone, protected head sets under the proposed operator, and full-support temperature-scaled softmax baselines.

Study design. We provide a model-analysis audit for attention substitution in pretrained transformers: a way to test whether a proposed row normalizer preserves the concentration, invariances, and head outputs expected by the frozen network. Polynomial substitutions make the audit sharp because they are monotone, analyzable, and easy to apply without retraining. A failed zero-shot audit is not an irreparability claim; it is evidence that any training, distillation, or conversion procedure must explicitly repair the broken row contract. ROWAN is a frozen diagnostic repair for the level trap, not an efficiency claim or deployment proposal: it removes dependence on the absolute score offset of each row while making explicit the cost of finite attention support. The resulting evidence ladder pairs each constructive result with a control, boundary case, or transfer test that identifies what future substitutes must preserve.

Contributions. (1) Level trap. We derive a target-mass identity for fixed-bias polynomial attention and show that it depends on the post-bias activation scale as well as score gap. In a fixed-protocol 14-checkpoint NIAH audit, canonical zero-bias $P_{4,0}$ falls below 30% on 12 of 14 core checkpoints. (2) Operator-specific head sets. On Mistral-7B-Instruct, retrieval-head and answer-logit rankings identify heads that look important under the original softmax model, yet preserving these ranked head sets provides little rescue; one operator-specific layer-0 protected set rescues the same narrow diagnostic, while NoLiMa and MMLU-STEM identify its transfer limit. (3) Score-gap active-set repair. We define ROWAN-(p, τ), denoted $C_{p,\tau}$ in formulas, prove it is a row-adaptive-bias polynomial map, and decompose it into an active-set approximation to temperature-scaled softmax plus an omitted-tail value error term. Frozen ROWAN-$(4, 5)$ ($C_{4,5}$) is a diagnostic design point: it repairs many NIAH and language-modeling failures, while temperature-scaled softmax and NoLiMa identify where full-support evidence still matters.

Relation to prior work. Recent long-context systems modify attention or KV selection for specific extrapolation or memory bottlenecks, or train sparse mechanisms end-to-end (Han et al., 2024; Li et al., 2024;

Table 1. Evidence ladder. The paper is organized as an audit protocol rather than a single-method win: each positive result is paired with a guardrail that narrows what the result licenses.

Audit question	Main evidence	Guardrail / boundary	Design lesson
Does rank preservation preserve attention mass?	Target-mass identity; $P_{4,0}$ collapses on 12/14 core NIAH cells.	Best-grid fixed-bias recipes can retrieve post-hoc on some cells, so this is preservation failure, not polynomial impossibility.	Report row-shift behavior and concentration, not rank alone.
Do softmax-era retrieval heads identify compatible heads?	Mistral top-ranked answer-logit heads do not rescue; a layer-0 protected pair rescues the narrow diagnostic.	The pair transfers negatively to NoLiMa/MMLU-STEM and is not a universal circuit.	Evaluate protected sets under the substitute operator.
Does row-centering repair the level trap?	Frozen ROWAN-$(4, 5)$ matches softmax on a 12-checkpoint NIAH transfer/no-harm panel and preserves PG-19/WikiText reasonably.	Temperature-scaled softmax remains the full-support positive control and often wins on harder retrieval.	Row-centering is necessary evidence, not a deployment certificate.
When does finite support fail?	Active-set and signed-tail identities decompose ROWAN into active-set softmax plus omitted-tail evidence.	NoLiMa shows paraphrased/distributed evidence can live outside the score-gap window.	Report omitted-tail value evidence and full-support temperature baselines.

Yuan et al., 2025). Efficient-attention and softmax-alternative work studies ReLU, polynomial, sigmoid, sparse, kernelized, recurrent, and stateful mechanisms, often with training, kernels, distillation, or conversion losses (Wortsman et al., 2023; Saratchandran et al., 2026; Kacham et al., 2024; Letourneau et al., 2025; Katharopoulos et al., 2020; Choromanski et al., 2021; Ramapuram et al., 2025; Zhang et al., 2024b; Lan et al., 2025; Bick et al., 2024). We ask a different counterfactual: if the pretrained transformer is held fixed and only its row-wise attention normalizer is replaced, which invariances and head outputs must be preserved? Retrieval-head work identifies heads that attend to or route retrieval-relevant tokens under the original softmax model (Wu et al., 2024; Michalak and Abreu, 2025; Zhang et al., 2025; Xiao et al., 2025; Tang et al., 2024). In particular, Wu-style retrieval-head scores are behavioral or prompt-protocol detectors of which softmax heads attend to retrieval tokens, whereas S_p is a row-map diagnostic: once the score row and a nominated key are fixed, it computes the substitute's majority-mass condition without running a behavioral ablation. Circuit, head-gating, logit-lens, induction-head, copy-suppression, and attention-sink studies show that component importance can be subtle or misleading (Conmy et al., 2023; Chowdhary et al., 2025; Nam et al., 2025; Li et al., 2025; Xiao et al., 2024; Zuhri et al., 2025; Belrose et al., 2023; Janiak et al., 2023; Olsson et al., 2022; McDougall et al., 2023). Our operator-specific question is which heads must keep softmax when all other heads use a specified substitute operator, and that is why retrieval-head rankings and substitution compatibility can diverge.

2. Row-wise score geometry: level trap and ROWAN

For a score row $z \in \mathbb{R}^n$ and a nonnegative pointwise map φ, write

$$A_\varphi(z)_i = \frac{\varphi(z_i)}{\sum_j \varphi(z_j)}$$

whenever the denominator is nonzero. Softmax is $A_{\exp}(z)_i = e^{z_i}/\sum_j e^{z_j}$. The fixed-bias ReLU-polynomial substitute is $A_{\varphi_{p,b}}$ with $\varphi_{p,b}(t) = [t+b]_+^p$, namely

$$P_{p,b}(z)_i = \frac{[z_i + b]_+^p}{\sum_j [z_j + b]_+^p}. \tag{1}$$

Theorem 2.1 (Pointwise row-shift contract; proof in Appendix N). Let $\varphi : \mathbb{R} \to (0, \infty)$ be continuous and define A_φ as above. If $A_\varphi(z + c\mathbf{1}) = A_\varphi(z)$ for every finite row z and scalar c, then $\varphi(t) = Ce^{\beta t}$ for some $C > 0$ and $\beta \in \mathbb{R}$. Conversely, every such exponential map is row-shift invariant.

This classical characterization (Aczél, 1966; Yellott, 1977) is the formal row contract behind the level trap. A fixed pointwise non-exponential substitute must either violate additive row-shift invariance or introduce row-dependent structure. Fixed-bias ReLU-polynomials take the first route; ROWAN takes the second by choosing the bias from the row maximum.

The theorem also clarifies the scope of the paper. We are not asking whether a polynomial or sparse normalizer can be trained from scratch, distilled, or adapted until it works. We are asking what happens when a pretrained network, whose downstream layers have

only seen softmax-normalized head outputs, is asked to consume a different row map without changing any weights. In that setting, row shift is not a cosmetic invariance: it determines which normalized vector downstream layers receive for score rows that are equivalent under softmax. A substitute can preserve the top key, the rank order, and even the raw score gap while still changing the attention output if it changes the way common row level affects concentration.

This is why we treat the frozen experiment as a preconversion audit rather than as an efficient-attention benchmark. If a candidate substitute fails here, the conclusion is not that the operator is unusable; it is that any later training or distillation has a concrete contract to repair. Conversely, a substitute that passes literal retrieval only after per-model tuning has not shown that it preserves the softmax object the frozen model consumes. The row contract gives a cheap first diagnostic before asking more expensive questions about adaptation, kernels, or throughput.

In implementation, rows whose unmasked polynomial denominator is zero are assigned the all-zero attention vector; we do not use a hidden softmax or uniform fallback. This convention makes zero-denominator rows part of the tested substitute rather than silently rescuing them. Let $*$ index a task-relevant key and write $r_i = [z_i + b]_+$. When $r_* > 0$,

$$P_{p,b}(z)_* = \frac{1}{1 + S_p(z,*)}, \qquad S_p(z,*) = \sum_{j \neq *} \left(\frac{r_j}{r_*}\right)^p. \tag{2}$$

We use "mass" for normalized attention weight, equivalently probability mass under the row's attention distribution. Thus $S_p < 1$ is exactly the condition that the nominated key receives more than half the substitute's attention mass. This is a row-level majority-mass statement, not a transformer-output theorem: it would become a downstream recall condition only under an aligned-binary value idealization, which the empirical claims below do not assume. Its practical use is therefore asymmetric. Large unsafe fractions are strong negative evidence against a zero-shot substitute, while small unsafe fractions merely direct attention to the value-direction, omitted-tail, and downstream-compatibility checks that follow. The ratios depend on the post-bias activation scale r_*: an additive row shift $z \mapsto z + c\mathbf{1}$ leaves softmax unchanged but can move $P_{p,b}$ from concentration to collapse.

The contrast with softmax is exact at the same row granularity. If a target-maximal row has at most A distractors and each satisfies $z_j \leq z_* - \Delta$ for $\Delta \geq 0$ and $\beta > 0$, then

$$\mathrm{softmax}(\beta z)_* \geq \frac{1}{1 + Ae^{-\beta\Delta}}. \tag{3}$$

Softmax concentration is controlled by the absolute gap Δ and inverse temperature β. On active polynomial distractors, the fixed-bias envelope below replaces $e^{-\beta\Delta}$ with $(1-\Delta/r_*)^p$, so the same raw gap is less protective when the post-bias activation scale r_* is large. This is the level trap in its simplest algebraic form, and it is only a row-level diagnostic: downstream success also depends on which rows and value directions carry the task.

Proposition 2.2 (Fixed-bias level dependence; proof in Appendix N). Let $D = \{j \neq * : r_j > 0\}$ be the active distractors in a target-score-maximal row. If $|D| \leq A$ and every active distractor satisfies $r_j \leq r_* - \Delta$ for $0 \leq \Delta \leq r_*$, then

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

Equality holds for A distractors tied at $r_* - \Delta$; if at least A active distractors satisfy $r_j \geq r_* - \Delta$, the same expression upper-bounds the target mass. Thus the governing quantity is the relative margin Δ/r_* rather than the raw gap Δ alone.

Corollary 2.3 (Degree envelope). For $A > 1$ and $0 < \Delta/r_* < 1$, a sufficient degree for $P_{p,b}(z)_* \geq 1/2$ is $p \geq \log A / \log(1/(1-\Delta/r_*)) = \Omega(r_* \log A/\Delta)$ as $\Delta/r_* \to 0$. Thus fixed degree and fixed bias cannot be row-shift invariant.

This target-maximal calculation is the minimal counterexample that exposes the row-level geometry; it is not an assumption that every empirical task row has the target as its score maximum. For real transformer rows, the behaviorally important quantities are the attention output and the value direction consumed downstream, which is why the compatibility and omitted-tail value audits below are part of the claim rather than decoration.

We measure operator-specific protected head sets by selective substitution. For head set H and substitute $\mathcal{O}$, let $\mathrm{Acc}_{\mathcal{O}}(H)$ be accuracy when heads in H retain softmax and all other heads use $\mathcal{O}$. When the all-substitution baseline is below softmax, the rescue ratio is

$$R_{\mathcal{O}}(H) = \frac{\mathrm{Acc}_{\mathcal{O}}(H) - \mathrm{Acc}_{\mathcal{O}}(\varnothing)}{\mathrm{Acc}_{\mathrm{sm}} - \mathrm{Acc}_{\mathcal{O}}(\varnothing)}. \tag{5}$$

A compatible protected head set is a high-$R_{\mathcal{O}}$ head set for a particular model–task–operator triple. Below, "protected set" is operator-specific unless stated otherwise; it is not a retrieval-circuit definition.

The constructive row-centered operator is ROWAN: Row-wise Active-set Normalization. In equations we write its formal map as

$$C_{p,\tau}(z)_i = \frac{[\tau + z_i - m]_+^p}{\sum_j [\tau + z_j - m]_+^p}, \quad m = \max_j z_j. \tag{6}$$

The support of $C_{p,\tau}$ is the score-gap active set $A_\tau(z) = \{i : m(z) - z_i < \tau\}$. It is defined by a gap from the row maximum, not by token position or local context. Thus ROWAN is a row-centered finite-support polynomial normalizer rather than a local-attention pattern. Relative to full-support temperature-scaled softmax, keys outside A_τ form the omitted tail.

Proposition 2.4 (Centering as row-adaptive bias). Let $b_\tau(z) = \tau - m(z)$. Then $C_{p,\tau}(z) = P_{p,b_\tau(z)}(z)$ for every score row. If $\delta_i = m(z) - z_i$ and $\delta_* < \tau$, then

$$C_{p,\tau}(z)_* = \frac{1}{1 + S^C_{p,\tau}(z,*)}, \tag{7}$$
$$S^C_{p,\tau}(z,*) = \sum_{j \neq *} \left(\frac{(\tau - \delta_j)_+}{\tau - \delta_*} \right)^p.$$

Both $C_{p,\tau}$ and $S^C_{p,\tau}$ are invariant under $z \mapsto z + c\mathbf{1}$.

The appendix proves the active-set bound. With $\beta = p/\tau$, $A_\tau = \{i : m(z) - z_i < \tau\}$, $s = \mathrm{softmax}(\beta z)$, and $Z_A = \sum_{i \in A_\tau} s_i$,

$$\|C_{p,\tau}(z) - s\|_1 \leq e^{2E_A} - 1 + 2(1 - Z_A), \tag{8}$$

where E_A is a curvature term inside the active set. For actual head outputs, the omitted tail is value-weighted and can be much smaller than the worst-case mass bound. This predicts the central operating envelope: $C_{p,\tau}$ can behave like temperature-scaled softmax when omitted values cancel, while full-support softmax retains an advantage when retrieval requires evidence from the omitted tail.

The signed version makes the boundary operational. For a fixed head and downstream readout direction u, let $g_i = u^\top v_i$, $\gamma_{\mathrm{sm}} = \sum_i s_i g_i$, $\gamma_C = \sum_i C_{p,\tau}(z)_i g_i$, and $T_u = \sum_{i \notin A_\tau} s_i g_i$. The appendix proves

$$\gamma_C = \frac{\gamma_{\mathrm{sm}} - T_u}{Z_A} + R, \qquad |R| \leq (e^{2E_A} - 1) \max_{i \in A_\tau} |g_i|. \tag{9}$$

Thus ROWAN keeps the active-set part of the temperature-softmax answer signal, renormalizes it by $1/Z_A$, and omits the signed tail T_u. If T_u is small, the answer signal is preserved up to curvature error. If positive answer evidence lies outside the score-gap active set, finite support creates a structural obstruction:

$$\gamma_{\mathrm{sm}} > 0 \text{ and } \max_{i \in A_\tau} g_i \leq 0 \quad \Longrightarrow \quad \gamma_C \leq 0 < \gamma_{\mathrm{sm}}. \tag{10}$$

Indeed, any nonnegative normalizer supported on A_τ forms a convex combination of $\{g_i : i \in A_\tau\}$. This is the formal version of the NoLiMa boundary below: row-centering can repair the level trap, but no degree choice recovers answer-positive evidence that the active set omits.

3. The level trap in fixed-bias substitution

Fixed-protocol stress test. The canonical fixed-bias recipe $P_{4,0}$ exposes the level trap across released checkpoints. In the core 14-checkpoint NIAH panel, whose model list, prompts, metric, and $P_{4,0}$ recipe were fixed in scripts before the run, $P_{4,0}$ falls below 30% on 12 of 14 cells. The robust outliers are the Gemma-2 instruction-tuned checkpoints; a post-hoc Falcon probe adds a third robust cell. The effect spans scale: Qwen3-32B and Llama-3.3-70B reach 0% while their softmax baselines are 100%, and Llama-3.1-8B remains in the same below-30% regime. Granite enters the audit with a degraded softmax NIAH baseline (53.4%), so we count it as a collapse cell only by within-checkpoint drop and do not lean on it for the headline. Table 2 reports representative core-audit and 5-panel cells; Table 7 gives the explicit core-audit inventory, and the packaged manifests record model revisions, prompt hashes, and regeneration commands.

This is a no-harm audit, not an attempt to make NIAH the hardest possible retrieval benchmark. The softmax rows are intentionally near-saturated on most checkpoints so that a collapse is interpretable as damage introduced by the substitution. The Gemma outliers and the best-grid column are therefore useful negative controls rather than nuisances: they show that the headline is not "polynomial attention can never retrieve," but "the canonical zero-bias drop-in does not preserve the row contract of many frozen checkpoints." A stronger substitute must survive this easy setting before its behavior on harder retrieval can be meaningful.

The audit also protects against a tempting but wrong interpretation of the numbers. The 12/14 headline is not a claim about model scale, architecture family, or a universal threshold at 30% exact match. It is a fixed-protocol prevalence statement for a named recipe. Reporting the robust Gemma cells, the degraded Granite baseline, and the post-hoc best-grid envelope keeps the claim falsifiable: the row-level theory predicts a failure mode, and the table records where that failure mode is present, repaired, or absent under nearby controls.

This stress test separates polynomial expressivity from zero-shot preservation. Some fixed-bias recipes can

be tuned to retrieve on some checkpoints, and task-matched LoRA repair can recover synthetic retrieval. The core issue is preservation: the pretrained network expected softmax's row-shift invariance, and $P_{4,0}$ violates it. Because "best poly" is selected from a grid, we use it as an expressivity check rather than as the evaluated zero-shot recipe. The Mistral 12.6% cell in Table 2 is the held-out Mistral-7B-Instruct compatibility/audit panel used for the protected-head-set story; the centered-control aggregate includes a separate fixed $n = 500$ Mistral-base panel where all-$P_{4,0}$ reaches 51.2%. We therefore treat the Mistral number as checkpoint- and panel-sensitive, while the cross-checkpoint conclusion is the broader 12/14 level-trap pattern.

4. Head compatibility is operator-specific

The level-trap audit suggests a natural rescue strategy: keep the softmax-era important heads unchanged. We test that strategy by ranking Mistral heads with two diagnostics measured in the original softmax model: Wu-style attention-to-needle evidence (Wu et al., 2024) and an answer-logit contribution score from the head output. Both identify high-scoring heads, but the compact main table foregrounds the direct answer-logit test: preserving the top $K = 80$ heads selected by this score can make NIAH worse than all-polynomial substitution, while preserving the bottom $K = 80$ nearly restores softmax behavior. Here "bottom" means low answer-direction score under the original softmax model, not semantic irrelevance. The rescue is the compatibility dissociation: heads that look unimportant to the softmax model can still be necessary for keeping the substituted network coherent. The appendix shows the same structural-versus-functional mismatch for Wu-style needle-attention rankings. A greedy compatibility search over candidate protected head sets finds one layer-0 pair, $(0, 17), (0, 28)$, that rescues the same narrow diagnostic far above same-size random controls (Table 3).

Table 2. Representative NIAH audit and 5-panel cells. $P_{4,0}$ is the frozen level-trap stress recipe; "best poly†" is the best fixed-bias recipe in the reported grid for that checkpoint and shows that polynomial attention can retrieve after tuning. OLMo-2 is a 5-panel row-centered cell rather than part of the core 14-checkpoint headline. †Post-hoc grid envelope, not an evaluated zero-shot recipe.

Checkpoint	softmax	$P_{4,0}$	best poly†
Mistral-7B-Inst	91.6	12.6	95.0
Qwen3-1.7B	100.0	0.2	49.5
Qwen3-32B	100.0	0.0	80.5
Llama-3.3-70B-Inst	100.0	0.0	99.0
OLMo-2-0425-1B	100.0	0.8	40.0
Granite-3.3-8B-Inst	53.4	22.8	29.6
Gemma-2-2B-it	100.0	96.4	100.0
Gemma-2-9B-it	100.0	97.8	100.0

This is the point where the audit moves beyond row mass. The relevant object is not simply whether a head attends to the needle under softmax, or whether its output has a large direct answer-logit component before substitution. It is whether downstream layers can still interpret the head output after every other head has changed operator. Selective substitution tests this compatibility directly: preserve a proposed set under softmax, substitute the rest, and ask how much of the lost behavior returns. That is why a low-ranked head set can be more protective than a high-ranked one; the score is not measuring semantic importance in isolation, but compatibility with a perturbed computation.

The result also gives a practical warning for head-selection heuristics. A retrieval-head detector can be correct about what a head does in the original model and still be the wrong tool for deciding which heads to protect under a new operator. Compatibility is counterfactual: it depends on the substitute used in the other heads, the residual write produced by the protected head, and the tolerance of downstream layers to that write. This is why the paper reports both the narrow rescue and the negative transfer cells rather than presenting the Mistral pair as a general circuit discovery.

Table 3. Mistral NIAH controls. "sm" is all-softmax; "all-P" is all-polynomial $P_{4,0}$; "prot." preserves the listed head set under softmax while substituting all other heads; R is Equation (5). Answer-logit rows use the head-output answer-direction magnitude $|c_*|$ ranking. $n = 200$ for $K = 80$ rows; $n = 500$ for the pair and random null. The random $K = 2$ row is the maximum over 20 all-head random draws; layer-0-conditioned nulls are in Appendix AX. Baselines and R are computed within each block because the two blocks use different prompt counts/splits.

Intervention	sm	all-P	prot.	R
Top answer-logit $K = 80$	95	14	1	−0.16
Bottom answer-logit $K = 80$	95	14	93	0.98
Pair $(0, 17), (0, 28)$	91.6	12.6	74.2	0.78
Random $K = 2$ (max/20 draws)	91.6	12.6	17.0	0.06

The pair is useful because it separates head salience from substitution compatibility. It is one validation-hardened Mistral protected set for a particular operator and diagnostic: the appendix excludes singletons under a family-wise screen and gives a matched held-out certificate for the pair, but also finds other layer-0 sufficient pairs at smaller n. Its transfer profile is diagnostic: the same pair gives 0.0% on NoLiMa-main

and 32.8% on MMLU-STEM, below the corresponding all-$P_{4,0}$ baselines (Table 60). The dissociation itself—structural detectors miss the substitution-critical protected set—reproduces across 7 of 9 tested families, but the compact Mistral form is model-family-specific: Llama-3.1-8B needs $K = 80$, OLMo-2 needs $K = 40$, and those protected sets span multiple layers (Appendix AX). What transfers is the abstraction, not the two-head mechanism: semantic retrieval-head identity is insufficient for predicting which set of heads keeps a substituted network coherent. The reusable check is therefore to evaluate protected sets under the substitute operator rather than importing retrieval-head rankings from the original softmax model.

5. ROWAN: row-centered active-set normalization

Selection protocol for (p, τ). To repair the row-level violation directly, we selected $C_{4,5}$ in two stages. A Mistral-only calibration grid over p and τ chose the smallest score-gap threshold that saturated the Mistral NIAH calibration cell at $p = 4$: $\tau = 1$ collapsed, $\tau = 3$ recovered most but not all of softmax, and $\tau \in \{5, 10\}$ matched softmax. We then froze $(p, \tau) = (4, 5)$ before the cross-model centered panel, PG-19 (Rae et al., 2020)/WikiText-103 (Merity et al., 2017) preservation tests, NoLiMa (Modarressi et al., 2025) transfer cells, MMLU-STEM (Hendrycks et al., 2021) rescue tests, and sparse-normalizer controls (Martins and Astudillo, 2016; Peters et al., 2019; Correia et al., 2019). This is a single-cell calibration, not a claim that $C_{4,5}$ is globally optimal; all subsequent rows treat ROWAN-$(4, 5)$ as a frozen diagnostic design point and evaluate its transfer against stronger full-support and bias controls.

Retrieval and language modeling. The constructive payoff is narrow but real. On a separate 12-checkpoint NIAH transfer/no-harm panel, frozen ROWAN-$(4, 5)$ stays within 1 pp of softmax on every measured row, including many rows where $P_{4,0}$ collapses (Table 8). It also keeps language modeling close to softmax: PG-19 ratios are 1.035–1.132× and WikiText-103 ratios are 0.983–1.117× across the five-family panel. These are preservation results, not a dominance claim. Here $\mathrm{softmax}_\beta$ denotes $\mathrm{softmax}(\beta z)$, so larger β means sharper attention. On NoLiMa-main ($n = 200$; Wilson intervals in Appendix BC), ROWAN-$(4, 5)$ is far above $P_{4,0}$ but below full-support temperature-scaled softmax: Mistral/Llama score 55.0/22.0% under ROWAN-$(4, 5)$, compared with 0.5/0.0% under $P_{4,0}$, 74.5/83.0% under softmax, 65.5/59.0% under $\mathrm{softmax}_{0.8}$, and 71.5/60.0% under $\mathrm{softmax}_{1.2}$, the best swept Mistral/Llama inverse temperature reported in the appendix. MMLU-STEM similarly treats ROWAN-$(4, 5)$ as a preservation sanity check: across five families it recovers $R = 0.77$–0.88 of the $P_{4,0}$ damage, while $\mathrm{softmax}_{0.8}$ is more variable ($R = 0.55$–0.99). The theory predicts this split: finite support helps when omitted values cancel or are redundant, but full-support softmax tails can matter when evidence is distributed or paraphrased.

Strong baselines. Temperature-scaled softmax is the strongest baseline because it keeps softmax's row-shift invariance and full support while changing concentration. The appendix sweep shows that $\beta = 1.2$ is often best on PG-19, while flatter $\beta = 0.6$ can collapse retrieval on OLMo-2 and Qwen3. This is the intended boundary condition of the audit: centering and a score-gap active set explain why a polynomial substitute can avoid the fixed-bias level trap, while temperature scaling exposes the cost of finite support. A calibrated global-bias control separates these mechanisms: $P_{4,b_{\mathrm{auto}}}$ rescues some Mistral cells but has catastrophic PG-19 ratios on several families, collapses NIAH on OLMo-2/Qwen3, and collapses NoLiMa on Mistral/Llama. Row-centering is doing something a single global offset cannot.

Finite-support boundary. The finite-support account makes three useful predictions. If omitted values are redundant or have low output norm, row-level L^1 mismatch can be large while next-token perplexity remains close to softmax; PG-19 and WikiText satisfy this condition. If the task is literal and the answer-bearing rows remain inside the active set with benign omitted-tail value error, ROWAN-$(4, 5)$ can match temperature-scaled softmax; NIAH mostly satisfies this condition. If retrieval requires evidence distributed across low-probability or paraphrased support, full-support temperature scaling should retain information that ROWAN-$(4, 5)$ omits; NoLiMa satisfies this condition. Thus ROWAN's role is diagnostic: it removes the fixed-bias level dependence while making the finite-support cost explicit.

6. Discussion: audit checks and operating envelope

The experiments reduce five plausible substitution heuristics to reusable audit checks. The fixed-bias audit shows when rank preservation fails to preserve concentration. The head-policy experiment separates softmax-era head salience from substitute-era head compatibility. The calibrated-bias comparison tests whether one global offset can replace row-centering. The temperature sweep tests the strongest full-support

concentration baseline. Finally, NoLiMa and the active-set diagnostics test whether exact-match retrieval recovery transfers to harder retrieval. The resulting recipe is compact: zero-shot substitutes must preserve softmax's row-wise invariances and must account for value contributions from the omitted tail.

Protocol status. We separate four evidential statuses. The 14-checkpoint $P_{4,0}$ audit is fixed-protocol evidence for a canonical zero-bias stress recipe. The "best poly" and calibrated-bias rows are post-hoc or control envelopes, used to distinguish polynomial expressivity from zero-shot preservation. The choice $C_{4,5}$ is a single-cell Mistral calibration that is then frozen before the 12-checkpoint transfer/no-harm panel, language-modeling preservation, NoLiMa, MMLU-STEM, and sparse-normalizer controls. The longer appendix analyses–mechanism probes, repair attempts, and additional retrieval variants–are preservation stress tests or hypothesis-generating diagnostics rather than extra main claims.

How to use the result. The practical audit follows directly from the theory. First, report the row contract: row-shift behavior, concentration curves, active-set sizes, and the zero-denominator convention. Second, mark evidential status explicitly: fixed protocol, single-cell calibration, frozen transfer, post-hoc envelope, or exploratory. Third, test operator-specific head compatibility with selective substitution and same-size or structure-matched nulls. Fourth, compare against fixed and swept temperature-scaled softmax, and measure omitted-tail value or signed-answer evidence when finite support is used. Finally, pair literal retrieval with language-modeling preservation, harder retrieval, downstream sanity checks, and manifest-level traceability.

For fixed-bias polynomial candidates, compute S_p on representative prompt rows before behavioral evaluation. It is cheap because it is computed from captured attention score rows rather than from a labeled retrieval benchmark or a many-sample head-ablation protocol. A large unsafe fraction, especially in task-relevant rows or heads, is an early warning, while a small unsafe fraction directs attention to value-direction and downstream compatibility checks. We use S_p as a cross-cell no-go screen, not as a within-cell prompt-collapse predictor: every audited cell with all-head measured-safe fraction $\leq 3\%$ collapsed, but within a single model–recipe cell S_p does not outpredict a one-line pre-softmax logit-margin baseline (Appendix AR). A protected head set that rescues NIAH but fails NoLiMa or MMLU-STEM identifies a compatibility boundary for the proposed operator.

Scope and falsifiability. The zero-shot setting is intentionally strict. It is not a claim that a failed substitute is impossible to train, distill, or adapt; it is a pre-conversion audit of what the frozen network already expects from softmax. A trained replacement would falsify the pessimistic version of our results by preserving language modeling, harder retrieval, and downstream behavior without protected softmax heads, while also matching full-support temperature baselines and passing the row-contract checks above. Conversely, a substitute that succeeds only on literal NIAH, only after post-hoc per-model tuning, or only by preserving a narrow protected set has not yet shown that it preserves the consumed attention outputs of the frozen model. This distinction also explains the role of the negative controls: the Gemma outliers, best-grid envelopes, calibrated-bias repairs, and temperature-softmax wins are not exceptions to be hidden; they mark which part of the row contract a substitute has or has not repaired.

7. Conclusion

Post-training attention substitution is governed less by named retrieval heads than by row-wise geometry and operator-conditioned compatibility. Fixed-bias polynomial attention falls into a level trap because it violates softmax's row-shift invariance; ROWAN removes that violation with a row-adaptive bias and score-gap active set, but temperature-scaled softmax and NoLiMa expose the remaining cost of finite support. The contribution is therefore an audit recipe and design target rather than a claim that one substitute dominates: future work should report substitution compatibility, row-shift behavior, omitted-tail evidence, and full-support temperature baselines. A stronger substitute would combine ROWAN's row-shift invariance with controlled value tails and the full-support evidence that paraphrased retrieval can require, while preserving language modeling under the frozen downstream network. This distinction is the main lesson of the audit: a substitute can pass a rank or approximation test while still failing the frozen model's consumed head outputs. The supplementary artifact gives the claim-to-artifact map, model/prompt manifests, and reproduce_main.sh --data-only hook that keep each headline number traceable to a named artifact, seed, and scoring script.

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.

Simran Arora, Sabri Eyuboglu, Michael Zhang, Aman Timalsina, Silas Alberti, Dylan Zinsley, James Zou, Atri Rudra, and Christopher Ré. Simple linear attention language models balance the recall-throughput tradeoff. In Proceedings of ICML, 2024. arXiv:2402.18668.

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.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. arXiv:2303.08112, 2023.

Aviv Bick, Kevin Y. Li, Eric P. Xing, J. Zico Kolter, and Albert Gu. Transformers to SSMs: Distilling quadratic knowledge to subquadratic models. arXiv:2408.10189, 2024.

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. arXiv:2304.01373.

Nicholas H. Bingham, Charles M. Goldie, and József L. Teugels. Regular Variation. Cambridge University Press, 1987.

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

Aleksandar Botev, Soham De, Samuel L Smith, Anushan Fernando, George-Cristian Muraru, Ruba Haroun, Leonard Berrada, Razvan Pascanu, Pier Giuseppe Sessa, Robert Dadashi, et al. RecurrentGemma: Moving past Transformers for efficient open language models. arXiv:2404.07839, 2024.

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.

Pratim Chowdhary, Peter Chin, and Deepernab Chakrabarty. K-MSHC: Unmasking minimally sufficient head circuits in large language models with experiments on syntactic classification tasks. arXiv:2505.12268, 2025.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In Proceedings of NeurIPS, 2023. arXiv:2304.14997.

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

Laurens de Haan and Ana Ferreira. Extreme Value Theory: An Introduction. Springer, 2006.

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

Aaron Grattafiori et al. The Llama 3 herd of models. arXiv:2407.21783, 2024. https://huggingface.co/meta-llama/Llama-3.2-1B.

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

Chi Han, Qifan Wang, Hao Peng, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang. LM-Infinite: Zero-shot extreme length generalization for large language models. In Proceedings of NAACL, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In Proceedings of ICLR, 2021. arXiv:2009.03300.

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.

Jett Janiak, Can Rager, Tom Dao, and James Lau. An adversarial example for direct logit attribution: memory management in GELU-4L. arXiv:2310.07325, 2023.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. arXiv:2310.06825, 2023. https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3.

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.

Disen Lan, Weigao Sun, Jiaxi Hu, Jusen Du, and Yu Cheng. Liger: Linearizing large language models to gated recurrent structures. In Proceedings of ICML, 2025.

Wonsuk Lee. On the invariants of softmax attention. arXiv:2605.02907, 2026.

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.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. SnapKV: LLM knows what you are looking for before generation. In Advances in Neural Information Processing Systems 37, 2024. arXiv:2404.14469.

Victoria R. Li, Jenny Kaufmann, Martin Wattenberg, David Alvarez-Melis, and Naomi Saphra. Can interpretation predict behavior on unseen data? arXiv:2507.06445, 2025.

Christos Louizos, Max Welling, and Diederik P. Kingma. Learning sparse neural networks through L_0 regularization. In Proceedings of ICLR, 2018.

Peng Lu, Ivan Kobyzev, Mehdi Rezagholizadeh, Boxing Chen, and Philippe Langlais. ReGLA: Refining gated linear attention. In Proceedings of NAACL, 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.

Callum McDougall, Arthur Conmy, Cody Rushing, Thomas McGrath, and Neel Nanda. Copy suppression: Comprehensively understanding an attention head. In Proceedings of BlackboxNLP, 2023. arXiv:2310.04625.

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

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In Proceedings of ICLR, 2017.

Felix Michalak and Steven Abreu. Some attention is all you need for retrieval. arXiv:2510.19861, 2025.

Microsoft. Phi-4-Mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs. arXiv:2503.01743, 2025. https://huggingface.co/microsoft/Phi-4-mini-instruct.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A. Rossi, Seunghyun Yoon, and Hinrich Schütze. NoLiMa: Long-context evaluation beyond literal matching. In Proceedings of ICML, 2025. arXiv:2502.05167.

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

Andrew Nam, Henry Conklin, Yukang Yang, Thomas Griffiths, Jonathan Cohen, and Sarah-Jane Leslie. Causal head gating: A framework for interpreting roles of attention heads in transformers. In Proceedings of NeurIPS, 2025. arXiv:2505.13737.

Chien Van Nguyen et al. Lizard: An efficient linearization framework for large language models. arXiv:2507.09025, 2025.

OLMo Team, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, et al. 2 OLMo 2 furious. arXiv:2501.00656, 2025. https://huggingface.co/allenai/OLMo-2-0425-1B.

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. Qwen2.5 technical report. arXiv:2412.15115, 2024.

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

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. Compressive transformers for long-range sequence modelling. In Proceedings of ICLR, 2020.

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, Jianqiao Zheng, Yiping Ji, Wenbo Zhang, and Simon Lucey. Rethinking attention: Polynomial alternatives to softmax in transformers. arXiv:2410.18613v3, 2026.

Shawn Tan, Songlin Yang, Aaron Courville, Rameswar Panda, and Yikang Shen. Scaling stick-breaking attention: An efficient implementation and in-depth study. In Proceedings of ICLR, 2025. arXiv:2410.17980.

Hanlin Tang, Yang Lin, Jing Lin, Qingsen Han, Shikuan Hong, Yiwu Yao, and Gongyi Wang. RazorAttention: Efficient KV cache compression through retrieval heads. arXiv:2407.15891, 2024.

Pavlo Vasylenko, Hugo Pitorro, André F. T. Martins, and Marcos Treviso. Long-context generalization with sparse attention. In Proceedings of ICLR, 2026. arXiv:2506.16640.

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.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In Proceedings of ICLR, 2023. arXiv:2211.00593.

Junxiong Wang, Daniele Paliotta, Avner May, Alexander M. Rush, and Tri Dao. The Mamba in the Llama: Distilling and accelerating hybrid models. In Proceedings of NeurIPS, 2024.

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

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. arXiv:2404.15574, 2024.

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

Guangxuan Xiao, Jiaming Tang, Jingwei Zuo, Junxian Guo, Shang Yang, Haotian Tang, Yao Fu, and Song Han. DuoAttention: Efficient long-context LLM inference with retrieval and streaming heads. In Proceedings of ICLR, 2025. arXiv:2410.10819.

Kai Yan, Zhan Ling, Kang Liu, Yifan Yang, Ting-Han Fan, Lingfeng Shen, Zhengyin Du, and Jiecao Chen. MIR-Bench: Can your LLM recognize complicated patterns via many-shot in-context reasoning? In NeurIPS Datasets & Benchmarks Track, 2025. arXiv:2502.09933.

Fanqi Yan, Huy Nguyen, Pedram Akbarian, Nhat Ho, and Alessandro Rinaldo. Sigmoid self-attention has lower sample complexity than softmax self-attention: a mixture-of-experts perspective. arXiv:2502.00281, 2025.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated Delta Networks: Improving Mamba2 with delta rule. In Proceedings of ICLR, 2025.

Can Yaras, Alec S. Xu, Pierre Abillama, Changwoo Lee, and Laura Balzano. MonarchAttention: Zero-shot conversion to fast, hardware-aware structured attention. arXiv:2505.18698, 2025.

Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. Differential Transformer. In Proceedings of ICLR, 2025. arXiv:2410.05258.

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.

Jingyang Yuan, Huazuo Gao, Damai Dai, Junyu Luo, Liang Zhao, Zhengyan Zhang, Zhenda Xie, Yuxing Wei, Lean Wang, Zhiping Xiao, Yuqing Wang, Chong Ruan, Ming Zhang, Wenfeng Liang, and Wangding Zeng. Native Sparse Attention: Hardware-aligned and natively trainable sparse attention. In Proceedings of ACL, 2025.

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

Michael Zhang, Simran Arora, Rahul Chalamala, Alan Wu, Benjamin Spector, Aaryan Singhal, Krithik Ramesh, and Christopher Ré. LoLCATs: On low-rank linearizing of large language models. arXiv:2410.10254, 2024.

Wuwei Zhang, Fangcong Yin, Howard Yen, Danqi Chen, and Xi Ye. Query-focused retrieval heads improve long-context reasoning and re-ranking. arXiv:2506.09944, 2025.

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. Limitations

The main protected-set result is a hardened Mistral NIAH case study plus cross-model operator-specificity checks, not a universal mechanistic taxonomy. The two-head Mistral set is specific to the model, task, and operator; it transfers negatively to NoLiMa and MMLU-STEM, and we do not replicate the full protected-set discovery certificate at 30B–70B scale. NIAH is controlled and useful for mechanism, but it has lexical shortcuts and can saturate under several operators. NoLiMa and MMLU-STEM reduce this concern but remain a small downstream slice.

ROWAN-(p, τ) is a diagnostic constructive probe, not a deployment certificate. It computes full score rows in our implementation, so we make no throughput claim. The practical risk is treating a narrow protected head set or ROWAN-$(4, 5)$ swap as a sufficient repair without transfer, language-modeling, and downstream audits; such a model could appear fixed on exact-match retrieval while failing paraphrased retrieval or broader reasoning. It is often close to $\mathrm{softmax}((p/\tau)z)$; in some regimes, especially NoLiMa and PG-19 with $\beta \to 1$, ordinary temperature-scaled softmax is stronger. We did not train models from scratch with ROWAN-(p, τ), learn per-row blends with temperature-scaled softmax, or evaluate a full downstream suite such as LongBench (Bai et al., 2024), BABILong (Kuratov et al., 2024), MIR-Bench (Yan et al., 2025a), or broader reasoning, instruction-following, and safety tasks. Multi-seed language-modeling preservation tops out at 8B-scale models on the five-family panel; the 70B PG-19 cells ($N = 3$ on Llama-3.3-70B for both ROWAN-$(4, 5)$ and $\mathrm{softmax}_{0.8}$) are exploratory rather than primary evidence.

The statistical protocol separates fixed-protocol, follow-up, and exploratory claims, but several appendix analyses remain hypothesis-generating. Random-set nulls with 20 draws detect large effects but do not replace exhaustive search except where explicitly certified. Benchmarks such as PG-19 may appear in training data for modern open checkpoints, so we interpret language-modeling results as within-checkpoint preservation tests rather than contamination-filtered generalization estimates.

B. Appendix roadmap

This appendix is a claim-indexed evidence atlas for the main paper. Appendices C, E and J record the complete table sources, model/prompt manifests, scoring rules, and reproducibility hooks behind every headline cell. Appendices N, BA, BB and BF support the level-trap and active-set claims with proofs, active-set diagnostics, and omitted-tail value measurements. Appendices AX to AZ and BE document the operator-specific protected-head-set intervention: Mistral protected-set search, singleton/pair certificates, QK/OV controls, and amplifier screens. Appendices AP and AQ contain trained-repair results treated as preservation stress tests rather than as an efficient-attention method. Local headings mark auxiliary or exploratory analyses; the main paper is self-contained and does not require appendix-only results for its central claims. The main claims also do not depend on the LoRA repair, RULER-Dwarkesh, amplifier-screen, or path-patching appendices; these are included as stress tests and diagnostics.

C. Full tables for main results

This appendix provides the detailed tables behind the compact main body. The main text reports only the evidence needed for the eight-page argument; full numerical variants remain below and in subsequent appendices.

Table 4. Submission-time claim-to-artifact map. Each main claim has a boundary condition and a named supplement trace, making the paper falsifiable without requiring reviewers to rerun model inference.

Claim	Key evidence	Boundary / strong baseline	Supplement trace
Fixed-bias polynomial attention has a level trap	Target-mass identity and core 14-checkpoint $P_{4,0}$ NIAH collapse	best-grid and calibrated-bias recipes can repair some cells, so the claim is about zero-shot substitution risk	fixed-bias logs
Retrieval-head rankings are not substitution compatibility	Mistral policy table: retrieval and answer-logit heads fail; one operator-specific protected head set rescues	Llama reproduces detector failure but not tiny-set rescue; the Mistral pair transfers negatively to NoLiMa/MMLU	head-policy and transfer logs
Row-centering repairs the diagnosed invariance violation	ROWAN-$(4, 5)$ fixed after Mistral calibration and evaluated on held-out models/tasks	temperature-scaled softmax keeps full support and often wins; ROWAN-$(4, 5)$ is a design point, not a dominance claim	centered audit
Finite support explains success and failure	PG-19/WikiText preservation plus active-set/omitted-tail value diagnostics	NoLiMa exposes paraphrased/full-support tail needs; sparsemax/entmax rule out generic sparsity	active-set logs

Table 5. Polynomial-unsafe row rate $\Pr(S_p \geq 1)$ at the longest context and $b = 0$, restricted to active-unsaturated rows. This is the compact version of the prevalence table.

Model	Suite	$p = 2$	$p = 4$	$p = 8$	A_{med}
Llama-3.2-1B	ind/cpy/mqar	20–30	10–18	5–10	2–3
Qwen3-1.7B	ind/cpy/mqar	94–97	76–80	37–49	540–1884
Gemma-4-E2B	ind/cpy/mqar	71–96	57–80	38–49	36–96
Pythia-1.4B	ind/cpy/mqar	77–99	65–98	51–95	148–1407

Table 6. Δ next-token accuracy (pp) at pad 500 on the literary-copying suite. No single pointwise-normalized recipe recovers softmax on all four audit models; unnormalized Wortsman variants uniformly collapse.

Variant	Llama	Qwen3	Gemma-4	Pythia
relu $p = 2, b = 0$	-52.5	-79.0	-70.5	-70.5
relu $p = 4, b = 0$	-54.5	-73.0	-27.0	-69.0
relu $p = 8, b_{\text{auto}}$	-25.0	-59.0	-5.0	-65.0
relu $p = 16, b = 0$	-69.0	-14.5	-45.0	-77.5
sigmoid $b = 0$	-57.5	-79.0	-72.5	-65.0
Wortsman variants	-64.0– -74.5	-79.0	-71.0– -74.0	-77.5

Table 7. Core 14-checkpoint NIAH audit inventory under canonical $p = 4, b = 0$ polynomial substitution. The headline is 12/14 canonical collapse: every core row except the two Gemma-2 instruction-tuned checkpoints falls below 30% exact match. Ranges indicate multiple checkpoints in the listed block; cells use $n = 200$–500 depending on checkpoint, with exact prompt counts, model revisions, and prompt hashes recorded in the packaged artifacts and model_manifest.tsv.

Core audit block	Checkpoints	softmax	$P_{4,0}$	reading
Mistral	7B-Instruct	91.6	12.6	collapse; grid can repair this cell
Llama-3.2 base	1B, 3B	100	3.4–5.0	collapse; partial via b_{auto}
Llama-3.2-Instruct	1B, 3B	100	0.0–1.2	collapse
Llama-3.1/3.3-Instruct	8B, 70B	100	0.0–12.0	collapse; 70B best-grid repair
Qwen2.5-Instruct	3B, 7B	98.4–100	0.0–0.2	collapse/partial
Qwen3	1.7B, 32B	100	0.0–0.2	collapse; best grid remains below softmax
Granite	3.3-8B-Instruct	53.4	22.8	collapse relative to a lower baseline
Gemma-2-it	2B-it, 9B-it	100	96.4–97.8	robust outliers

Table 8. ROWAN-$(4, 5)$ cross-checkpoint NIAH panel. This table gives the per-checkpoint numbers behind the main-text NIAH transfer/no-harm claim. The recipe is fixed at $p = 4, \tau = 5$ with no per-model calibration. Cells use $n = 500$ except Gemma-2-9B, Gemma-4-E2B, and Qwen2.5-7B, which use $n = 200$; a post-hoc Falcon row is packaged but excluded from the 12-cell panel. This centered panel is not identical to the twelve collapse cells from the 14-checkpoint $P_{4,0}$ audit (Table 7); it is a transfer/no-harm panel that includes non-collapse rows such as Gemma-2-9B-it. The claim is within ≤ 1 pp of softmax on every measured row, not that all twelve rows collapsed.

Checkpoint	softmax	$P_{4,0}$	ROWAN-$(4, 5)$
Mistral-7B-v0.3 base	91.6	51.2	91.6
OLMo-2-0425-1B	100.0	0.8	99.2
Llama-3.1-8B-Instruct	100.0	12.0	100.0
Llama-3.2-1B	100.0	3.2	100.0
Llama-3.2-3B-Instruct	100.0	0.0	100.0
Qwen3-1.7B	100.0	0.2	100.0
Qwen2.5-7B-Instruct	98.5	0.0	97.5
Granite-3.3-8B-Instruct	53.4	22.8	53.4
Phi-4-mini-Instruct	100.0	0.2	100.0
Pythia-1.4B	100.0	0.0	99.4
Gemma-4-E2B	100.0	5.0	100.0
Gemma-2-9B-it	100.0	97.5	100.0

Table 9. Δ NIAH accuracy (pp) at pad 128 versus 200/200 softmax baselines on the four-model audit.

Variant	Llama	Qwen3	Gemma-4	Pythia
relu $p = 2, b = 0$	-19.0	-100	-100	-100
relu $p = 2, b_{\text{auto}}$	-8.5	-100	-100	-100
relu $p = 4, b = 0$	-95.5	-100	-96.0	-100
relu $p = 4, b_{\text{auto}}$	0.0	-99.5	-100	-100
relu $p = 8, b = 0$	-100	-100	-1.0	-100
relu $p = 8, b_{\text{auto}}$	0.0	-100	0.0	-100
relu $p = 16, b = 0$	-99.5	-50.5	-26.0	-100
sigmoid / Wortsman	-96.5– -100	-100	-100	-100

Table 10. Matched NIAH argmax-concentration contrasts. Low safe fraction is a useful negative screen; high all-head safe fraction is not sufficient.

Model	Recipe	Δ pp	all-head safe	top-K safe
Qwen3	relu $p = 2, b = 0$	-100	2.9	0.0
Gemma-4	relu $p = 2, b = 0$	-100	2.0	0.0
Pythia	relu $p = 4, b_{\text{auto}}$	-100	0.1	0.0
Qwen3	relu $p = 8, b_{\text{auto}}$	-100	66.3	21.9
Llama	relu $p = 2, b = 0$	-19.0	72.3	30.4
Llama	relu $p = 4, b_{\text{auto}}$	0.0	55.7	21.2
Llama	relu $p = 8, b_{\text{auto}}$	0.0	86.4	51.6
Gemma-4	relu $p = 8, b_{\text{auto}}$	0.0	18.3	6.1

D. Extended related work

Softmax characterizations and length limits. Softmax has a long lineage in choice theory and statistics and recent invariant analyses (Lee, 2026): 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). We restate the classical shift-invariance characterization in Appendix N: the only pointwise normalized map invariant to additive row shifts is exponential; fixed-bias alternatives therefore introduce level dependence. 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 2.3, 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. (2026) 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. (2025b) 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 mass allocation in ways that fall outside the pointwise impossibility result of Corollary 2.3; the row-ratio diagnostic itself extends to sigmoid (Appendix Y), 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., 2024a) train ϕ such that $\phi(Q)\cdot\phi(K)\approx\exp(QK)$. Other recall-throughput and subquadratic architectures, including Zoology (Arora et al., 2023), RecurrentGemma (Botev et al., 2024), Based (Arora et al., 2024), and Gated Delta Networks (Yang et al., 2025), are orthogonal attempts to reduce the same long-context bottleneck. These are kernel approximations or memory architectures 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 use would be to apply the same row-mass screen to the post-feature-map score row $\phi(Q)\cdot\phi(K)$; whether passing that screen predicts behavior would still require training and validation infrastructure beyond the present experiments.

Linear and kernelized attention. Katharopoulos et al. (2020); Choromanski et al. (2021) are the canonical linear/kernelized attention references; Arora et al. (2023; 2024); Zhang et al. (2024a;b); Nguyen et al. (2025); Yaras et al. (2025) develop the neighboring design space of recall-oriented subquadratic layers, learned feature maps, and pretrained conversion. They make different design choices (kernel feature maps, causal-mask compatibility, learned adaptation) and are not zero-shot pointwise-on-logit drop-ins.

Pretrained-LLM conversion and repair. LoLCATs (Zhang et al., 2024b), Liger (Lan et al., 2025), ReGLA (Lu et al., 2025), MOHAWK (Bick et al., 2024), Lizard (Nguyen et al., 2025), MonarchAttention (Yaras et al., 2025), and Mamba-in-Llama (Wang et al., 2024) are the strongest adjacent repair/conversion baselines. They show that pretrained transformer attention can sometimes be converted to linear, gated, SSM, or hybrid alternatives with attention transfer, low-rank adaptation, distillation, or continual pretraining. This makes our LoRA section low-novelty as an adaptation method. Its role in this paper is narrower: it tests whether rows that fail the zero-shot S_p screen can be recovered for a specified retrieval workload, and whether that recovery survives task transfer, source shift, model-family checks, and language-modeling preservation.

E. Empirical setup details

Model inventory. The experiments are organized by panel rather than by a single fixed model list. The core fixed-bias NIAH audit covers fourteen released

decoder checkpoints summarized in Table 7: Mistral-7B-Instruct; Llama-3.2 1B/3B base and instruction variants plus Llama-3.1-8B-Instruct and Llama-3.3-70B-Instruct; Qwen2.5-3B/7B-Instruct and Qwen3-1.7B/32B (Qwen Team, 2024; 2025); Granite-3.3-8B-Instruct; and the robust Gemma-2 instruction-tuned 2B/9B outliers. Phi-4-mini and Pythia-1.4B (Biderman et al., 2023) are auxiliary regime controls, not part of the 12/14 headline. The row-centered NIAH panel uses twelve checkpoints; its packaged centered-control aggregate includes a separate Mistral-7B-v0.3 base cell and OLMo-2-0425-1B (OLMo Team et al., 2025), while the protected-head-set audit uses Mistral-7B-Instruct. The five-family language-modeling panel uses Mistral-7B-Instruct, OLMo-2-0425-1B, Llama-3.1-8B-Instruct, Qwen3-1.7B, and Granite-3.3-8B-Instruct. The 70B PG-19 row is a small $N = 3$ multi-seed exploratory extension on a single model (Llama-3.3-70B). Older diagnostic tables in Appendices C, F and G retain an early four-family audit (Llama-3.2-1B, Qwen3-1.7B, Gemma-4-E2B, Pythia-1.4B); those cells are explicitly labeled as diagnostic and are not the source of the 12/14-checkpoint headline. Exact checkpoint revisions and model-card URLs are recorded in model_manifest.tsv.

Architectures and loading. We use Hugging Face Transformers with open-weight decoder LMs loaded in bfloat16 and attn_implementation="eager". The hook intercepts each attention score row immediately before the model's softmax, after any model-native transformations such as masking or softcapping. It then either calls the original softmax or substitutes a row normalizer such as $P_{p,b}$, ROWAN-(p, τ), sparsemax, entmax, sigmoid, or a temperature-scaled softmax control. The value projections, output projections, residual stream, layer norms, MLPs, and decoding code path are otherwise untouched. Bit-identity between the hooked-softmax path and vanilla forward is verified in Appendix L.

Scoring and prompt suites. NIAH uses the single-template first-token exact-match probe described in Appendix H; needles are sampled per prompt from single-token common nouns and are checked not to appear in the haystack outside the planted position. NoLiMa uses substring scoring on the benchmark's paraphrased-needle setting. PG-19 and WikiText-103 are evaluated as language-modeling preservation tests with the token counts and seed counts reported in the corresponding tables. MMLU-STEM is evaluated as multiple-choice accuracy over the fixed STEM subset with paired tests for operator comparisons. Greedy generation is used only where the task is formulated as next-token or short-string retrieval; multiple-choice tasks use option scoring.

Fixed-bias grid and calibration. The stress recipe is always $P_{4,0}$. Grid-envelope columns use a fixed diagnostic grid over polynomial degree and bias, including the auto-calibrated bias b_{auto}. The auto-bias 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 $-b$. Calibration is done once per model on text not used in any prompt suite. Because the grid envelope is selected after observing the grid cells, it is reported only as a descriptive upper envelope, never as the headline fixed-bias result.

Numerical stability. S_p is computed in float64 using the relative form $S_p = \sum_{j \neq *} (r_j / r_*)^p$; we do not compute r^p at large magnitude before normalization. High-degree polynomial attention is also checked with a log-space implementation in Appendix O. For models with attention-logit softcaps, diagnostics use the post-cap row, which is the same input seen by the original softmax.

Statistical protocol. Binary retrieval cells report exact-match proportions, with Wilson intervals in the appendix where intervals matter. The Mistral protected-set certificate separates search prompts from held-out validation prompts and uses one-sided Clopper–Pearson bounds with Bonferroni correction to exclude all singletons under the stated retention floor. Random-set nulls are always reported with their candidate pool, draw count, and prompt set; comparisons across different pools or n values are not pooled. Language-modeling preservation reports means over seeds and seed standard deviations where multiple seeds are run.

F. 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 5 should be interpreted as an OOD stress reading rather than as an in-distribution prevalence claim.

G. Per-model ablation details

This section expands the group-ablation summaries across all four models on the literary and NIAH suites.

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 bottom-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 AA) 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 is consistent with benign behavior under the tested polynomial substitutions in this panel, and the behavioral test agrees. The result does not turn S_p into a certificate; it shows that low-S_p rows and benign downstream value directions coincide in this Llama panel, while the per-head causal signal is not separable 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 flags Pythia as a no-go case for this zero-shot substitution setting; 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 12 and 26 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.

H. 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 AD) 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 AC) asks for one of two distinct planted needles based on a final-position cue; the long-context variant (Appendix AD) extends pad to 512. Both use the same per-prompt random-needle construction.

I. Matched NIAH diagnostic: extended discussion

Table 11 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 W and the recipe-matched top-K alternative is in Appendix V.

Table 11. 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 4. "meas" is $\Pr(S_p < 1)$; "pred" is $\Pr(p \geq p^{\star}_{\text{exact}})$ from Corollary 2.3. "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 V and W.

			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. 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 X) 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 6 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.

J. Reproducibility

Code and data release. The supplementary artifact contains three manifest files (table_sources.yml, model_manifest.tsv, prompt_manifest.tsv) that together map the reported model-dependent cells to raw artifacts, HuggingFace revision hashes, and regeneration commands; an aggregation script (reproduce_main.sh --data-only) that regenerates the main-table report from the raw artifacts; frozen prompt parquets with content hashes for every retrieval suite (NIAH pad=0/128, RULER UUID, RULER-Dwarkesh, NoLiMa-main, NoLiMa-Hard); the protected-set discovery descent script (e3_ablation/find_min_softmax_cohort.py), the diagnostic hook with bit-identity tests, the per-head ablation runner, the Bonferroni screen script (cp_screen_followup.py), and the operator-similarity computation (compute_y_op_cohort.py); the eval driver (verify_mssc_cohort.py); and raw per-row parquets for the main-paper cells and packaged appendix audits listed in the supplement. A public release can additionally include the experimental Triton kernel (triton_selective_attn.py; not a paper contribution but useful for reproduction at scale), the LoRA repair adapters and per-objective training scripts (with op-

timizer/lr/batch/seeds disclosed), and the Modal job configurations used for cloud reruns.

Model revisions. Exact model revisions for the core 14-checkpoint audit, five-family panel, and auxiliary spot checks are recorded in the supplementary manifest. The early four-family diagnostic panel used the following immutable Hugging Face commit hashes (downloaded 2026-05-07): meta-llama/Llama-3.2-1B at 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; its exact revision is likewise in the manifest.

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 AA) uses seeds $\{0, \ldots, 19\}$.

Hook pseudocode. The diagnostic hook is a $\sim$30-line modification of the standard transformers eager and SDPA attention forward functions; the substitution recipe is the normalized row map in Section 2. 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 E).

Audit log. Appendix L traces each numerical claim to the specific script that produced it, including the bias-selection audits 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 $\sim$36 hours.

Experimental scaffolding. This subsection documents an opt-in implementation detail; it is not a contribution of the paper. The per-head protected-set intervention runs through a Python-level hook on the eager attention forward, which materializes the full $[B, H, S, S]$ score tensor twice (softmax path and polynomial path) and torch.where-merges by head. The Python hook is the source of truth for every number in the main text, including the 14-checkpoint audit, the protected-set and operator-transfer tables, the BOS-pathway tests, the downstream consumer interventions, the LoRA repair, and the PG-19 perplexity numbers.

We have also implemented a fused Triton kernel (e3_ablation/triton_selective_attn.py) that takes Q, K, V, a per-head boolean mask, polynomial degree p, and bias b, and applies softmax or $\mathrm{ReLU}(z+b)^p$-normalized attention per head in a single pass over keys with online normalization. The kernel exists to provide wall-clock relief for protected-set search at small K, not to serve as a deployment-grade implementation. Correctness: 6/6 test cases pass against the Python reference (bf16: max $|\mathrm{err}| \leq 4 \times 10^{-3}$; fp32: $\leq 1.5 \times 10^{-3}$, the limits set by reduction-order differences between torch.matmul and tl.dot, not bit-identity). Behavioral agreement on Mistral: $K = 20$, $n = 10$ prompts the Triton path and the Python hook agree 10/10 ($\Delta = 0$ pp accuracy); on $K = 40$, $n = 20$ they agree 20/20. Steady-state end-to-end speedup at Mistral $S \approx 1571$, $K = 20$ over the same Python hook: 1.74× (596 ms → 343 ms per prompt). We do not compare against FlashAttention-2 or FlashAttention-3: those are softmax-only kernels and the relevant comparison would be FA-2 on the softmax baseline plus our kernel on the selective-polynomial path, which we have not measured. Honest framing: this is experimental scaffolding for protected-set search, validated for correctness against the Python reference; we do not present it as a kernel-paper contribution and have not benchmarked it against production attention implementations. The kernel is disabled by default until it is validated across $p \in \{2, 8, 16\}$ recipes and across Granite/Llama/OLMo/GPT-NeoX families; Gemma's tanh softcap is detected and falls back to the Python hook automatically.

K. 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-head threshold (Olsson et al., 2022), 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.

L. 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 3 to 4 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.

Auto-bias selection validation: Qwen3. 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 incorrectly selected $b = 0$ instead. We replaced the picker with $\max(\text{biases}, \text{key} = \text{abs})$ 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 5 and the phase diagnostic are bit-identical (this column never depended on the picker). The Qwen3 $b = b_{\text{auto}}$ row in Table 20 changed from a mislabeled-$b = 0$ value to the correct $b = -3.36$ value; all reported headline numbers in Sections 3 to 4 use the post-fix code path.

Auto-bias selection validation: Gemma-4. The replacement picker $\max(\text{biases}, \text{key} = \text{abs})$ 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 the copying group-ablation summary: 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.

Column-access validation. Two ablation drivers originally used attribute-style column access for a field named head. Because this name can collide with a library method, the released scripts use explicit bracket access and include a regression check for the affected tables.

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 12 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 (Ap-

pendix AA); 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 12. Decisiveness of the diagnostic ranking at each K. "Decisive" means that top-K Δ-accuracy is strictly worse than every random-K sample. The Gemma-4-E2B starred values use the original literary-suite ranking at $b = -2$; at the corrected $b = +1.93$, the $K = 80$ cell becomes marginal, while the $K \leq 40$ cells remain decisive.

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

M. Calibrated bias by family

We use a 5%-mass-budget calibration: on held-out generic text, choose b so that the average attention row places at most 5 % of softmax mass on positions with score below $-b$. Since S_p uses $r_j = \mathrm{clamp}(z_j + b, 0, \tau)$, $b = 0$ and $b = b_{\mathrm{auto}}$ are distinct operating points. We report both: $b = 0$ (Tables 5 and 19) is calibration-free and supports the prevalence claim, while b_{auto} (Table 20) is complementary. The calibrated values are $+3.5547$ for Llama-3.2-1B, -3.3594 for Qwen3-1.7B, $+1.9258$ for Gemma-4-E2B, and $+119.7500$ for Pythia-1.4b. The wide spread is an architectural fingerprint, not a tuned headline hyperparameter.

N. Theory: proofs

Classical shift-invariance characterization

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 $\varphi(z) = Ce^{\beta z}$.

Softmax gap control

For Equation (3), factor the target score out of the denominator. If each distractor satisfies $z_j \le z_* - \Delta$ and there are at most A distractors, then

$$\sum_j e^{\beta z_j} \le e^{\beta z_*} \left(1 + Ae^{-\beta\Delta}\right).$$

Therefore $\mathrm{softmax}(\beta z)_* \ge (1 + Ae^{-\beta\Delta})^{-1}$. This comparison uses the full row support of softmax; the polynomial bound in Proposition 2.2 counts only active ReLU-polynomial distractors.

Proof of Proposition 2.2

For every active distractor in the lower-bound case, $r_j \le r_* - \Delta$, so

$$\left(\frac{r_j}{r_*}\right)^p \le \left(1 - \frac{\Delta}{r_*}\right)^p.$$

There are at most A such terms, hence

$$S_p(z, *) = \sum_{j \in D} \left(\frac{r_j}{r_*}\right)^p \le A \left(1 - \frac{\Delta}{r_*}\right)^p.$$

Using $P_{p,b}(z)_* = 1/(1 + S_p(z, *))$ gives Equation (4). Equality holds if and only if there are exactly A active distractors and all have $r_j = r_* - \Delta$.

For the converse, if at least A active distractors satisfy $r_j \ge r_* - \Delta$, then the same calculation reverses at the summation step:

$$S_p(z, *) \ge A \left(1 - \frac{\Delta}{r_*}\right)^p,$$

which gives the corresponding upper bound on $P_{p,b}(z)_*$. The sufficient-degree condition follows by requiring the lower bound to be at least 1/2 and rearranging. □

Lemma N.1 (Direct sign-compatibility criterion). Consider one head, target coordinate $*$, value-projected contributions c_j along an answer direction, and normalized polynomial ratios $q_j = (r_j/r_*)^p$ with $r_* > 0$ and $c_* \ne 0$. The substituted head preserves the target sign iff

$$\Psi_p \equiv -\frac{\sum_{j \ne *} q_j c_j}{c_*} < 1.$$

Equivalently, for any head set H with nonzero softmax answer projection $Y_{\mathrm{sm}}(H)$, the direct sign class is preserved by operator $\mathcal{O}$ iff $\Omega_{\mathcal{O}}(H) < 1$, where $\Omega_{\mathcal{O}}(H) = -(Y_{\mathcal{O}}(H) - Y_{\mathrm{sm}}(H))/Y_{\mathrm{sm}}(H)$.

Lemma N.2 (Aligned binary recall toy). In the aligned toy setting where the target value contributes positively and all distractor values contribute equally in the opposite answer direction, polynomial recall succeeds iff the target mass exceeds the distractor mass, equivalently $\sum_{j \ne *} (r_j/r_*)^p < 1$.

Proofs for Lemma N.1 and its special cases

Since $q_j = (r_j/r_*)^p \geq 0$ for all j (with $r_* > 0$ by hypothesis), the denominator $1+\sum_{j\neq *} q_j$ is strictly positive. Therefore

$$\operatorname{sign}(y_p) \;=\; \operatorname{sign}\Big(c_* + \sum_{j\neq *} q_j c_j\Big).$$

Write $S \equiv \sum_{j\neq *} q_j c_j$. We require $\operatorname{sign}(c_*+S) = \operatorname{sign}(c_*)$, which we verify by case on the sign of c_*.

Case 1: $c_* > 0$. Then $\operatorname{sign}(c_*+S) = +1$ iff $c_*+S > 0$ iff $S > -c_*$. Dividing by $c_* > 0$ preserves the inequality: $S/c_* > -1$, equivalently $-S/c_* < 1$, equivalently $\Psi_p < 1$.

Case 2: $c_* < 0$. Then $\operatorname{sign}(c_*+S) = -1$ iff $c_*+S < 0$ iff $S < -c_* = |c_*|$. Dividing by $c_* < 0$ flips the inequality: $S/c_* > -c_*/c_* = -1$, equivalently $-S/c_* < 1$, equivalently $\Psi_p < 1$.

In both cases the criterion is $\Psi_p < 1$. The two cases yield the same threshold because the sign flip from dividing by a negative number cancels the sign flip in the inequality direction; the algebra absorbs both. The hypothesis $c_* \neq 0$ excludes the degenerate case where sign-preservation is ill-posed (target contribution vanishes). □

For the head-set/operator identity, write $\Delta_{\mathcal{O}}(H) = Y_{\mathcal{O}}(H) - Y_{\mathrm{sm}}(H)$. By definition,

$$\Omega_{\mathcal{O}}(H) = -\frac{\Delta_{\mathcal{O}}(H)}{Y_{\mathrm{sm}}(H)}.$$

Multiplying by $Y_{\mathrm{sm}}(H)$ and rearranging gives

$$Y_{\mathcal{O}}(H) = Y_{\mathrm{sm}}(H) + \Delta_{\mathcal{O}}(H) = Y_{\mathrm{sm}}(H)(1 - \Omega_{\mathcal{O}}(H)).$$

Because $Y_{\mathrm{sm}}(H) \neq 0$, the sign of $Y_{\mathcal{O}}(H)$ equals the sign of $Y_{\mathrm{sm}}(H)$ exactly when the scalar multiplier $1 - \Omega_{\mathcal{O}}(H)$ is positive, i.e. exactly when $\Omega_{\mathcal{O}}(H) < 1$. The magnitude statement follows by taking absolute values:

$$|Y_{\mathcal{O}}(H)| = |Y_{\mathrm{sm}}(H)|\,|1 - \Omega_{\mathcal{O}}(H)|.$$

If $|Y_{\mathrm{sm}}(H)| \geq M$ and $0 \leq \Omega_{\mathcal{O}}(H) < 1$, then $|1 - \Omega_{\mathcal{O}}(H)| = 1 - \Omega_{\mathcal{O}}(H)$ and therefore $|Y_{\mathcal{O}}(H)| \geq M(1 - \Omega_{\mathcal{O}}(H))$.

Corollary N.3 (Operator-typing is geometric). Fix a head set H and prompt with $Y_{\mathrm{sm}}(H) \neq 0$. Let $\mathcal{S}_H = \prod_{h\in H} \Delta^{n-1}$ be the product of per-head attention simplices. The direct-compatible set

$$\mathcal{C}(H) = \left\{ a \in \mathcal{S}_H : \operatorname{sign} \sum_{h,j} a_j^{(h)} c_j^{(h)} = \operatorname{sign} Y_{\mathrm{sm}}(H) \right\}$$

is the relative intersection of $\mathcal{S}_H$ with an open half-space. Two operators have different direct-compatibility classes iff their head-set attention vectors lie on opposite sides of the value hyperplane $\sum_{h,j} a_j^{(h)} c_j^{(h)} = 0$.

Proof. For fixed values $c_j^{(h)}$, the map $a \mapsto \sum_{h,j} a_j^{(h)} c_j^{(h)}$ is linear on the ambient vector space of head-set attention vectors. The condition that this linear functional have the same sign as $Y_{\mathrm{sm}}(H)$ is therefore an open half-space condition. Intersecting that half-space with the product of simplices gives $\mathcal{C}(H)$. The boundary inside $\mathcal{S}_H$ is the hyperplane section where the linear functional is zero. By Lemma N.1, crossing this boundary is equivalently crossing $\Omega_{\mathcal{O}}(H) = 1$. □

Corollary N.4 (Operator-similarity preserves only direct sign class). For any two operators $\mathcal{O}_1, \mathcal{O}_2$ and head set H with $Y_{\mathrm{sm}}(H) \neq 0$, define

$$s_{\mathcal{O}_1,\mathcal{O}_2}(H) = \frac{Y_{\mathcal{O}_1}(H) Y_{\mathcal{O}_2}(H)}{Y_{\mathrm{sm}}(H)^2}.$$

Then

$$s_{\mathcal{O}_1,\mathcal{O}_2}(H) = (1 - \Omega_{\mathcal{O}_1}(H))(1 - \Omega_{\mathcal{O}_2}(H)).$$

Thus $s_{\mathcal{O}_1,\mathcal{O}_2}(H) > 0$ iff the two operators share the same direct-attribution sign class relative to softmax; it makes no claim about task-level rescue $R_{\mathcal{O}}$.

Proof. Apply Lemma N.1 twice:

$$\frac{Y_{\mathcal{O}_k}(H)}{Y_{\mathrm{sm}}(H)} = 1 - \Omega_{\mathcal{O}_k}(H), \qquad k \in \{1, 2\}.$$

Multiplying the two ratios gives the displayed identity. Since $Y_{\mathrm{sm}}(H)^2 > 0$, the sign of $s_{\mathcal{O}_1,\mathcal{O}_2}(H)$ is the product of the two direct-sign multipliers. Positive product means the multipliers share sign; negative product means they differ. The statement is purely about the direct projection Y and does not include downstream residual, MLP, or later-attention effects. □

N.1. Diffuse-sink amplification

Theorem N.5 (Diffuse-sink amplification bound). Let an attention head write value vectors $u_j \in \mathbb{R}^d$ at a fixed query position. Let

$$y^{\mathrm{sm}} = \sum_j a_j^{\mathrm{sm}} u_j, \qquad y^{\mathcal{O}} = \sum_j a_j^{\mathcal{O}} u_j,$$

with $y^{\mathrm{sm}} \neq 0$. Suppose $j^* = \arg\max_j z_j$, $a_{j^*}^{\mathrm{sm}} = \varepsilon \in (0, 1)$, $a_{j^*}^{\mathcal{O}} \geq 1 - \delta$, and $\|u_{j^*}\|_2 > 0$. Define

$$\bar{u} = \frac{1}{1-\varepsilon} \sum_{j \neq j^*} a_j^{\mathrm{sm}} u_j, \qquad R_u = \frac{\max_{j \neq j^*} \|u_j\|_2}{\|u_{j^*}\|_2}.$$

Then

$$\frac{\|y^{\mathcal{O}}\|_2}{\|y^{\mathrm{sm}}\|_2} \geq \frac{(1-\delta) - \delta R_u}{\|\varepsilon u_{j^*} + (1-\varepsilon)\bar{u}\|_2 / \|u_{j^*}\|_2}. \tag{11}$$

The bound is nontrivial when $(1 - \delta) > \delta R_u$.

Proof. First rewrite the softmax output by separating the argmax key from the tail:

$$y^{\mathrm{sm}} = a_{j^*}^{\mathrm{sm}} u_{j^*} + \sum_{j \neq j^*} a_j^{\mathrm{sm}} u_j = \varepsilon u_{j^*} + (1-\varepsilon)\bar{u}.$$

For the substitute operator, also separate the argmax key:

$$y^{\mathcal{O}} = a_{j^*}^{\mathcal{O}} u_{j^*} + \sum_{j \neq j^*} a_j^{\mathcal{O}} u_j.$$

By the reverse triangle inequality,

$$\|y^{\mathcal{O}}\|_2 \geq a_{j^*}^{\mathcal{O}} \|u_{j^*}\|_2 - \left\| \sum_{j \neq j^*} a_j^{\mathcal{O}} u_j \right\|_2.$$

The tail term is bounded by the triangle inequality:

$$\left\| \sum_{j \neq j^*} a_j^{\mathcal{O}} u_j \right\|_2 \leq \sum_{j \neq j^*} a_j^{\mathcal{O}} \|u_j\|_2 \leq \left(\sum_{j \neq j^*} a_j^{\mathcal{O}} \right) \max_{j \neq j^*} \|u_j\|_2.$$

Since $a_{j^*}^{\mathcal{O}} \geq 1 - \delta$, the substitute tail mass is at most δ. Therefore

$$\|y^{\mathcal{O}}\|_2 \geq (1-\delta)\|u_{j^*}\|_2 - \delta \max_{j \neq j^*} \|u_j\|_2 = \big((1-\delta) - \delta R_u\big)\|u_{j^*}\|_2.$$

Dividing this lower bound by the exact expression

$$\|y^{\mathrm{sm}}\|_2 = \|\varepsilon u_{j^*} + (1-\varepsilon)\bar{u}\|_2$$

gives Equation (11). □

Corollary N.6 (Tail-cancelled inverse-mass regime). Under the assumptions of Theorem N.5, suppose additionally that $(1-\delta)-\delta R_u \geq \gamma > 0$ and

$$\|(1-\varepsilon)\bar{u}\|_2 \leq \kappa\varepsilon\|u_{j^*}\|_2$$

for some $\kappa \geq 0$. Then

$$\frac{\|y^{\mathcal{O}}\|_2}{\|y^{\mathrm{sm}}\|_2} \geq \frac{\gamma}{(1+\kappa)\varepsilon}.$$

If $u_{j^*} \perp \bar{u}$, the denominator can be sharpened to $\varepsilon\sqrt{1+\kappa^2}$ under the same tail-smallness condition. Orthogonality alone is not enough for a $1/\varepsilon$ law; the averaged softmax tail must also be small in value space.

Proof. By the triangle inequality and the tail-smallness assumption,

$$\|y^{\mathrm{sm}}\|_2 = \|\varepsilon u_{j^*} + (1-\varepsilon)\bar{u}\|_2 \leq \varepsilon\|u_{j^*}\|_2 + \kappa\varepsilon\|u_{j^*}\|_2 = (1+\kappa)\varepsilon\|u_{j^*}\|_2.$$

Theorem N.5 gives $\|y^{\mathcal{O}}\|_2 \geq \gamma\|u_{j^*}\|_2$. Dividing yields the first bound. If $u_{j^*} \perp \bar{u}$, then Pythagoras gives

$$\|y^{\mathrm{sm}}\|_2^2 = \varepsilon^2\|u_{j^*}\|_2^2 + \|(1-\varepsilon)\bar{u}\|_2^2 \leq \varepsilon^2(1+\kappa^2)\|u_{j^*}\|_2^2,$$

which gives the sharpened denominator. □

Pointwise monomial dominance

Theorem N.7 (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.

Proof. 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 Lemma N.2

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

N.2. Scoped critical-head sufficient condition

Proposition N.8 (Scoped critical-head sufficient condition). Consider a single attention layer with H heads producing per-head answer-direction contributions $\alpha^{(h)} = \mathbf{u}^\top W_O^{(h)} \mathbf{V}^{(h)\top} \mathbf{a}^{(h)}$ along an answer direction $\mathbf{u}$, where $\mathbf{a}^{(h)} \in \Delta^{N-1}$ is the row attention vector and $\mathbf{V}^{(h)} \in \mathbb{R}^{N\times d}$ is the value matrix. Let $H^\star \subseteq H$ be answer-bearing heads with $S^\star = \sum_{h\in H^\star} \alpha_{\mathrm{softmax}}^{(h)} > 0$. Suppose: (i) the softmax residual outside $H^\star$ is bounded by $|\sum_{h\notin H^\star} \alpha_{\mathrm{softmax}}^{(h)}| \leq (1-\epsilon)S^\star$; (ii) the substituted answer-bearing contribution is directly retained, $\sum_{h\in H^\star} \alpha_{\mathrm{poly}}^{(h)} \geq \lambda S^\star$ for some $\lambda > 0$; and (iii) non-bearing heads are bounded after substitution, $|\sum_{h\notin H^\star} \alpha_{\mathrm{poly}}^{(h)}| \leq (1+\kappa)(1-\epsilon)S^\star$. Then $y_{\mathrm{poly}} > 0$ whenever $\epsilon > 1 - \lambda/(1+\kappa)$.

This is a scoped sufficient condition, not a transformer-output theorem and not a contribution. Its only role is to make explicit why all-head averages can be misleading: the behaviorally relevant quantity is retained contribution on answer-bearing heads plus bounded leakage elsewhere.

N.3. Bounded-level non-equivariance gap

Proposition N.9 (Bounded-level non-equivariance gap). Fix $A \geq 1$, $p > 0$, $b \in \mathbb{R}$, margin $\Delta > 0$, and a maximum active target level $R > \Delta$. Consider rows $z_*(s) = s - b$, $z_j(s) = s - b - \Delta$ $(j = 1, \ldots, A)$ with $s \in (\Delta, R]$. Then softmax target mass is constant in s: $m_{\mathrm{sm}} = 1/(1 + Ae^{-\Delta})$, while the polynomial target mass is $m_p(s) = 1/(1 + A(1 - \Delta/s)^p)$. Consequently

$$\sup_{s \in (\Delta, R]} \left| m_p(s) - m_{\mathrm{sm}} \right| \;\geq\; \tfrac{1}{2}\left(1 - \frac{1}{1 + A(1 - \Delta/R)^p}\right).$$

Proof. Softmax is shift-invariant in s, so $m_{\mathrm{sm}} = e^{s-b}/(e^{s-b} + Ae^{s-b-\Delta}) = 1/(1 + Ae^{-\Delta})$ does not depend on s. For the polynomial map, factor s^p from numerator and denominator: $m_p(s) = 1/(1 + A(1 - \Delta/s)^p)$. As $s \downarrow \Delta$, $1 - \Delta/s \to 0$, so $m_p(s) \to 1$. At $s = R$, $m_p(R) = 1/(1 + A(1 - \Delta/R)^p)$. For any constants x, y and c, $\max(|x - c|, |y - c|) \geq |x - y|/2$. Apply with $x = 1$, $y = m_p(R)$, $c = m_{\mathrm{sm}}$ to get the bound. □

This bound is tight only on the active target-maximal class (target above every distractor) and the specific tied-distractor row geometry used; without those qualifiers the target mass need not even attain $1/(A + 1)$. Empirically the bound is loose for Gemma-2 ($\tau = 50$ softcap) where the actual observed gap is well within bounds, suggesting the softcap is doing additional work beyond the bounded-level argument — consistent with the Gemma-2 substitution-robustness question raised in Limitations.

N.4. Cancellation-free sufficient condition $R_p < 1$

Corollary N.10 (Value-weighted sufficient condition). Under the assumptions of the single-head special case of Lemma N.1, let

$$R_p \;\equiv\; \frac{\sum_{j \neq *} q_j |c_j|}{|c_*|}.$$

If $R_p < 1$ then $\mathrm{sign}(y_p) = \mathrm{sign}(c_*)$.

Proof. The polynomial head output along u has positive denominator $1 + \sum_{j \neq *} q_j$, so its sign is the sign of $N = c_* + \sum_{j \neq *} q_j c_j$. By the triangle inequality $|\sum q_j c_j| \leq \sum q_j |c_j|$, so $|N| \geq |c_*| - \sum q_j |c_j| = |c_*|(1 - R_p)$, which is positive if $R_p < 1$. Thus N does not cross zero from c_*, so $\mathrm{sign}(N) = \mathrm{sign}(c_*)$. □

This is a stricter cancellation-free analog of the Ψ_p criterion in Lemma N.1: it gives the same conclusion ($\mathrm{sign}(y_p) = \mathrm{sign}(c_*)$) without relying on signed cancellation among distractor contributions. Practically, R_p is useful as a worst-case screen when adversarial cancellation is a concern; Ψ_p is the appropriate criterion when distractor signs distribute as expected.

Proof of Proposition N.8

Write $R = \sum_{h \notin H^\star} \alpha^{(h)}_{\mathrm{softmax}}$. Assumption (i) gives $y_{\mathrm{softmax}} = S^\star + R \geq \epsilon S^\star > 0$. Under substitution, write $\widetilde{S}^\star = \sum_{h \in H^\star} \alpha^{(h)}_{\mathrm{poly}}$ and $\widetilde{R} = \sum_{h \notin H^\star} \alpha^{(h)}_{\mathrm{poly}}$. Assumptions (ii)–(iii) give

$$y_{\mathrm{poly}} = \widetilde{S}^\star + \widetilde{R} \geq \lambda S^\star - (1 + \kappa)(1 - \epsilon) S^\star.$$

This is strictly positive when $\lambda > (1 + \kappa)(1 - \epsilon)$, equivalently $\epsilon > 1 - \lambda/(1 + \kappa)$. □

The proposition intentionally assumes retained answer-bearing contribution directly. A row-level condition such as $S_p < 1$ can justify such retention only under additional value-alignment assumptions like the toy setting in Lemma N.2; Lemma N.1 above drops the aligned-binary assumption and gives an exact per-head signed criterion in terms of value-projected contributions, at the cost of requiring a target answer direction. Empirically, we use S_p to localize candidate critical heads and then test them by intervention.

Theorem N.11 (Unbounded fixed-bias non-equivariance gap). Fix $A \geq 1$, $p > 0$, and $b \in \mathbb{R}$. Over target-maximal rows with A tied distractors and nonzero polynomial denominator,

$$\sup_z |P_{p,b}(z)_* - \mathrm{softmax}(z)_*| = \frac{A}{A+1}.$$

Moreover, on a fixed-softmax-level slice there are rows for which the gap is $A/(2(A+1))$. Thus no fixed-bias polynomial map can be uniformly close to softmax on this class while preserving softmax's row-shift invariance.

N.5. Proof of Theorem N.11

Lower bound (construction). Fix $A \geq 1$, $p > 0$, and any $b \in \mathbb{R}$. Let $\Delta_s \to \infty$ with $\Delta_s/s \to 0$ (concretely, $\Delta_s = \sqrt{s}$ for $s \geq 1$). Set

$$z_*(s) = s - b, \qquad z_j(s) = s - b - \Delta_s \quad (j = 1, \ldots, A).$$

Then $r_* = [z_*(s) + b]_+ = s$ and $r_j = [z_j(s) + b]_+ = s - \Delta_s$. Softmax target mass:

$$\mathrm{softmax}(z(s))_* \;=\; \frac{1}{1 + Ae^{-\Delta_s}} \;\xrightarrow{s\to\infty}\; 1.$$

Polynomial target mass:

$$P_{p,b}(z(s))_* \;=\; \frac{1}{1 + A(1 - \Delta_s/s)^p} \;\xrightarrow{s\to\infty}\; \frac{1}{A+1}.$$

Hence $\sup_s |P_{p,b}(z(s))_* - \mathrm{softmax}(z(s))_*| \to A/(A+1)$.

Upper bound (matching). Restrict to target-maximal rows with $z_* + b > 0$ (non-degenerate domain: otherwise $P_{p,b}$ is undefined as $0/0$). Then $[z_* + b]_+ \geq [z_j + b]_+ \geq 0$ for every j, so for any tied-distractor row with A distractors,

$$\tfrac{1}{A+1} \;\leq\; P_{p,b}(z)_* \;=\; \tfrac{[z_*+b]_+^p}{[z_*+b]_+^p + \sum_{j=1}^{A} [z_j+b]_+^p} \;\leq\; 1,$$

and identically $\frac{1}{A+1} \leq \mathrm{softmax}(z)_* \leq 1$ (since softmax of a maximal-target row is monotone in Δ). Both values lie in the interval $[1/(A+1),\, 1]$ of length $A/(A+1)$, so $|P_{p,b}(z)_* - \mathrm{softmax}(z)_*| \leq A/(A+1)$ for every z in the domain. Combined with the construction above, $\sup_z |P_{p,b}(z)_* - \mathrm{softmax}(z)_*| = A/(A+1)$ exactly.

Fixed-softmax-level corollary. Fix $\Delta = \log(A+2)$, so softmax stays at $(A+2)/(2(A+1))$ for all s, while polynomial mass moves from 1 (as $s \downarrow \Delta$) to $1/(A+1)$ (as $s \to \infty$): both endpoints sit at distance $A/(2(A+1))$ from softmax. The upper bound for the fixed-softmax-level case is again $A/(2(A+1))$ by the same interval-diameter argument applied to the restricted level set $\{z : \mathrm{softmax}(z)_* = (A+2)/(2(A+1))\}$. □

N.6. Direct logit attribution can miss compositional heads

A constructive existence claim that motivates the structural-functional dissociation observation: there exist transformer heads whose direct logit attribution (DLA) along an answer direction is zero even though their causal effect on the output along that direction is $\Omega(1)$. Direct-logit-attribution caveats of this kind are discussed by Janiak et al. (2023), and the MSC heads $(0, 17)$ and $(0, 28)$ on Mistral are empirical near-instances (bottom-2 % on $|c_*|$, causally essential under polynomial substitution).

Proposition N.12 (DLA-blindness for compositional heads). There exist 2-layer transformers and attention heads h_0 such that the direct logit attribution $\alpha^{(h_0)} = u^\top \sum_j a_j^{(h_0)} W_O^{(h_0)} V_j^{(h_0)}$ along an answer direction u is exactly zero, while the ablation-causal contribution $\beta^{(h_0)} = u^\top \big(\mathbf{logit}_q(T) - \mathbf{logit}_q(T \mid h_0 \to \mathbf{0})\big)$ is bounded below by an arbitrary positive constant under the experimenter's control. Any head-ranking method that orders heads by $|\alpha^{(h)}|$ alone therefore ranks such h_0 in the bottom $O(\epsilon_{\mathrm{num}})$ regardless of $|\beta^{(h_0)}|$.

Construction. Let T have $L = 2$ layers, single head per layer, residual dimension $d = 2$, input sequence (BOS, x, y) at positions $(0, 1, 2)$, and final query at $q = 2$. Take $u = (1, 0)^\top$. Set the layer-0 head h_0:

$$W_V^{(0)} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad W_O^{(0)} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix},$$

with $W_Q^{(0)}, W_K^{(0)}$ producing softmax score gap $\sigma \equiv \mathrm{score}(q, \mathrm{BOS}) - \max_{j \neq 0} \mathrm{score}(q, j) \geq \log\big(2(T-1)/\eta\big)$ for any chosen $\eta \in (0,1)$, so that $a_{\mathrm{BOS}}^{(0)} \geq 1-\eta$ at q (such $W_Q^{(0)}, W_K^{(0)}$ exist for any $\eta > 0$ since softmax-attention scores are unbounded). Take embeddings $e_{\mathrm{BOS}} = (1,0)$, $e_x = e_y = (0,0)$. Then h_0's contribution to the residual stream at q is $W_O^{(0)} W_V^{(0)} (a_{\mathrm{BOS}}^{(0)} e_{\mathrm{BOS}}) = (0,\, a_{\mathrm{BOS}}^{(0)})^\top$, and exactly $\alpha^{(h_0)} = u^\top (0,\, a_{\mathrm{BOS}}^{(0)})^\top = 0$ regardless of $a_{\mathrm{BOS}}^{(0)}$, since the entire dim-2 contribution is orthogonal to u.

For the layer-1 head h_1, set $W_K^{(1)} = \left(\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ (extracting the dim-2 component of the residual stream) and $W_Q^{(1)} = (0,1)^\top$. When h_0 is active, the residual stream at q has dim-2 component $a_{\mathrm{BOS}}^{(0)} \geq 1-\eta$, so h_1's key-query inner product at q is $\geq 1-\eta$ versus 0 at positions BOS, x, y; choosing the same softmax-score scaling for h_1 gives $a_q^{(1)} \geq 1-\eta$. When h_0 is ablated to $\mathbf{0}$, the residual at q has dim-2 component 0, so h_1's keys at all positions are 0 and h_1 attends uniformly ($a_j^{(1)} = 1/T$ for every j). Choose $W_V^{(1)}, W_O^{(1)}$ so that the final logit along u in the "h_1 attends to q" regime is 1 and in the "h_1 attends uniformly" regime is 0. Then $\beta^{(h_0)} \geq 1-\eta$ while $\alpha^{(h_0)} = 0$. Taking $\eta \to 0$ gives $\beta^{(h_0)} \to 1$ with $\alpha^{(h_0)}$ identically zero. The construction therefore separates DLA from causal contribution by a constant $\gamma \in (0,1)$ at any finite-gap softmax precision $\eta < 1-\gamma$. □ □

This is an existence proof, not a tightness result: it shows that $\alpha = 0$, $\beta = \Omega(1)$ heads can exist by construction. The construction is contrived (hand-tuned W_V, W_O, single head per layer) and the result should be read as "DLA can miss compositional heads," not "DLA must miss them on real models." Its purpose in this paper is to justify the empirical observation that the Mistral MSC heads (bottom-2 % $|c_*|$, top-0.5 %/bottom-2 % Ω_p, 42× /9× polynomial amplification) sit in exactly the regime the proposition predicts: small direct logit attribution, large compositional sensitivity to downstream heads. The qualitative message — DLA is a zero-order residual-stream path term (Wang et al., 2023; Janiak et al., 2023), and some compatibility heads may matter through higher-order paths — motivates causal validation beyond attribution. Activation/path patching (Conmy et al., 2023) would be needed to trace a circuit; our protected-set intervention establishes functional load under substitution, not a full path-level mechanism.

N.7. Tail-index law for distractor concentration

Corollary 2.3 is a worst-case sufficient envelope; on real rows the slack between predicted and measured S_p is often large (the lower-bound diagnostic). The following distributional refinement explains why the slack scales with the upper-tail behavior of the distractor-ratio distribution.

Theorem N.13 (Tail-index law). Let $X_1, \ldots, X_A \in [0,1]$ be i.i.d. distractor-to-target ratios $X = r_j / r_*$, and let $S_{A,p} = \sum_{j=1}^{A} X_j^p$. Assume the upper tail near 1 is regularly varying:

$$\Pr(X \geq 1-t) \;\sim\; c\, t^{\alpha} \qquad \text{as } t \downarrow 0$$

for some constants $c > 0$, $\alpha > 0$. Then $\mathbb{E}[X^p] \sim c\,\Gamma(\alpha+1)\, p^{-\alpha}$ and hence $\mathbb{E}[S_{A,p}] \sim Ac\,\Gamma(\alpha+1)\, p^{-\alpha}$. If $A_p \to \infty$, then $S_{A_p,p} \to 0$ in probability when $A_p/p^\alpha \to 0$, and $S_{A_p,p} \to \infty$ when $A_p/p^\alpha \to \infty$. The distributional concentration threshold is therefore $p_{\mathrm{crit}} \asymp (Ac\,\Gamma(\alpha+1))^{1/\alpha}$.

Sketch. By the tail-integral identity $\mathbb{E}[X^p] = p\int_0^1 x^{p-1} \Pr(X \geq x)\, dx$, substitute $x = 1 - y/p$ to obtain $\int_0^p (1 - y/p)^{p-1} \Pr(X \geq 1 - y/p)\, dy$. For fixed y, $(1-y/p)^{p-1} \to e^{-y}$ and the tail assumption gives $\Pr(X \geq 1-y/p) \sim c(y/p)^\alpha$. Dominated convergence under Potter bounds yields $p^\alpha\, \mathbb{E}[X^p] \to c\int_0^\infty e^{-y} y^\alpha dy = c\Gamma(\alpha+1)$ (Bingham et al., 1987; de Haan and Ferreira, 2006). Summing over A i.i.d. distractors gives the expectation $\mu_{A,p} \equiv \mathbb{E}[S_{A,p}] \sim Ac\Gamma(\alpha+1)p^{-\alpha}$.

Upper branch ($A_p/p^\alpha \to 0$): $\mu_{A_p,p} \to 0$, so Markov $\Pr(S_{A_p,p} \geq \varepsilon) \leq \mu_{A_p,p}/\varepsilon \to 0$ for any $\varepsilon > 0$.

Lower branch ($A_p/p^\alpha \to \infty$): $\mu_{A_p,p} \to \infty$. By the same tail argument applied to X^{2p}, $\mathbb{E}[X^{2p}] \sim 2^{-\alpha} c\Gamma(\alpha+1) p^{-\alpha}$, hence $\mathrm{Var}(S_{A_p,p}) \leq A_p\, \mathbb{E}[X^{2p}] \sim A_p\, 2^{-\alpha} c\Gamma(\alpha+1) p^{-\alpha}$. The relative variance

$$\frac{\mathrm{Var}(S_{A_p,p})}{\mu_{A_p,p}^2} \;\sim\; \frac{2^{-\alpha}\, p^\alpha}{A_p\, c\,\Gamma(\alpha+1)} \;=\; O\big(p^\alpha / A_p\big) \;\to\; 0,$$

so by Chebyshev $S_{A_p,p}/\mu_{A_p,p} \to 1$ in probability, and $\mu_{A_p,p} \to \infty$ implies $S_{A_p,p} \to \infty$ in probability. □

Theorem N.13 is parallel to Corollary 2.3, not a replacement. Corollary 2.3 is a max-ratio worst-case envelope. Under the same regularly-varying row model the sampled maximum $M_A = \max_j X_j$ satisfies $1 - M_A \asymp A^{-1/\alpha}$, so plugging M_A into the worst-case envelope yields $p^\star \asymp A^{1/\alpha} \log A$. The distributional mass threshold of Theorem N.13 is $p_{\mathrm{crit}} \asymp A^{1/\alpha}$ up to constants. Thus, on this random-row model, the tail-average theorem is typically a log-factor less conservative than applying the max-ratio worst-case envelope to a typical row. This explains the observed conservative slack in the lower-bound diagnostic on safer models without contradicting Corollary 2.3: the corollary remains the adversarial row certificate, while Theorem N.13 describes average mass under an additional distributional assumption. At our $A \sim 10^2$–10^3, estimating α from the top 1–5 % of distractor ratios (Hill / Pickands / moment estimators) is genuinely noisy; we therefore treat Theorem N.13 as theoretical context explaining the slack envelope, not as a deployable per-head predictor.

O. Precision-controlled rerun of $p=16$ cells

An earlier reading of the $p = 16$ cells noted "a high-degree substitution/implementation interaction we have not isolated." We ruled out the precision-artifact hypothesis by reimplementing polynomial attention in log space,

$$a_j \propto \exp\big(p\,(\log r_j - \log r_{\max})\big),$$

which avoids overflow/underflow in r^p for high p and wide-dynamic-range r. We rerun the NIAH suite ($n = 200$, four pad configs) at $p = 16$, $b \in \{0, b_{\text{auto}}\}$ on Llama-3.2-1B, Qwen3-1.7B, Gemma-4-E2B in three precisions: the original bf16 naive (Table 9), stable log-space at fp32, and stable log-space at fp64. Synthetic validation shows the naive and stable functions agree to fp32 epsilon for all (p, b) except $p = 16, b = -3.36$ (Qwen3's b_{auto}), where the per-row target mass disagreement is up to 0.99 on randn-scaled inputs.

Table 13. $p=16$ NIAH accuracy ($n=200$ per cell) across precisions. "bf16" is the original naive implementation. "fp32"/"fp64" use the log-space stable variant at the indicated precision. All three precisions give the same outcome within Wilson noise floor on every cell, ruling out a precision artifact: the $p = 16$ behavior on each model reflects row geometry, not numerical implementation.

Model	Bias	pad	bf16	fp32	fp64
Llama-1B	$b=0$	0	0.0	0.0	0.0
		128	0.5	0.0	0.0
	b_{auto}	0	1.5	1.5	1.5
		128	5.0	6.0	5.0
Qwen3-1.7B	$b=0$	0	6.0	6.0	6.5
		128	49.5	51.5	50.5
	b_{auto}	0	2.0	8.5	7.5
		128	6.5	3.0	9.5
Gemma-4-E2B	$b=0$	0	25.5	25.5	25.5
		128	74.0	72.5	74.5
	b_{auto}	0	97.0	97.0	97.0
		128	99.0	100.0	100.0

Reading the table. The three precisions agree on every cell within Wilson noise (max disagreement ≈ 6 pp at Qwen3 b_{auto}, all in $[3, 10]$ %, all behaviorally indistinguishable from collapse). The Qwen3 b_{auto} case is particularly informative: synthetic validation showed 0.99 target-mass disagreement between naive and stable, yet behavioral retrieval collapses identically across all three precisions. The synthetic disagreement therefore did not propagate to a behavioral recovery — the polynomial substitution at $p = 16, b_{\text{auto}}$ on Qwen3 is row-geometrically incapable of preserving retrieval, regardless of how the per-row mass is computed.

For Llama, the $b = 0$ case at $p = 16$ is uniformly 0 % across precisions; b_{auto} is uniformly ≤ 6 %. For Gemma, b_{auto} is uniformly ≥ 99 % — the $p = 16$ Gemma cell was already preserving in bf16. The implication: the earlier framing "we have not isolated this" is replaced by "the row-level theory is consistent with behavior at all three precisions; the $p=16$ collapse on Llama is geometric, not numerical."

P. Cross-family validation

The development audit (Llama-3.2-1B, Qwen3-1.7B, Gemma-4-E2B, Pythia-1.4b) is small. We extend the prevalence diagnostic to three additional decoder LMs from distinct training and architecture lineages: Mistral-7B-Instruct-v0.3 (Jiang et al., 2023) (Mistral family, 7B), OLMo-2-0425-1B (OLMo Team et al., 2025) (OLMo family, fully open recipe, 1B), and Phi-4-mini-instruct (Microsoft, 2025) (Microsoft Phi family, 3.8 B, dense). All three load under the diagnostic hook with max-diff 0 vs. vanilla forward on a generic-text input (bit-identity verified). The auto-calibrated bias b_{auto}: Mistral +3.30, OLMo-2 +0.25, Phi-4-mini +3.63.

Prevalence at $p = 2, b = 0$ (induction $L = 4096$). Table 14 reports the polynomial-unsafe rate alongside the existing four-model audit.

Table 14. Cross-family extension of Table 5 on induction at $L = 4096$, $b = 0$, active-unsaturated rows. The family pattern of the original four-model audit extends cleanly. Three additional models confirm the partition: Mistral-7B and Phi-4-mini cluster with Llama in the safe corner (small A, high ρ, ~ 40 % unsafe at $p = 2$); OLMo-2-1B clusters with Gemma/Qwen3 in the unsafe corner (high A, low ρ, ≥ 85 % unsafe at $p=2$). The 7-model audit lands 3 safe / 4 unsafe with no model breaking the (A, ρ) predictor.

Model	$p=2$	$p=4$	$p=8$	A_{med}	ρ_{med}
Original four-model audit:					
Llama-3.2-1B	19.6	10.3	4.7	2	0.73
Qwen3-1.7B	94.0	79.3	48.5	1884	0.21
Gemma-4-E2B	71.3	56.9	37.6	36	0.14
Pythia-1.4b	99.2	98.3	95.1	1407	0.03
Cross-family extension (this work):					
Mistral-7B-Instr	40.0	21.7	10.1	7	0.57
Phi-4-mini	39.2	21.9	11.0	5	0.55
OLMo-2-0425-1B	86.2	70.5	44.7	117	0.12

Reading the table. The seven-model audit clusters cleanly along the diagonal of (A, ρ) space: Llama, Mistral, and Phi-4-mini anchor the safe corner (small A, high ρ, low unsafe rate); Pythia and Qwen3 anchor the unsafe corner (large A, low ρ, high unsafe rate); Gemma and OLMo-2 sit in the middle (moderate-to-large A, low ρ, high unsafe rate). The family pattern is

therefore not a small-audit artifact: at three additional families spanning 1–7 B parameters and distinct training pipelines (Mistral instruct, OLMo-2 fully-open recipe, Phi-4-mini dense), the diagnostic landed where Cor. 2.3 predicts based on the row-level (A, ρ) geometry.

Cross-family NIAH. Table 15 reports the substitution sweep at pad = 128 on the three new models. The behavioral pattern matches the prevalence prediction: Mistral and Phi-4-mini preserve at relu_p4_bauto and relu_p8_bauto (joining Llama and Gemma in the safe-recipe partition); OLMo-2 collapses on relu_p4_bauto and only partially recovers at relu_p8_bauto (40 %, joining Qwen3 in the no-preserving-recipe partition). Phi-4-mini is notable: it preserves at three polynomial cells (including the uncalibrated relu_p2_b0 at 99.5 %), making it the most polynomial-friendly model in our audit.

Table 15. Cross-family NIAH at pad = 128 (n = 200). Sample-perfect cells in bold (Wilson 95 % lower bound 98.1 %). Mistral preserves at relu_p4_bauto/relu_p8_bauto at the model's own softmax baseline of 95 % (i.e., the recipe matches softmax). Phi-4-mini preserves at relu_p2_b0/relu_p4_bauto/relu_p8_bauto (3 cells, sample-perfect). OLMo-2 has no sample-perfect recipe; its best cell, relu_p8_bauto, drops −60 pp.

Model	softmax	relu_p2_b0	relu_p4_bauto	relu_p8_bauto	sigmoid_b0
Mistral-7B-Instr	95.0	39.0	**95.0**	**95.0**	5.5
Phi-4-mini	**100.0**	99.5	**100.0**	**100.0**	0.0
OLMo-2-0425-1B	**100.0**	0.0	0.0	40.0	0.0

Q. Concentration envelopes across attention map families

Table 16. Concentration envelopes for pointwise attention maps. A is active-distractor count, $\Delta = z_* - \max_{j \neq *} z_j$ the score margin, $r_i = [z_i + b]_+$ the post-bias ReLU level for polynomial attention, 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. 2.3). 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)	$\max(z+b,0)^p$	$1/(1+S_p) \geq 1/(1+A(1-\Delta/r_*)^p)$
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

R. Lower-bound validation: extended

Corollary 2.3 predicts that to keep target mass $\widetilde{a}_* \geq 1/2$ against A active distractors with relative margin $\rho = \Delta/r_*$, 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 $p \geq p^*$. This appendix gives the full classification breakdown.

Table 17 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.

Table 17. Lower-bound classification at degree p. "Agree" is overall predicted/measured agreement (over all evaluated active-unsaturated rows). "FN at $p = 2$" has rate $\text{FN}/(\text{FN}+\text{TP}_{\text{unsafe}})$, 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 $\text{FP}/(\text{FP}+\text{TN}_{\text{safe}})$, 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 ($S_p \leq 1+\epsilon$), the FN count drops to 0 for every model at every $p \in \{2, 4, 8, 16\}$ (Appendix U), confirming that the bound is conservative up to numerical tolerance, exactly as Corollary 2.3 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 $b_{\text{auto}} = +119.75$.

The asymmetry in Table 17 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 by the bound being tight: Pythia's measured-unsafe rate is > 99 % at $p = 2$ (Table 5), so a constant-"unsafe" classifier already achieves ~ 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.

S. $p = 4$ 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 18 reports the result.

Table 18. 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

T. Per-suite tables

Tables 19 and 20 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 19. Per-suite diagnostic at the longest context with $b=0$, restricted to active-unsaturated rows.

Model	Suite	n_{rows}	unsafe %	med Δ/r_*	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 20. 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/r_*	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

U. 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 2.3,

$$p^{\star}_{\text{worst}}(A,\rho) = \frac{\log A}{\log(1/(1-\rho))}, \quad \rho \in [0, 0.99].$$

not the small-ρ approximation $(r_*/\Delta)\log A$. 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, $p \geq p^{\star}_{\text{worst}}$ implies $A(1-\rho)^p \leq 1$, hence $S_p \leq 1$. Corollary 2.3 certifies this weak row-mass pass condition; our strict measured-safe criterion uses $S_p < 1$, 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 11, 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 17, 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 Table 11 reports the conservative, strict number) and disclose the boundary-only nature of the residual FN here. The takeaway is that the possible approximate-threshold concern—does the implementation use the small-ρ approximation, or are false negatives 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.

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

Table 11 reports top-K measured-safe and predicted-safe percentages restricted to the fixed $K = 40$ heads chosen by the Section 4 causal ranking—computed once per model from the E1 induction parquet at $(p = 2, b = b_{\text{auto}})$ and reused for every recipe in the table. An alternative is to recompute the top-K head set per recipe at the recipe's own (p, b) (a "recipe-matched" top-K); we report this alternative in Table 21 for transparency.

Table 21. Top-K measured-safe and predicted-safe percentages on the matched NIAH rows, comparing the fixed Section 4 causal head set (used in Table 11) 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

We use the fixed causal head set in Table 11 because the paper's narrative is about behaviorally critical heads (the heads tested by ablation in Section 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 this semantics, the fixed causal set is the more meaningful comparison.

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

Table 11 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 11: 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 22 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 4 and 6 relies on is well-separated relative to its bootstrap CI width: the within-Llama non-monotonicity (72.3 $\rightarrow$ 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.

Table 22. Prompt-level 95 % bootstrap intervals for the headline cells of Table 11. "med" is the bootstrap median (matches Table 11); "[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]

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

X. Answer-key-target diagnostic

Table 11 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 11 and report it in Table 23. 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 2.3 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 23 restricts to those triples and reports answer-key meas-safe and pred-safe percentages: they match the all-head columns of Table 11 to within 5 pp on every cell. For example, Llama relu_p8_bauto has 86.4 % meas-safe in Table 11 (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 4, is that NIAH retrieval flows through a small fraction of (layer, head) pairs, and Table 11 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 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 4 top-$K = 40$ head set.

Table 23. Answer-key-target diagnostic on the same NIAH rows as Table 11. "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 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

Y. 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 row-mass condition is exactly $S_\phi < 1$. The polynomial degree bound (Corollary 2.3) is specific to $\phi(r) = r^p$, but the row-ratio diagnostic itself is general. Theorem N.7 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 $S_\phi < 1$ row-mass condition is general; the degree-ordering theorem is not.

We compute S_σ on the same NIAH rows as Table 11, using $\phi(r) = \sigma(r) = 1/(1 + e^{-r})$ with $r = z + b$ and target $* = \arg\max z$ (no clamp; sigmoid is everywhere positive). Table 24 reports the result. The row-mass screen aligns with the substitution sweep's behavior: at $b = 0$ and $b = b_{\text{auto}}$, every model has ≤ 0.5 % measured-safe rows, consistent with both sigmoid recipes (sigmoid_b0, sigmoid_bauto) collapsing on every model in Table 9 (NIAH, −100 pp on 7 of 8 cells), and also performing poorly on the literary-

completion suite in Table 6 (−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 24. Sigmoid-target diagnostic S_σ on the matched NIAH rows. "% meas safe" is the fraction of rows with $S_\sigma < 1$. The row-mass screen flags sigmoid substitution as a no-go case on every model; this is consistent with the sigmoid_b0 and sigmoid_bauto rows of both Table 9 (NIAH) and Table 6 (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

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

the phase diagnostic colors heads by the sink/local/diffuse taxonomy of Appendix K, while Section 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 25. 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 bottom-K row gives the analogous bottom-of-ranking composition.

		top-K			bottom-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

Table 25 shows a clean correspondence: the top-K heads—the heads the causal ablation in Section 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 bottom-K heads are predominantly sink. This is the architectural reason behind the bottom-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. The phase taxonomy and Section 4's causal ranking therefore tell the same story from two angles.

AA. Random-K decisiveness with 20 draws

The main NIAH group-ablation summary 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 26 reports the result.

Table 26. Random-K decisiveness on NIAH at pad=128 with 20 draws (vs the 5 draws of the NIAH group-ablation summary). "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 NIAH summary 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.

AB. 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 27 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 5.

Table 27. 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 5): $\geq 95\,\%$ 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/r_*	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

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 19), 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 5 reports for the smaller models.

AC. 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 28 reports the substitution sweep at pad=32 on all four models.

Table 28. 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 9) 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 9: Qwen3 collapses on every recipe (softmax $\rightarrow$ 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.

AD. Long-context NIAH (pad=512, $\sim$6K tokens)

We extend the NIAH suite to pad=512, producing prompts of $\sim$6 K tokens—roughly 4× longer than the pad=128 ($\sim$1.5 K) suite of Table 9. Pythia trained at 2048 tokens and is excluded from this run. Table 29 reports the substitution sweep on Llama, Qwen3, and Gemma-4.

Table 29. Long-context NIAH at pad=512 ($\sim$ 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 9 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 4—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 9, −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 ($\sim$ 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.

AE. 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 H; total prompt length is $\sim$1500 tokens; 200 prompts per cell with deterministic per-prompt random plants (seed 7000).

Table 30. Adversarial-decoy NIAH at pad = 128 ($\sim$ 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 9 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 locks onto "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 disambiguation 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 9 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.

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

To test whether the architectural partition depends on the single-token English-noun needle of Section 4, 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 31. UUID-needle NIAH at pad = 128, exact-match accuracy on 200 prompts. Comparison columns show the corresponding English-noun cell from Table 9.

	UUID-needle			English-noun (Table 9)		
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.

AG. Llama-3.2-3B prevalence spot check

Mirroring the Qwen3-8B prevalence diagnostic of Appendix AB on the Llama family. Table 32 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 5. Llama-3.2-3B's auto-calibrated bias is $b_{\text{auto}} = +4.50$, larger than Llama-3.2-1B's +3.55.

Table 32. Llama-3.2-3B unsafe-row rate at $L = 4096$, $b = 0$, active-unsaturated rows, side-by-side with Llama-3.2-1B (Table 19). Prevalence is similar at the 1.4–3 B Llama scale, consistent with the diagnostic stability we observe on the substitution sweep (Appendix AH).

Suite	$p=2$ (3B)	$p=2$ (1B)	$p=4$ (3B)	$p=8$ (3B)	med Δ/r_*	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 AH and AI.

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

We extend the substitution sweep of Table 9 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 (Grattafiori et al., 2024) ($b_{\text{auto}} = +4.50$). All three Llama models are evaluated on the same English-noun NIAH suite at pad = 128 ($n = 200$, prompts ∼1500 tokens). Table 33 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.

Table 33. 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

AI. Llama-3.2-3B group head ablation

We replicate the group-ablation experiment 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 34 reports top-$K$, bottom-$K$, 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 34. 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 12: 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 bottom-K polynomialization at every K tested

on both suites, confirming the per-model finding from Section 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.

AJ. Per-head ablation: retrieval is distributed

The group-ablation results are decisive at $K \geq 20$ on Qwen3 and Gemma-4: top-K polynomialization drops behavior -80 to -100 pp while bottom-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 35 summarizes.

Table 35. 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 retrieval behavior intact. Combined with the $K \geq 20$ decisiveness of the NIAH group-ablation summary, this implies the substitution-sensitive behavior is distributed: no single head carries the effect, but a 20-head subset does.

Model	heads tested	nonzero Δ	Pearson r	Spearman ρ
Qwen3-1.7B	40 (top-20 + bottom-20 by S_p)	0	undef.	undef.
Gemma-4-E2B	40 (top-20 + bottom-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 protected set of high-S_p heads rather than a per-head failure-mode predictor.

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

Appendix AA 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 the copying group-ablation summary. Table 36 reports.

The decisiveness claim of the copying group-ablation summary 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.

Table 36. 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

AL. $n = 500$ 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 37 reports.

Table 37. $n = 500$ 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

AM. 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 9 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 38 reports.

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 ar-

Table 38. 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 9). 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	—	—	—

chitectural 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 9); the 0 % Gemma cells in this table reflect the wrong-degree choice.

AN. 2026 post-cutoff mixed-source retrieval audits

The original RULER-template stress check used a uniformly post-cutoff Dwarkesh haystack (Appendix AO). That is useful for contamination control but source-narrow: one could reasonably ask whether repair succeeds only on an interview/blog distribution. We therefore build balanced 2026 post-cutoff mixed-source suites and matched no-Dwarkesh suites.

Source construction. We gather 2026 post-cutoff sentences from independent or non-model-provider sources: NASA, NASA Science, NOAA, UN News, EU press, WHO, NIST, Federal Reserve, Smithsonian, and several arXiv categories (CS/AI, q-bio, econ, physics, astronomy), plus Dwarkesh as one source. We cap retained sentences per source at 300 before sampling, and each prompt is constructed by round-robin sampling across sources. Google/DeepMind sources are excluded because provider-internal training exposure would be hard to rule out for Gemma-family models.

Table 39. Source balance in 2026 post-cutoff retrieval audits. Fractions are aggregate sentence-source fractions over generated prompts. Balanced suites include Dwarkesh as one equal source slot; no-Dwarkesh suites remove it entirely.

Suite	Prompts	Sources/prompt	Source-fraction range
Balanced single-key	100	15	6.61–6.76 % (Dwarkesh 6.64 %)
Balanced multikey	50	15	6.51–6.78 % (Dwarkesh 6.66 %)
No-Dwarkesh single-key	100	14	7.03–7.22 %
No-Dwarkesh multikey	50	14	7.02–7.24 %

Table 45 reports the resulting Qwen3 multikey-adapter behavior. An earlier mixed-source pilot reached 97/100 polynomial+LoRA on single-key, but the run may have used a pre-rebalance parquet and lacks source-count metadata; we therefore do not use it for paper claims. The paper-facing balanced and no-Dwarkesh results are the four rows in Table 45.

AO. RULER-Dwarkesh: post-cutoff retrieval probe

AO.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.

AO.2. Substitution sweep

Table 40 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 40. RULER-template substitution sweep with custom post-cutoff Dwarkesh haystack ($n = 50$, ~ 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 41). 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 ~ 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 2.3: 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) goes to 0 % uniformly. The diagnostic at $p = 16$ predicts lower unsafe-row fraction than $p = 8$ (Table 41), but behavioral substitution fails. A precision-controlled rerun in log-space fp32 and fp64 (Appendix O) gives the same outcome as bf16 on every $p = 16$ cell, ruling out numerical precision. The collapse therefore reflects row geometry: at the active-distractor counts and relative margins on Llama, the relu^{16} envelope sharpens too aggressively for the answer-bearing heads to retain $S_p < 1$ on retrieval-bearing rows. The behavioral failure is consistent with the row-level theory once one accounts for which rows actually carry the behaviorally relevant retrieval signal (Proposition N.8).

AO.3. Matched diagnostic on RULER-Dwarkesh prompts

We compute per-row S_p on the same RULER-Dwarkesh prompts using the diagnostic of Section 2. Table 41 reports the all-head unsafe-row fraction at the model-specific b_{auto}.

Table 41. 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 40). 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 X and AJ).

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

AO.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 42 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 %),

Table 42. Multi-key RULER-template probe with custom post-cutoff Dwarkesh haystack (4 keys, $n = 50$, ~ 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

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 4 and Appendices X and AJ) that retrieval is 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 $n = 50$. The $p = 16$ collapse on the Llama family is resolved as geometric in Appendix O.

AP. Repair details

This appendix records the LoRA-repair training details, per-cell repair numbers, and the objective-specific intervention analysis demoted from the body Appendix AP.

Training setup. The audit so far is zero-shot. A natural follow-up is whether lightweight trained repair can recover what the substitution loses. This section should be read against LoLCATs/Liger/ReGLA: low-rank adaptation after attention conversion is not novel as a recipe. We use it here only as a diagnostic stress test for irreparability and preservation. We test on relu_p4_b0 (Llama drops −95.5 pp on NIAH zero-shot, Qwen3 −100 pp): replace softmax with the log-space-stable $\mathrm{ReLU}^4(z)/\sum \mathrm{ReLU}^4(z)$ map at every layer; train LoRA adapters of rank $r = 8$ ($\alpha = 16$, dropout 0.05) on q_proj, k_proj, o_proj with the polynomial attention active in both directions; LM loss on PG-19 (Rae et al., 2020) until the corpus is exhausted (~ 10 M tokens). Optimizer AdamW, learning rate 5×10^{-5} with linear warmup over the first 50 steps then constant, batch size 1 sequence per step with gradient accumulation 8 steps (effective batch 8), sequence length 2048, bf16 mixed-precision, ~ 10 min on a single A6000. Seeds: $\{0, 1, 2\}$ on Qwen3 single-key (reported in Table 45); a single seed 0 on Llama (the recovery is sharp enough that the 0/5M/10M-token intermediate checkpoints already saturate, and we did not deepen seed variance). PG-19 is likely train-seen for modern LMs and is therefore a poor source for clean repair claims; we keep these runs only as a preservation and sanity audit, and use 2026 post-cutoff retrieval suites for source-shift checks. Trainable parameters: 1.4 M (Llama, 0.11 %) / 2.5 M (Qwen3, 0.15 %). Table 43 reports NIAH accuracy at pad = 128 in three states.

Table 43. Trained-repair NIAH at pad = 128, $n = 200$ prompts. Each row reports softmax baseline, zero-shot relu_p4_b0 (no training), and the same polynomial recipe after LoRA repair on the PG-19 corpus (~ 10 M tokens until exhausted, ~ 10 min on a single A6000). Both audited models recover the polynomial trained-task cell from a recipe that the zero-shot audit reported as completely collapsing. This table does not claim that the LoRA preserves softmax behavior under all evaluations; Table 44 shows the bounded-repair failure modes. (Gemma-4 is omitted: the multimodal Gemma-4 attention replaces standard nn.Linear with a custom Gemma4ClippableLinear module that the PEFT LoRA wrapper does not recognize; we leave a clean Gemma repair to future work.)

Model	Softmax	Zero-shot	LoRA-repaired
Llama-3.2-1B	**100.0**	2.5	**100.0**
Qwen3-1.7B	**100.0**	0.0	**100.0**

Reading the table. Both Llama (lower-risk cluster, 19.6 % unsafe rows at $p = 2$) and Qwen3 (higher-risk cluster, 94 % unsafe rows at $p = 2$) recover the polynomial trained-task cell to the softmax baseline (100 % NIAH) after ~ 10 M tokens of LoRA training, despite both collapsing to ≤ 2.5 % zero-shot. The 5 M-token intermediate checkpoint already reaches 100 % on both models, so the recovery is fast and not at the budget margin. The diagnostic's no-go signal predicts zero-shot failure, not geometric irreparability on the matched training distribution. A small set of ≤ 2.5 M trainable parameters (< 0.15 % of base parameters) is sufficient to bridge the polynomial-substitution gap on both lower-risk and higher-risk models, with a training budget that is $< 10^{-5}$ of pretraining. This reframes the contribution of S_p: it remains a conservative negative screen for zero-shot substitution (every cell with ≤ 3 % measured-safe in the audit collapsed −100 pp behaviorally; Table 10), but the same cells can be recovered on the trained task with lightweight adaptation. The practical implication: S_p tells a deployer "you cannot drop in this substitute as-is," and the repair experiments add a second requirement: verify that the

LoRA-modified model still preserves the softmax behaviors that matter outside the training distribution.

Robustness of the repair: bounded recovery. The training-task table (Table 43) shows full recovery on the suite the LoRA was trained on. A critical reader will ask whether the recovery extends outside the training task. We first characterize a PG-19 language-modeling repair setting in Table 44 (full data in Appendix AQ): (i) harder retrieval at the same task type (RULER niah_single_3 Dwarkesh, $L = 4\,\mathrm{K}$, single-key UUIDs); (ii) harder retrieval at multi-key (RULER multi-key Dwarkesh, $L = 4\,\mathrm{K}$, four needles per prompt); (iii) the LoRA-active model evaluated under softmax (no polynomial substitution); (iv) asymmetry across models. The headline finding is that generic PG-19 repair has bounded recovery, and the bounds are model-dependent: Llama generalizes gracefully on single-key retrieval, while the original Qwen3 PG-19 adapter is unsafe off the matched training cell.

Table 44. Bounded recovery of PG-19 trained repair. Each cell reports %-correct retrieval. Notation: ⇊ means the LoRA degrades the underlying model (softmax-with-LoRA is worse than softmax-without-LoRA). Llama-1B's PG-19 LoRA generalizes across context length (RULER niah L4k softmax: 82 → 93); the original Qwen3 PG-19 LoRA is unsafe outside the matched training cell. This table motivates the task-matched repair experiments in Table 45.

Suite (model)	sm	zs	sm+L	zs+L
Llama-3.2-1B				
NIAH pad=128 ($n=200$, train)	100	2.5	100	**100**
RULER niah_s3 L4k ($n=100$)	82	0	**93**	53
RULER multi-key ($n=50$)	64	0	46⇊	30
Qwen3-1.7B				
NIAH pad=128 ($n=200$, train)	100	0	91↓	**100**
RULER niah_s3 L4k ($n=100$)	100	0	0⇊	0
RULER multi-key ($n=50$)	96	0	0⇊	0

Columns: sm = softmax baseline; zs = zero-shot relu_p4_b0; sm+L = softmax with LoRA active (no substitution); zs+L = relu_p4_b0 with LoRA active. The Qwen3 trained-task sm+L value is $182/200 = 91\,\%$, so even the easy NIAH suite shows a softmax-preservation warning. ↓ marks mild LoRA degradation; ⇊ marks catastrophic degradation.

The pattern is precise. (i) On the trained task, both models recover the polynomial cell to softmax baseline (100 %), but Qwen3 already shows mild softmax degradation when the same adapter is left active without substitution (100 % → 91 %). (ii) Llama's PG-19 LoRA generalizes on a harder same-task probe (RULER niah L4k): softmax+LoRA at 93 % is 11 pp above the no-LoRA softmax baseline (82 %), and polynomial+LoRA recovers to 53 % from a 0 % zero-shot floor. (iii) The original Qwen3 PG-19 LoRA is unsafe off the matched training cell, reaching 0 % on both RULER single-key and multikey. (iv) Even Llama's LoRA degrades multi-key softmax (from 64 % down to 46 %), so the gentle Llama generalization is itself bounded by task type as well as context length.

Task-matched Qwen3 repair changes the boundary. The PG-19 result above does not imply that Qwen3 is irreparable. We therefore train Qwen3 adapters directly on synthetic retrieval objectives. Table 45 separates the key comparison: a single-key adapter repairs single-key retrieval but fails multikey transfer, while a multikey-trained adapter substantially repairs both multikey retrieval and single-key preservation, including balanced and no-Dwarkesh 2026 post-cutoff source audits.

Language-modeling preservation is currently negative. This is the most important repair result. A critical repair claim must ask whether the adapter preserves ordinary language modeling under the substituted attention. Table 46 gives the current PG-19 validation audit for Qwen3 adapters. PG-19 is likely train-seen by 2026-era models, so we use it only as a preservation stress test, not as clean benchmark evidence. The result is unfavorable: all repaired polynomial models remain far worse than the base softmax model, and the single-key retrieval adapter worsens polynomial perplexity relative to the unrepaired polynomial baseline. The adapters preserve softmax perplexity when run without substitution, so the failure is specific to polynomial-attention language modeling rather than generic LoRA damage. Operationally, this argues for full retraining, attention-transfer, or feature-map engineering rather than claiming that narrow LoRA repair converts a pretrained LM.

Objective-specific intervention. The lesson is not that LoRA "solves" polynomial substitution. The lesson is that repair is an objective-specific intervention. A repair objective that matches the intended retrieval workload can recover substantial behavior under polynomial attention; a mismatched objective can fail badly even when it looks perfect on its training cell; and even successful retrieval repair can fail preservation. This is a stronger and more useful conclusion than the original binary story: S_p is a conservative zero-shot screen; trained repair can close the target-task gap; and the repaired model must then pass transfer, source-robustness, model-breadth, and language-modeling preservation audits.

Table 45. Task-matched Qwen3 repair transfer. Exact-match retrieval under softmax and relu_p4_b0 with the indicated LoRA active. "Balanced" suites use 2026 post-cutoff sources with Dwarkesh as one equal source slot (6.64–6.66% of context); "no-Dwarkesh" removes it entirely. The single-key adapter is task-locked on multikey transfer, while the multikey adapter substantially tracks the softmax ceiling across source shifts.

Adapter / suite	Source mix	Softmax+LoRA	Poly+LoRA	Reading
single-key adapter, train L2k	Dwarkesh	199/200	200/200	matched training recovery
single-key adapter, hard single L4k	Dwarkesh	100/100	99/100	same-task transfer
single-key adapter, hard multikey L4k	Dwarkesh	49/50	6/50	task transfer fails
multikey adapter, train L2k	Dwarkesh	196/200	199/200	matched multikey recovery
multikey adapter, hard multikey L4k	Dwarkesh	46/50	47/50	indistinguishable at $n=50$
multikey adapter, hard single L4k	Dwarkesh	100/100	99/100	preserves single-key
multikey adapter, single L4k	balanced 2026 post-cutoff	100/100	93/100	source transfer, imperfect
multikey adapter, multikey L4k	balanced 2026 post-cutoff	48/50	46/50	indistinguishable at $n=50$
multikey adapter, single L4k	no-Dwarkesh	99/100	95/100	Dwarkesh not required
multikey adapter, multikey L4k	no-Dwarkesh	49/50	47/50	indistinguishable at $n=50$

Table 46. Qwen3 PG-19 preservation audit. Perplexity on 262k PG-19 validation tokens. "Base poly" is the unrepaired relu_p4_b0 model. Retrieval adapters preserve softmax PPL but do not preserve language modeling under polynomial attention.

Adapter	Base sm	Base poly	LoRA sm	LoRA poly
single-key seed 0	17.8	1.18×10^4	17.9	9.73×10^5
single-key seed 1	17.8	1.18×10^4	18.1	2.71×10^4
KL 0.02 single-key	17.8	1.18×10^4	17.7	6.21×10^5
multikey adapter	17.8	1.18×10^4	18.3	1.97×10^4

AQ. Robustness of trained repair

This appendix records the remaining robustness axes for trained repair. The current evidence is deliberately narrow: PG-19-trained LoRA can restore the matched PG-19 NIAH diagnostic, and task-matched RULER training can restore harder UUID retrieval, but we do not yet claim that a single repaired model is a general softmax replacement. A critical reader should still ask about rank sensitivity, substitution-recipe transfer, cross-family coverage, and preservation of general language-modeling behavior. These are left for future work rather than treated as completed claims in this paper.

AQ.1. Rank ablation

The original repair plan included LoRA ranks $r \in \{2, 4, 8\}$ for Llama-1B and Qwen3-1.7B under relu_p4_b0. We now treat this as an efficiency audit, not as part of the central claim. The body result already shows that a modest $r = 8$ adapter can recover matched diagnostics; lower-rank results would determine how cheap that repair can be, but cannot by themselves answer whether repair transfers across tasks.

AQ.2. Recipe transfer

Recipe transfer remains an open safety question. The strongest version would train under one substitution, such as relu_p4_b0, and evaluate under more aggressive or qualitatively different substitutions such as relu_p2_b0, relu_p8_bauto, and sigmoid gating. Without those cells, this paper does not claim that the adapter learns a substitution-agnostic repair; the supported claim is task-matched recovery for the substitution used during training.

AQ.3. Cross-family + harder retrieval

Cross-family repair is also still an audit requirement. OLMo-2-0425-1B remains useful because it is an open 2025 model in the unsafe diagnostic cluster. Pythia-1.4B, though dated by 2026 standards in absolute capability, occupies a privileged empirical role in the protected-set taxonomy: it is the only confirmed Regime A″ (reversed-ρ, alignment) cell, where top-$|c_*|$ heads are causally essential (protected-set intervention: keeping only top-80 softmax rescues retrieval to 94 %; Table 3) and structural detectors $|c_*|$/Wu correctly identify the causally important protected set. This regime contrast is what licenses the falsification claim — without an A″ counterexample, "structural detectors miss the compatible protected set" could be a ρ-sign artifact rather than a regime-dependent phenomenon. Newer families (Llama-3.x, Qwen3, Granite, Mistral, OLMo-2) carry the modern-scale claims; Pythia carries the regime-taxonomy claim that no current modern model in our panel replicates.

AQ.4. Harder retrieval (RULER niah_single_3 dwarkesh)

The harder UUID-retrieval suite is part of the main evidence rather than an appendix-only diagnostic. Ta-

ble 45 reports that single-key training recovers single-key retrieval but fails multi-key transfer, while multi-key training recovers both Dwarkesh and balanced/no-Dwarkesh 2026 post-cutoff source audits. That result is the main reason the paper direction shifted from "universal repair" toward "diagnose collapse, then train and audit task-matched repair."

AR. Diagnostic bake-off: full table and permutation null

Table 47. Diagnostic bake-off, NIAH pad=128, balanced cells. Per-prompt ROC-AUC for predicting collapse under zero-shot relu_pp_b0. We restrict to (model, recipe) cells with ≥ 5 collapses and ≥ 5 survivors so within-cell ROC is well-defined; this yields six cells spanning two recipes and three model families. Sign convention: higher predictor value $\Rightarrow$ higher collapse probability. "Wins" counts cells where the predictor achieves ROC-AUC ≥ 0.55 in the within-cell direction.

Predictor	Mistral$_{p=4}$	Mistral$_{p=2}$	Llama-3B$_{p=2}$	Granite$_{p=4}$	Granite$_{p=2}$	Llama-1B$_{p=2}$	median	wins
Polynomial-specific (this work)								
S_p median log	**0.649**	0.482	**0.616**	0.418	0.399	0.482	0.482	2/6
S_p unsafe fraction	**0.607**	**0.556**	0.416	0.464	0.489	0.471	0.480	2/6
Generic concentration baselines								
Hedgehog spikiness	0.483	0.456	0.436	0.506	0.481	0.484	0.482	0/6
Pre-softmax logit margin	**0.684**	**0.553**	**0.592**	0.371	0.350	0.474	0.514	3/6
Top-1 softmax mass	0.521	0.490	0.405	0.474	0.474	0.529	0.482	0/6
Attention entropy	0.465	0.452	0.314	0.500	0.455	**0.560**	0.460	1/6
Median target rank	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0/6

Table 48 reports the within-cell permutation null for every (cell, predictor) combination of the Table 47 bake-off. For each cell, 1000 random shuffles of the binary collapse/survive label are applied; for each predictor we recompute ROC-AUC (sign-corrected) on the shuffled labels and report the one-sided p-value (fraction of null ROCs $\geq$ observed). Bonferroni correction is applied across the full 54-test grid ($\alpha = 0.05/54 \approx 9.3 \times 10^{-4}$).

Three observations:

- Only one cell–predictor combination survives Bonferroni: pre-softmax logit margin on Mistral-7B-Instruct $p = 4$ (observed ROC 0.684, $p_{\text{perm,one}} < 0.001$). The corresponding S_p median log test on the same cell is per-cell significant ($p = 0.003$) but not Bonferroni; S_p unsafe fraction reaches per-cell significance at $p = 0.027$.
- All three Llama-3.2-3B-Instruct $p = 2$ values that initially appeared as "wins" (logit margin 0.59, S_p median 0.62, S_p unsafe 0.42) sit inside the per-cell permutation null band. The Llama-3B$_{p=2}$ cell is severely imbalanced (87 collapses vs 13 survivors) so the null distribution is wide.
- Reversals on Granite at $p = 2$ (logit margin ROC 0.35, $p = 0.99$) show the predictor is statistically significant in the opposite direction on that cell — consistent with the main text's observation that Granite's row-level diagnostics decouple from per-prompt collapse.

The honest reading: the bake-off rules out the strong claim that Hedgehog spikiness or entropy are uniformly better diagnostics within-cell (every Hedgehog cell sits inside null), but does not establish S_p or pre-softmax logit margin as universal cell-conditional predictors. The single Bonferroni-robust signal supports Mistral$_{p=4}$ only.

Cell	Predictor	Obs ROC	Null 5%	Null 95%	p_{one}	n_{coll}	n_{surv}
Llama-3.2-1B $p=2$ ($n=200$, 39/161).							
	S_p unsafe frac	0.471	0.415	0.584	0.71	39	161
	S_p median log	0.482	0.413	0.585	0.64	39	161
	entropy	0.560	0.418	0.589	0.13	39	161
	top-1	0.529	0.419	0.587	0.29	39	161
	top-5	0.558	0.415	0.586	0.14	39	161
	top-20	0.568	0.414	0.583	0.09	39	161
	Hedgehog spikiness	0.484	0.411	0.584	0.59	39	161
	logit margin	0.474	0.417	0.585	0.70	39	161
Llama-3.2-3B-Instruct $p=2$ ($n=100$, 87/13).							
	S_p unsafe frac	0.416	0.362	0.633	0.85	87	13
	S_p median log	0.616	0.365	0.650	0.09	87	13
	entropy	0.314	0.358	0.638	0.99	87	13
	top-5	0.279	0.360	0.632	1.00	87	13
	Hedgehog spikiness	0.436	0.358	0.636	0.77	87	13
	logit margin	0.592	0.356	0.646	0.15	87	13
Mistral-7B-Instruct-v0.3 $p=4$ ($n=200$, 169/31).							
	S_p unsafe frac	0.607	0.406	0.591	0.027	169	31
	S_p median log	0.649	0.415	0.588	0.003	169	31
	entropy	0.465	0.414	0.595	0.75	169	31
	top-5	0.459	0.408	0.595	0.78	169	31
	Hedgehog spikiness	0.483	0.411	0.593	0.62	169	31
	logit margin	**0.684**	0.403	0.595	$< 10^{-3}$	169	31
Mistral-7B-Instruct-v0.3 $p=2$ ($n=200$, 110/90).							
	S_p unsafe frac	0.556	0.432	0.575	0.11	110	90
	S_p median log	0.482	0.427	0.566	0.65	110	90
	entropy	0.452	0.434	0.570	0.90	110	90
	Hedgehog spikiness	0.456	0.432	0.567	0.85	110	90
	logit margin	0.553	0.428	0.567	0.10	110	90
Granite-3.3-8B-Instruct $p=4$ ($n=100$, 81/19).							
	S_p unsafe frac	0.464	0.377	0.622	0.66	81	19
	S_p median log	0.418	0.381	0.615	0.86	81	19
	entropy	0.500	0.377	0.624	0.49	81	19
	Hedgehog spikiness	0.506	0.384	0.620	0.46	81	19
	logit margin	0.371	0.383	0.620	0.96	81	19
Granite-3.3-8B-Instruct $p=2$ ($n=100$, 68/32).							
	S_p unsafe frac	0.489	0.396	0.599	0.56	68	32
	S_p median log	0.399	0.392	0.603	0.94	68	32
	entropy	0.455	0.396	0.603	0.76	68	32
	Hedgehog spikiness	0.481	0.403	0.598	0.62	68	32
	logit margin	0.350	0.401	0.601	0.99	68	32

Table 48. Permutation null for Table 47. For each of 6 balanced cells × 9 predictors = 54 tests we report observed within-cell ROC-AUC (sign-corrected), 5 %/95 % percentiles of the $n=1000$ permutation null, and one-sided p-value (fraction of null $\geq$ observed). Bold = Bonferroni-significant at $\alpha=0.05/54$; italic = per-cell $p<0.05$ but not Bonferroni. Only logit margin on $\text{Mistral}_{p=4}$ survives Bonferroni; two other $\text{Mistral}_{p=4}$ predictors are per-cell significant. Rows for the constant target-rank predictor and redundant top-K duplicates are omitted for compactness.

AS. Wu vs $|c_*|$ structural-detector agreement (full)

The Wu-protocol retrieval-head identification of Wu et al. (2024) captures, for each (layer, head), the mean softmax weight from the final query position to the needle position across 100 NIAH pad=128 prompts. We compare these per-head scores to the per-head mean $|c_*|$ from Ψ_p at $p=4$, $b=0$. Table 49 summarizes across the five models for which we have both captures.

The Granite case (zero Wu $\cap$ bottom-$|c_*|$ overlap up to $K=160$, Wu top-80 mean rank 1002.5 in a 1280-head model) is the cleanest single-model evidence that two mechanistically distinct structural detectors can converge on the same population. The Mistral causal protected-set test (Tab. 3) shows that this co-identified population is functionally peripheral for polynomial substitution, which is the structural-vs-functional dissociation. The Gemma-2-2B reversal had been a candi-

Model	$\rho(\text{Wu}, \lvert c_*\rvert)$	p	$K=80$ overlap top-$\lvert c_*\rvert$	overlap bottom-$\lvert c_*\rvert$	Wu top-80 mean rank	median
Mistral-7B-Inst	+0.517	4×10^{-71}	25 / 6.2 exp	2	750.7	512.5
Llama-3.2-3B-Inst	+0.449	1.4×10^{-34}	29 / 9.5 exp	6	483.2	336.0
Pythia-1.4B	+0.389	2.7×10^{-15}	33 / 16.7 exp	4	259.3	192.0
Granite-3.3-8B	+0.321	4.2×10^{-32}	28 / 5.0 exp	0	1002.5	640.0
Gemma-2-2B-it	−0.145	0.037	26 / 30.8 exp	35	97.7	104.0

Table 49. Wu vs $|c_*|$ structural-detector agreement across five models, niah_pad=128. Spearman ρ between Wu mean attention-at-needle and per-head mean $|c_*|$. "$K=80$ overlap top" is the size of Wu-top-80 $\cap$ $|c_*|$-top-80 versus the random expectation K^2/H. "Wu top-80 mean rank" is the average rank of Wu's top-80 heads within the full $|c_*|$ distribution (median across all heads $=H/2$). Four models show strong positive agreement with Wu top-K concentrated in high-$|c_*|$ heads; Gemma-2-2B reverses, with Wu top-80 skewing toward low-$|c_*|$. The single-model reversal demonstrates that "structural retrieval head" is a detector-dependent property across architectures.

date test for which detector (if either) aligns with function, but the Wu-ranked protected-set intervention on Gemma-2-2B (Table 50) reveals the test is degenerate: Gemma-2-2B is polynomial-impervious (no failure mode under all-polynomial substitution at the easy NIAH cell), so the protected-set intervention shows uniform $\geq 93\,\%$ retention on every variant, and the structural-vs-functional question has no traction.

AT. Gemma-2-2B-it protected-set intervention (substitution-robust regime)

Table 50 reports both the $|c_*|$-ranked and Wu-ranked protected-set intervention on Gemma-2-2B-it. The two-row block emphasizes that Gemma-2-2B-it is functionally distinct from Mistral, Granite, and Llama: polynomial substitution does not fail on the easy NIAH cell (all-polynomial accuracy 96 %), so no protected-set intervention reveals a functional dissociation. We include this null cell to demarcate the regime: the structural-vs-functional dissociation finding applies only where polynomial substitution actually fails, which is 12/14 checkpoints in our core 14-checkpoint substitution sweep (Table 7) but not the Gemma-2-it variants.

K	top-K poly	all-except-top-K poly	bottom-K poly	all-except-bottom-K poly	random-K poly	all-except-random-K poly
Gemma-2-2B-it Wu-ranked, 208 heads, softmax 100 %, all-polynomial 96 %.						
20	100 %	93 %	100 %	97 %	100 %	95 %
40	100 %	98 %	100 %	100 %	100 %	100 %
80	100 %	100 %	100 %	100 %	100 %	100 %
160	100 %	100 %	100 %	100 %	100 %	100 %
Gemma-2-2B-it $\lvert c_*\rvert$-ranked, 208 heads, softmax 100 %, all-polynomial 96 %.						
20	100 %	100 %	100 %	93 %	100 %	99.7 %
40	100 %	100 %	100 %	98 %	100 %	91.7 %
80	100 %	100 %	100 %	100 %	100 %	100 %
160	90 %	100 %	100 %	100 %	100 %	100 %
Gemma-2-9B-it $\lvert c_*\rvert$-ranked, 672 heads, softmax 100 %, all-polynomial 99 %.						
40	100 %	99 %	100 %	99 %	100 %	98.3 %
80	100 %	100 %	100 %	99 %	100 %	98.7 %
160	100 %	97 %	100 %	100 %	100 %	99 %
320	100 %	100 %	100 %	100 %	100 %	100 %

Table 50. Gemma family protected-set intervention, NIAH pad=128, $p = 4$, $b = 0$, $n = 100$ prompts. Every variant retains $\geq 90\,\%$ across both Gemma-2-2B (all-polynomial 96 %) and Gemma-2-9B (all-polynomial 99 %). Both ranking sources give the same null answer because there is no functional failure to dissociate from. Gemma-2 is in a substitution-robust regime at both scales tested; structural-vs-functional retrieval-head dissociation cannot be tested in this regime.

AU. Wu-ranked vs $|c_*|$-ranked Mistral protected set: detector sharpness

The Wu-vs-$|c_*|$ correlation $\rho = +0.52$ on Mistral-7B-Instruct establishes statistical agreement on the rank of heads. To test whether the two rankings carve out the same functional protected set, we run the protected-set intervention with both ranking sources on the same model under identical $p = 4$, $b = 0$ conditions. Table 51 reports both side by side.

The dissociation direction is preserved across both rankings: top-K poly is mostly harmless and all-except-top-K poly is catastrophic, in both. But the sharpness differs substantially. With $|c_*|$ ranking, keeping only bottom-K softmax (all-except-bottom-K poly) recovers 87–95 % across K; with Wu ranking, the same operation recovers only 24–88 % across K. The complement (all-except-top-K poly) is also stricter for $|c_*|$ (1–6 %) than for Wu (4–55 %). Wu's bottom-K protected set contains some causally important heads that $|c_*|$'s bottom-K excludes; conversely Wu's top-K retains some functional content that $|c_*|$'s top-K does not.

This is not an inconsistency with the Wu corroboration: the rank correlation says "the two detectors mostly agree on who looks like a retrieval head," and the protected set overlap numbers say "the top-K populations have ~ 1.7–$4\times$ the random overlap." But the functional consequences of substitution depend on the exact head identity at the protected set boundary, and the boundary differs by ranking. The takeaway: $|c_*|$ (task-specific, derived from Ψ_p) is a sharper structural localizer for the polynomial-substitution dissociation than Wu (generic, attention-only). Wu remains the right tool for confirming that the dissociation isn't a $|c_*|$-derivation artifact, but the sharpest causal evidence comes from $|c_*|$.

K	top-K poly	all-except-top-K poly	bottom-K poly	all-except-bottom-K poly
$\lvert c_*\rvert$-ranked Mistral, softmax 95 %, all-polynomial 14 %.				
40	95 %	**3** %	95 %	**87** %
80	95 %	**1** %	82 %	**93** %
160	95 %	**3** %	90 %	**91** %
320	95 %	**6** %	55 %	**95** %
Wu-ranked Mistral, softmax 95 %, all-polynomial 14 %.				
40	95 %	12 %	95 %	24 %
80	95 %	4 %	95 %	41 %
160	95 %	14 %	95 %	80 %
320	78 %	55 %	77 %	88 %

Table 51. Wu-ranked vs $|c_*|$-ranked protected-set intervention on Mistral-7B-Instruct, NIAH pad=128, $p = 4$, $b = 0$, $n = 100$ prompts. Same model, same intervention, two different head rankings (Spearman $\rho(\text{Wu}, |c_*|) = +0.52$). The dissociation direction holds in both; the magnitude is sharper under $|c_*|$ ranking (all-except-bottom-K recovers 87–95 % vs Wu's 24–88 %).

AV. Phase diagram: ρ across (model, suite, p) cells

The layer-permutation phase summary reports the layer-permutation-tested Spearman $\rho(\text{flip rate}, |c_*|)$ across 108 cells: 14 models, 2 suites (niah pad=128 and RULER niah_dwarkesh), and 4 polynomial degrees $p \in \{2, 4, 8, 16\}$. The full per-cell breakdown is in the supplementary artifact (phase_diagram_table.tsv). We classify regime as A ($\rho \leq -0.30$ and permutation-significant), A" ($\rho \geq +0.30$ and permutation-

significant), A' ($0.15 \le |\rho| < 0.30$), and B ($|\rho| < 0.15$ or not permutation-significant).

Three cell-level findings worth highlighting:

- Granite-3.3-8B RULER: ρ flips from $+0.41$ at $p=2$ (A") to -0.42 at $p=8$ (A) within the same (model, suite). This is a recipe-driven regime reversal — the same head ranking inverts which heads survive substitution as the polynomial degree changes.

- Phi-4-mini-instruct: $\rho \approx 0$ on niah pad=128 across all p (Regime B-ish), $\rho = +0.55$ on RULER at $p=2$ (Regime A"). This model occupies opposite regimes by suite.

- Gemma-2-2B-it: $\rho \approx -0.39$ on niah pad=128 (A-direction at low $|\rho|$), $\rho \approx 0$ on RULER (B) at $p \ge 4$. Suite-driven regime change.

10 within-model regime transitions across 5 distinct models (Llama-3.2-3B-Instruct, Mistral-7B-Instruct-v0.3, Phi-4-mini-instruct, gemma-2-2b-it, granite-3.3-8b) make explicit that regime is a (model, suite, recipe)-cell property, not a per-model attribute. The driver, per Corollary 2.3, is the (model, suite, p) cell's position relative to the worst-case sufficient degree $p^\star \asymp \log A / \log(1/(1-\rho))$: harder suites have larger A and higher required $p^\star$, pushing substitution closer to its concentration boundary and tilting cells from Regime A toward Regime A" as p grows.

AW. Per-cell Ψ_p correlation table and p-sweep table

Table 52. Headline Ψ_p cells. We tested 123 (model, suite, p) cells across 11 models, 2 suites, and 4 recipes. Each cell's p-value is from a 1000-shuffle layer-permutation null on per-layer flip-rate versus $|c_*|$. Aggregate: 45/123 permutation-significant, 7 Bonferroni. The full table is in the supplementary artifact (psip_permutation_null.parquet).

Cell (model, suite, p)	n_{lay}	ρ	p_{perm}	note
Bonferroni-significant supports: 7 cells				
Gemma-2-2b base, niah pad=128 $p=8$	26	-0.58	$< 10^{-3}$	–
Qwen-2.5-3B base, niah pad=128 $p=4$	36	-0.57	$< 10^{-3}$	–
Mistral-7B base, niah pad=128 $p=4$	32	-0.56	$< 10^{-3}$	–
Granite-8B-Inst, niah pad=128 $p=16$	40	-0.55	$< 10^{-3}$	–
Mistral-7B-Inst, niah pad=128 $p=2$	32	-0.55	$< 10^{-3}$	–
Granite-8B-Inst, niah pad=128 $p=8$	40	-0.51	$< 10^{-3}$	–
Qwen-2.5-3B base, niah pad=128 $p=8$	36	-0.51	$< 10^{-3}$	–
Significant reversals: 4 cells				
Pythia-1.4b, niah pad=128 (all p)	24	$+0.44$–$+0.46$	0.017–0.020	all p
Phi-4-mini-Inst, RULER $p=2$	32	$+0.55$	0.002	–
Phi-4-mini-Inst, RULER $p=4$	32	$+0.35$	0.045	–
Granite-8B-Inst, RULER $p=2$	40	$+0.41$	0.005	within-model flip

AX. Greedy minimum sufficient compatible protected-set (MSC) search

Terminology and certification status (global). "Minimum sufficient" here is in the algorithmic sense: the greedy procedure halts at the smallest protected head set the descent can find. Some artifact names still say cohort for historical reasons; in prose, the object is an operator-specific protected head set. Promoting it to a global minimum requires excluding all smaller singleton sets and lower-bounding the candidate pair. The certification proposition is: if no singleton clears the θ-floor and our two-head set does, then the two-head set is a minimum θ-sufficient compatible protected set (Proposition AX.1 below; family-wise version in Proposition AX.3). The table below summarizes the hardening tests. The all-head random null and the layer-0-conditioned null answer different questions; the latter is intentionally harder because the discovered pair lies in an enriched layer-0 pool. All evaluated smaller/plausible competitors fail the 72.2 % floor. The full exhaustive 1024-singleton sweep at $n = 50$ excludes 1023/1024 singletons directly under the family-wise Bonferroni bound $\alpha/H_{\text{tot}} \approx 4.9 \times 10^{-5}$; the lone above-threshold singleton $(0, 17)$ was retested at $n = 500$ and excluded with $U_\beta = 0.628$. Thus singleton exclusion is established at family-wise $\alpha = 0.05$; the joint singleton-exclusion plus pair-sufficiency certificate below is a 90% joint certificate ($\alpha_1 + \alpha_2 = 0.10$) at the matched $n = 500$ floor. The protected-set identity is not unique within layer 0, as shown below.

Mistral MSC hardening

Table 53. Mistral MSC hardening. NIAH pad=128, $p=4$, $b=0$. The greedy pair clears the $\theta = 0.76 \cdot \text{Acc}_{\text{sm}}$ floor; evaluated singletons and same-KV competitors fail. The text gives the Bonferroni singleton screen and held-out retests.

Test	Result	n	Reading
Held-out MSC $\{(0,17),(0,28)\}$	74.2% vs suffmax 91.6%	500	clean (substantial recovery)
All-polynomial baseline	12.6%	500	collapse floor
Layer-0-conditioned random-2 null	max 17%, $p_{\text{emp}}=0.0020$	200	beats local null
Layer-0-conditioned random-2 null	max 42 %	500	harder local null; layer 0 is enriched
99-seed random-2 null	max 43%, mean 14.9%, $p=0.0100$	200	MSC 76.5%, 0/99 random $\geq$ protected set
99-seed random-2 null	MSC 61 %, max 34 %, $p=0.01$	100	strengthened null rejection
Singleton $\{(0,17)\}$	62 % ($n=100$); 54.2 % ($n=500$)	100/500	fails (R_c rank-preserve alone)
Singleton $\{(0,19)\}$	28 % ($n=100$); 33.8 % ($n=500$)	100/500	fails (KV4 within-group)
Singleton $\{(0,28)\}$	29 %	100	fails (R_c sign-canceller alone)
Singleton $\{(1,24)\}$	44.8 % ($U_{\beta}=0.628$)	500	fails (top-4 candidate at L0–10)
Singleton $\{(1,24)\}$	39.4 % ($U_{\beta}=0.43$)	500	fails (top-4 candidate at L0–10)
Same-KV pair $\{(0,17),(0,19)\}$	66 %	100	fails (KV4+KV4 redundancy)
Cross-KV pair $\{(0,17),(0,28)\}$	84 %	100	clears (KV4+KV7 spanning)
Exhaustive layer-0 K=2 pair $\{(0,17),(0,22)\}$	87 % (096-pair best)	100	strongest sufficient pair, beats greedy MSC
Exhaustive layer-0 K=2 pair $\{(0,19),(0,22)\}$	74 % (096-pair third)	100	clears floor without $\{0,17\}$
Exhaustive 1024-singleton family-wise Bonferroni screen	1023/1024 excluded at $n=50$; $\{0,17\}$ retested at $n=500$ ($U_\beta=0.628$)	50/500	all 1024 singletons excluded at family-wise $\alpha=0.05$; pair lower bound gives joint $K=2$ size certificate at $\alpha_1+\alpha_2=0.10$

Singleton certification (global, all 1024 singletons family-wise). The three direct singleton tests on the MSC's own members and the highest-amplification non-member all fail the θ-floor at $n = 100$. The asymmetry $(0, 17) \gg (0, 19) \gg (0, 28)$ is informative: $(0, 17)$ alone carries most of the protected-set signal but is not sufficient, and $(0, 28)$ alone is barely above its random-1 control — consistent with the rank/sign-canceller role of Proposition AX.5.

The exhaustive 1024-singleton Clopper–Pearson screen at $n = 50$ excludes 1023/1024 singletons directly at

the family-wise Bonferroni level $\alpha/H_{\text{tot}} \approx 4.9\times10^{-5}$; only $(0,17)$ ($n = 50$ observed 0.56, $U_\beta = 0.809$) was above the 0.722 floor at $n = 50$ and was retested at $n = 500$ where it observed 54.2% ($U_\beta = 0.628$, below floor). The four highest-observed singletons retest at $n=500$ as $(0,17)=54.2\%$ ($U_\beta=0.628$), $(0,22)=44.8\%$ ($U_\beta = 0.536$), $(0,19) = 33.8\%$ ($U_\beta = 0.424$), $(1,24) = 39.4\%$ ($U_\beta=0.482$). All 1024 singletons therefore certify family-wise singleton exclusion at $\alpha=0.05$ (1020 excluded directly by the $n = 50$ screen, 4 excluded by the $n = 500$ retest). The global $K = 2$ minimum-size statement also requires the pair-sufficiency lower bound given in the joint certificate below.

On Mistral, KV-spanning is necessary among the evaluated high-amplification candidates. The same-KV pair $\{(0,17),(0,19)\}$ (both in KV group 4 under Mistral's GQA: 32 query heads, 8 KV heads, 4 queries per group) achieves 66 % — 4 pp above $(0,17)$ alone and 6 pp below the θ-floor. The cross-KV pair $\{(0,17),(0,28)\}$ (KV groups 4 and 7) clears the floor at both $n=500$ (74.2 %) and $n=100$ (81 %). Adding $(0,28)$ to $(0,17)$ delivers +12 pp; adding $(0,19)$ delivers only +4 pp. We report the KV-spanning observation as Mistral-specific and specific to this protected set among the evaluated candidates; whether the GQA-spanning constraint generalises across families and survives substrate change is left to follow-up.

Protected-set identity is plural, not unique. The exhaustive $\binom{32}{2}=496$-pair layer-0 search at $n=100$ identifies three K=2 protected sets clearing the $\theta = 0.73$ floor: $\{(0,17),(0,22)\}$ at 87% (strongest, KV groups $4+5$), $\{(0,17),(0,28)\}$ at 81% (the greedy MSC, KV groups $4+7$), and $\{(0,19),(0,22)\}$ at 74% (KV groups $4+5$). The greedy descent's stopping point is therefore one of multiple K=2 sufficient protected sets at layer 0, not the unique one. No single head is necessary across all three protected sets: $(0,17)$ appears in two of the three but not in $\{(0,19),(0,22)\}$, and $(0,22)$ appears in two of three but not in $\{(0,17),(0,28)\}$. The K=2 minimum sufficiency claim is supported by the joint certificate below (singleton exclusion plus pair lower bound), but the identity of the K=2 minimum protected set is plural. The fourth-ranked pair $\{(0,17),(0,27)\}$ at 69% is below the floor at $n=100$; whether it clears at $n=500$ is unresolved.

Formal apparatus.

Proposition AX.1 (Certification of a minimum θ-sufficient compatible protected set). Let $\mathcal{H}$ denote the set of all attention heads in a model, and for $H\subseteq\mathcal{H}$ let $\mathrm{Acc}(H)$ denote retrieval accuracy with heads in H kept softmax and heads in $\mathcal{H}\setminus H$ polynomialized. Let $\mathrm{Acc}_{\text{sm}}\equiv\mathrm{Acc}(\mathcal{H})$. Fix $\theta\in(0,1]$ and call H θ-sufficient if $\mathrm{Acc}(H)\geq\theta\,\mathrm{Acc}_{\text{sm}}$. Suppose (i) $\mathrm{Acc}(\varnothing)<\theta\,\mathrm{Acc}_{\text{sm}}$, (ii) for every $h\in\mathcal{H}$, $\mathrm{Acc}(\{h\})<\theta\,\mathrm{Acc}_{\text{sm}}$, and (iii) there exists $H_2=\{h_a,h_b\}$ with $\mathrm{Acc}(H_2)\geq\theta\,\mathrm{Acc}_{\text{sm}}$. Then H_2 is a minimum θ-sufficient compatible protected set.

Proof. H_2 is θ-sufficient by (iii). Any protected set G with $|G|<|H_2|=2$ satisfies either $|G|=0$ (so $G=\varnothing$ and (i) gives $\mathrm{Acc}(G)<\theta\,\mathrm{Acc}_{\text{sm}}$) or $|G|=1$ (so $G=\{h\}$ for some $h\in\mathcal{H}$ and (ii) gives $\mathrm{Acc}(G)<\theta\,\mathrm{Acc}_{\text{sm}}$). In both cases G is not θ-sufficient, so no protected set strictly smaller than H_2 is θ-sufficient. □ □

Threshold sensitivity. For any protected set H define its empirical sufficiency ratio $\theta_H\equiv\mathrm{Acc}(H)/\mathrm{Acc}_{\text{sm}}$. A protected set H is θ-sufficient iff $\theta\leq\theta_H$. A two-head protected set H_2 is therefore certified minimum at threshold θ whenever

$$\max\Big\{\theta_\varnothing,\ \max_{h\in\mathcal{H}}\theta_{\{h\}}\Big\}\ <\ \theta\ \leq\ \theta_{H_2}.$$

The certificate is not tied to a single threshold value; it holds on an interval. The proof follows from Proposition AX.1 by dividing both sides of each (in)equality by $\mathrm{Acc}_{\text{sm}}>0$. On the current $n=100$ Mistral run, the evaluated pair clears $\theta_{H_2}=0.85$ (81/95), the evaluated singletons have $\theta_{\{h\}}\leq0.65$ (62/95), and the same-KV pair has $\theta=0.69$; the certified interval over evaluated protected sets is $(0.69,0.85]$ and the $\theta=0.76$ floor lies firmly inside it. Finite-sample caveat: if the full singleton sweep produces a near-floor singleton, the left endpoint should be replaced by its one-sided binomial upper confidence bound (Lemma AX.2) before claiming global minimality.

Lemma AX.2 (One-sided Clopper–Pearson singleton exclusion). Let a singleton protected set be evaluated on n independent prompts with $x<n$ successes and unknown true success probability q. The one-sided $(1-\alpha)$-confidence upper bound is

$$U_\alpha(x,n)\ =\ \mathrm{BetaInv}(1-\alpha;\,x+1,\,n-x).$$

If $U_\alpha(x,n)<\theta\,\mathrm{Acc}_{\text{sm}}$, then with confidence $1-\alpha$ the singleton's true success rate is below the θ-sufficiency floor, so the singleton is excluded as θ-sufficient.

Proposition AX.3 (Family-wise singleton exclusion with empirical pair sufficiency). Let $|\mathcal{H}|=H_{\text{tot}}$, fix $\theta\in(0,1]$ and $\alpha\in(0,1)$. Suppose (i) $\mathrm{Acc}(\varnothing)<\theta\,\mathrm{Acc}_{\text{sm}}$; (ii) for every singleton $\{h\}\in\mathcal{H}$ evaluated on n_h prompts with x_h successes, the Bonferroni-corrected Clopper–Pearson upper bound satisfies $U_{\alpha/H_{\text{tot}}}(x_h,n_h)<\theta\,\mathrm{Acc}_{\text{sm}}$; (iii) there exists $H_2=\{h_a,h_b\}$ with empirical $\mathrm{Acc}(H_2)\geq\theta\,\mathrm{Acc}_{\text{sm}}$ on a held-out prompt set. Then with family-wise confidence $1-\alpha$, no singleton in $\mathcal{H}$ is θ-sufficient. Together with the empirical pair success in

(iii), this gives an empirical minimum-size-2 protected set certificate; a fully probabilistic certificate for pair sufficiency additionally needs a lower confidence bound on H_2 and is given in Proposition AX.4.

Proof. Let $E_h = \{q_h > U_{\alpha/H_{\text{tot}}}(x_h, n_h)\}$ for each singleton. By one-sided Clopper–Pearson, $\Pr(E_h) \leq \alpha/H_{\text{tot}}$. By the union bound, $\Pr(\bigcup_h E_h) \leq \alpha$. So with probability $\geq 1-\alpha$, every singleton satisfies $q_h \leq U_{\alpha/H_{\text{tot}}}(x_h, n_h) < \theta\,\text{Acc}_{\text{sm}}$, hence no singleton is θ-sufficient. Assumption (i) excludes the empty protected set, and (iii) records the held-out empirical pair success. □ □

Numerical instantiation on Mistral. With $H_{\text{tot}} = 1024$ and $\alpha = 0.05$, per-test $\alpha/H_{\text{tot}} \approx 4.9\times10^{-5}$. The exhaustive 1024-singleton Clopper–Pearson screen at $n = 50$ excludes 1023/1024 singletons under the family-wise Bonferroni U_β bound; only the top singleton $(0, 17)$ (observed $28/50 = 0.56$, $U_\beta = 0.809$) sat above the 0.722 floor at $n = 50$. We retested the four highest-observed singletons at $n = 500$ and observed $(0, 17) = 54.2\%$, $(0, 22) = 44.8\%$, $(0, 19) = 33.8\%$, $(1, 24) = 39.4\%$; the family-wise Clopper–Pearson upper bounds (computed via $\beta = \alpha/H_{\text{tot}}$ on the exact binomial CDF) are 0.628, 0.536, 0.424, 0.482 respectively, all below the 0.722 floor. Combined with the 1020 low-observed singletons that pass the $n = 50$ Bonferroni screen directly (1024 total minus the 4 retested), all 1024 singletons certify family-wise exclusion at $\alpha = 0.05$. The global $K = 2$ minimum-size certificate uses this singleton exclusion plus the pair lower bound in Proposition AX.4.

Proposition AX.4 (Joint singleton-exclusion and pair--sufficiency certificate). Fix $\theta \in (0, 1]$, $\alpha_1, \alpha_2 \in (0, 1)$, and set the retention floor $f \equiv \theta\,\text{Acc}_{\text{sm}}$. Suppose (i) the empty protected set is below the floor deterministically: $\text{Acc}_{\mathcal{O}}(\varnothing) < f$; (ii) for every singleton $\{h\} \in \mathcal{H}$ evaluated on n_h prompts with x_h successes, the Bonferroni-corrected one-sided Clopper–Pearson upper bound satisfies $U_{\alpha_1/H_{\text{tot}}}(x_h, n_h) < f$; (iii) the candidate pair $H_2 = \{h_a, h_b\}$ is evaluated on a held-out prompt set of size n_2 with x_2 successes, giving a one-sided Clopper–Pearson lower bound $L_{\alpha_2}(x_2, n_2) \geq f$. Then with probability at least $1-\alpha_1-\alpha_2$, the empty protected set and all singletons in $\mathcal{H}$ are below the floor and H_2 is f-sufficient. Therefore the minimum sufficient protected set size is 2.

Proof. The empty protected set is excluded deterministically by (i). Let $E_{1,h} = \{q_h > U_{\alpha_1/H_{\text{tot}}}(x_h, n_h)\}$ for each singleton with true success rate q_h. By one-sided Clopper–Pearson, $\Pr(E_{1,h}) \leq \alpha_1/H_{\text{tot}}$; by the union bound, $\Pr(\bigcup_h E_{1,h}) \leq \alpha_1$. Let $E_2 = \{q_2 < L_{\alpha_2}(x_2, n_2)\}$ with true pair success rate q_2; one-sided Clopper–Pearson gives $\Pr(E_2) \leq \alpha_2$. A second union bound gives $\Pr(\bigcup_h E_{1,h} \cup E_2) \leq \alpha_1 + \alpha_2$. On the complement, every singleton has $q_h \leq U_{\alpha_1/H_{\text{tot}}} < f$ (so no singleton is f-sufficient) and $q_2 \geq L_{\alpha_2} \geq f$ (so the pair is f-sufficient). Combined with the deterministic empty-set exclusion (i), the minimum sufficient protected set size is 2. □ □

The Mistral instance: $\alpha_1 = 0.05$ Bonferroni U bounds on all 1024 singletons are below 0.722, and the pair $\{(0, 17), (0, 28)\}$ at $n_2 = 500$ has $L_{0.05} = 0.708 \geq 0.696$ (the matched $n = 500$ floor). The joint certificate at $\alpha_1 + \alpha_2 = 0.10$ gives a 90% joint-confidence minimum-protected-set-size-2 statement.

Both singleton-exclusion versions are now established: the per-test version directly at $n = 50$ (1023/1024 singletons with Wilson 95% UB below 0.69), and the family-wise version through the $n = 500$ retest of the four highest-observed singletons (all four Clopper–Pearson U_β below 0.628).

Statistical note. At $n = 50$, a singleton observed at $x/50 = 62\,\%$ has $U_{0.05} \approx 73.5\,\%$, above the $72.2\,\%$ floor: Lemma AX.2 cannot exclude a near-threshold singleton at $n = 50$, requiring an $n \geq 200$ follow-up. At $n = 100$ the same $62\,\%$ observation has $U_{0.05} \approx 70.1\,\%$ which does exclude. We accordingly treat the 1024-singleton sweep at $n = 50$ as a screening pass and schedule $n = 500$ follow-ups for any singleton whose $U_{0.05}$ overlaps the floor (script: e3_ablation/cp_screen_followup.py).

Greedy search trace and converged protected set

The protected-set analysis establishes that a $|c_*|$-bottom-K softmax-protected head set (rest polynomialized) achieves substantial NIAH retrieval on Mistral ($K = 80$: $94\,\%$ retention vs softmax $96\,\%$). A natural follow-up question is: what is the smallest softmax-protected head set sufficient for retrieval? We implement a greedy knockout algorithm in e3_ablation/find_min_softmax_cohort.py:

1. Initialize the softmax-protected head set H to the bottom-K_{seed} $|c_*|$ heads (which is verified to work at $K = 40$ for Mistral).
2. Each round: for each head $h \in H$, evaluate $\text{Acc}(H \setminus \{h\})$ (polynomial-substitute all other heads). Pick the removal whose resulting protected set has the highest accuracy.
3. If that accuracy $\geq \theta \cdot \text{Acc}_{\text{softmax}}$ (default $\theta = 0.8$), accept the removal; else stop.

4. Output: minimum sufficient compatible protected set $H_{\min}$ and the round-by-round trace.

The CPU-only counterpart based on the $\Omega_p(H) < 1$ direct-compatibility criterion gives a theoretical minimum of $|H_{\min}| = 1$ on both Mistral and Granite (37/1024 Mistral heads and 27/1280 Granite heads have per-head sign-preservation rate $= 1.00$ over the prompt set; greedy minimum-$\Omega_p(H)$ search terminates at size 1 at every threshold). The empirical (retrieval-accuracy-preserving) $H_{\min}$ is expected to be larger because retrieval requires margin in addition to sign.

Converged Mistral protected set. On Mistral-7B-Instruct (NIAH pad=128, $p = 4$, $b = 0$, $n = 20$ search prompts, accept-floor $\theta = 0.76 = 0.8 \cdot \mathrm{Acc}_{\mathrm{softmax}}$), the greedy descent converged to a $|H_{\min}| = 2$ protected set: $\{(\ell = 0, h = 17), (\ell = 0, h = 28)\}$. The trace shows 38 monotonically accepted single-head removals from the bottom-40 $|c_*|$ seed, achieving search-set accuracy 80 % (vs softmax baseline 95 %); the algorithm rejected and halted on the next-best removal, whose resulting accuracy (55 %) fell below the θ floor. Both surviving heads are in layer 0, consistent with the layer-range protected set result that layer 0 alone preserves 94 % retrieval — the MSC refines that 32-head set down to its 2 individually causally important members.

Model	Setting	$\lvert H_{\min}\rvert$	SM acc	MSC acc	All-poly acc	Rand-K max (p)	Heads
Mistral-7B-Inst	search $n=20$	2	95 %	80 %	—	—	(0,17), (0,28)
Mistral-7B-Inst	held-out $n=500$	2	91.6 %	74.2 %	12.6 %	17.0 % ($p=0.048$)	(0,17), (0,28)

Table 54. MSC search converged on Mistral-7B-Instruct (NIAH pad=128, $p=4$, $b=0$, threshold $\theta=0.8\cdot\mathrm{Acc}_{\mathrm{softmax}}=0.76$). Row 1: search-set ($n=20$) accuracy at descent termination. Row 2: held-out validation on $n=500$ disjoint NIAH prompts; the same $\{(0,17),(0,28)\}$ protected set recovers 74.2 % retrieval (81 % of softmax baseline) while polynomializing everything collapses to 12.6 %. Across 20 random $K=2$ control draws the maximum is 17.0 % and the mean is 12.5 % (median 12.8 %); 0/20 random samples match the MSC protected set, giving randomization $p = 1/21 \approx 0.048$. Cross-suite and cross-model extensions are reported separately as scoping or exploratory follow-ups, not as part of this Mistral certificate.

Structural detectors miss the pair. Cross-referencing $(0, 17)$ and $(0, 28)$ against the three structural rankings used in the main text:

- $|c_*|$ ranking (high-to-low, global): rank 1022/1024 for $(0, 17)$ and 1006/1024 for $(0, 28)$ — both in the bottom 2 %. Within layer 0, $(0, 17)$ ranks 30/32 and $(0, 28)$ ranks 14/32.
- Wu-style attention-at-needle (global): rank 597/1024 for $(0, 17)$ and 949/1024 for $(0, 28)$ — neither head is identified as a Wu retrieval head. Within layer 0, $(0, 17)$ ranks 2/32 on Wu and $(0, 28)$ ranks 23/32.
- Theorem D's Ω_p mean (global): rank 5/1024 for $(0, 17)$ (top 0.5 %) and rank 1002/1024 for $(0, 28)$ (bottom 2 %). The two protected-set members sit at opposite extremes of the per-head sign-flip distribution: $(0, 17)$ has Ω_p mean 22.81 and flips on 52 % of prompts (highly vulnerable, top-5 globally), while $(0, 28)$ has Ω_p mean -6.28 and flips on only 20 % (highly stable, bottom-2 % globally).

Two consequences. First, no single structural detector identifies this protected set: $|c_*|$ and Wu rank both heads below the top half (in fact in the bottom 7 % or worse on Wu), and Ω_p would flag $(0, 17)$ as a sign-flip risk but rate $(0, 28)$ as among the most stable heads in the model. A structural reseeding of MSC search from top-K $|c_*|$, top-K Wu, or top-K Ω_p would converge to a different (and larger) protected set, none of which would contain both $(0, 17)$ and $(0, 28)$ together. Second, the protected set is functionally — not structurally — defined. Greedy descent from bottom-$|c_*|$ works only because the seed protected set is in the right neighborhood (early-layer compatibility heads), but the destination is two specific heads that no rank-based heuristic singles out. This is a direct falsification of "find the retrieval heads and protect them" as a recipe for polynomial substitution, in this setting.

BOS-argmax diffuse mass. What do $(0, 17)$ and $(0, 28)$ actually do? Inspecting the per-prompt attention pattern in the existing Ψ_p captures: both heads attend hardest to position 0 (the BOS / start-of-sequence token) on every prompt — their argmax_pos is exactly 0 on 100/100 NIAH pad=128 prompts. But their actual mass on BOS is small ($(0, 17) = 0.42$ %, $(0, 28) = 8.31$ % averaged over $n = 50$ NIAH prompts in a direct attention-pattern measurement): the attention distribution is diffuse, with BOS only weakly dominant. But 22 other layer-0 heads also satisfy argmax_pos$=0$ on every prompt (24/32 layer-0 heads total are BOS-argmax heads), yet only $(0, 17)$ and $(0, 28)$ end up in the MSC. The distinguishing feature is what happens to the head's output when softmax is replaced with polynomial: the ratio $|y^{(h)}_{\mathrm{poly}}|/|y^{(h)}_{\mathrm{sm}}|$ averaged over prompts is $\mathbf{42.4}\times$ for $(0, 17)$, $\mathbf{8.6}\times$ for $(0, 28)$, and a median of only $\mathbf{1.22}\times$ for the remaining 22 layer-0 BOS-argmax heads. These two heads are 7–35× outliers in polynomial amplification.

The Mistral-specific picture is then: under softmax, $(0, 17)$ and $(0, 28)$ are BOS-argmax heads where BOS is the single highest-mass position on every prompt,

but the tail of non-BOS positions still carries $\sim 99.6\,\%$ and $\sim 91.7\,\%$ of the mass respectively, spread thinly across ~ 4000 keys (per-key mass $\sim 10^{-4}$ to 10^{-3}). Under polynomial substitution at $p = 4$, the polynomial map's smaller dynamic range no longer suppresses these per-key tails so steeply, and the head's value contributions at each non-BOS position get aggregated into a $42\times/9\times$ amplification of $|y|$. These two heads are not the Wu-style retrieval heads; they are compatibility heads whose diffuse-tail-plus-BOS-argmax pattern is unusually distorted by polynomial substitution. Protecting them preserves much of the layer-0 softmax behavior needed for this NIAH cell. We therefore present anomalous polynomial amplification of BOS-argmax diffuse heads as a Mistral-family mechanism candidate, not as a complete downstream circuit trace. The layer-0 amplification summary plots the full layer-0 amplification distribution and highlights the MSC members as the two extremes.

Why $(0, 19)$ has the highest amplification yet is not in the MSC. The layer-0 amplification ranking is $(0, 19) = 65\times$, $(0, 17) = 42\times$, $(0, 28) = 8.6\times$ — and the $n = 50$ probe shows $(0, 19)$ is also a near-BOS-argmax head (argmax_pos $= 0$ on $44/50$ prompts, BOS mass $0.14\,\%$ — diffuse, not concentrated). Yet only $(0, 17)$ and $(0, 28)$ end up in the MSC. Mistral uses GQA with 32 query heads sharing 8 KV heads (4 query heads per KV group). Within layer 0, $(0, 17)$ and $(0, 19)$ sit in the same KV group (group 4, along with $(0, 16)$ and $(0, 18)$), so their value tensors are shared; substituting one while keeping the other softmax still uses the same V projection, making $(0, 19)$ functionally redundant with respect to $(0, 17)$. By contrast $(0, 28)$ is in KV group 7, a different group entirely. The MSC selects one head per causally important KV group, not the top-K by any per-head score — KV-group-level coverage is a hidden structural constraint we did not anticipate. Inspecting per-prompt S_p and sign-preservation statistics for the three heads distinguishes them sharply. Under the clamped concentration diagnostic, $(0, 17)$ has $S_p = 0$ on every prompt: relative to its score argmax, all other keys fall in the diagnostic's inactive branch, so the diagnostic sees a clean single-key row. This is a property of the clamped diagnostic, not of the unclamped $P_{4,0}$ map applied to the network. On these rows, the original softmax write is diffuse: BOS is only the incidental score peak ($0.42\,\%$ mass), with $\sim 99.6\,\%$ spread thinly over ~ 4000 non-BOS keys whose value directions largely cancel, so $\|y_{\mathrm{sm}}\|$ is small. Polynomial reweighting sharpens this diffuse tail unevenly, and the value directions no longer cancel to the same degree, producing a $42\times$ larger residual-stream write in a direction away from the softmax-mean manifold. The qualitative change for $(0, 17)$ is therefore a large off-manifold output-norm jump from a diffuse softmax write, not an increase in the head's attention entropy and not a collapse onto BOS — consistent with the BOS-mask null, where masking BOS on the protected set leaves retrieval unchanged. $(0, 19)$, by contrast, already carries significant non-BOS mass under softmax (S_p mean 3.69, $S_{p,\mathrm{safe}} = 0.28$), so the same substitution induces a smaller off-manifold displacement despite its larger raw amplification. The MSC therefore tracks an operator-induced output-direction transition, not raw amplification alone. $(0, 28)$ enters the MSC by a different mechanism: it has the strongest Pearson correlation between y_{sm} and y_{poly} across prompts ($+0.65$ vs $+0.11$ for $(0, 17)$ and $+0.04$ for $(0, 19)$) and an $80\,\%$ sign-preservation rate (H_+ rank-preserver). The empirical MSC is therefore one rank-preserver ($(0, 28)$, H_+) and one off-manifold amplifier ($(0, 17)$, H_-) — consistent with Proposition AX.5 (a head with high Ω_h can be useful when its signed contribution Y_h/Y_H has the right sign and magnitude) and a structure that no single rank-based metric (amplification, $|c_*|$, Ω_p, Wu) singles out. We emphasize this is an empirical Mistral observation, not a universal partition of every minimum sufficient protected set — see the counterexample below.

On Mistral, low-$|c_*|$ heads are systematically more polynomial-amplified within layer 0. Spearman correlation between per-head $|c_*|$ and amplification within Mistral's layer 0 is $\rho = -0.45$ ($p = 0.010$, $n = 32$). The same anti-correlation appears in Mistral's late layers ($\rho = -0.46$ at layer 22, -0.46 at layer 23, both $p < 0.01$); the global correlation is weak ($\rho = -0.09$) because it averages over layers with opposite or null relationships. The anti-correlation does not extend trivially to other models in the panel: Pythia-1.4b (Regime A″, alignment) instead shows $\rho = +0.51$ ($p = 0.046$, $n = 16$) in layer 0 — the opposite sign, consistent with its regime — while Llama-3.2-3B, Phi-4-mini, Qwen-2.5, Granite, and both Gemma-2 checkpoints show null correlations at layer 0. Whether the causally important layer in those models (Phi-4-mini layer 1, Granite layers 23–24, etc.) carries an analogous anti-correlation is a per-model empirical question left to follow-up. This per-layer anti-correlation has a clean interpretation: heads with small direct logit attribution along u have small $|y_{\mathrm{sm}}|$ but their polynomial reweighting can shift their $|y_{\mathrm{poly}}|$ by a large multiplicative factor (the same absolute leakage that is invisible to a high-$|c_*|$ head's normalization becomes a 10–40× ratio for a head whose denominator is small). This is precisely why low-$|c_*|$ heads can be both structurally invisible to DLA-style

detectors (the dissociation observed in the main text) and causally important under polynomial substitution (the MSC). The MSC heads $(0, 17)$ at rank 3/32 on $|c_*|$ (bottom) and rank 2/32 on amplification (top) are extreme exemplars of this anti-correlation, not idiosyncratic outliers.

Rank/sign duality. The Mistral MSC contains one head from each extreme of the Ω_p distribution: $(0, 28)$ at the bottom-2 % (mean $\Omega_p = -6.28$, individual sign-preservation rate 80 %) and $(0, 17)$ at the top-0.5 % (mean $\Omega_p = 22.81$, individual sign-preservation rate 48 %). This is not numerology — it is consistent with a simple algebraic fact about how per-head Ω_p values aggregate into a protected set-level $\Omega_p(H)$.

Proposition AX.5 (Signed-weighted-average form of $\Omega_p(H)$). For each head h with $Y_h \equiv Y_{\mathrm{sm}}(\{h\}) \neq 0$, define per-head quantities $\Delta_h \equiv Y_{\mathrm{poly}}(\{h\}) - Y_{\mathrm{sm}}(\{h\})$ and $\Omega_h \equiv -\Delta_h / Y_h$. For a protected set H with $Y_H \equiv \sum_{h \in H} Y_h \neq 0$,

$$\Omega_p(H) = \sum_{h \in H} w_h \, \Omega_h, \qquad w_h = \frac{Y_h}{Y_H}, \qquad \sum_{h \in H} w_h = 1.$$

The weights w_h are signed (since Y_h can be of either sign): $\Omega_p(H)$ is a signed weighted average of per-head Ω_h, not in general a convex combination.

Proof. From $\Omega_p(H) = -\sum_h \Delta_h / Y_H$ and $\Omega_h Y_h = -\Delta_h$ we have $-\sum_h \Delta_h = \sum_h \Omega_h Y_h$, so $\Omega_p(H) = \sum_h (Y_h / Y_H) \, \Omega_h$. The weights sum to 1 since $\sum_h Y_h / Y_H = 1$. They are signed because no constraint forces $Y_h \geq 0$. □

The proposition is the rank-vs-sign duality made precise: a head with $\Omega_h \gg 1$ (individually failing the protected set sign criterion) can still pull $\Omega_p(H)$ below 1 if its signed weight w_h has the right magnitude and sign relative to the other protected-set members. There is no need for a universal rank-preserver/sign-canceller partition of every minimum sufficient protected set.

Operator-graded version. Lemma N.1 is the same algebra with the operator left symbolic. Let $\Delta_{\mathcal{O}}(H) = Y_{\mathcal{O}}(H) - Y_{\mathrm{sm}}(H)$. Then $Y_{\mathcal{O}}(H) = Y_{\mathrm{sm}}(H) + \Delta_{\mathcal{O}}(H)$. Sign preservation is equivalent to

$$(Y_{\mathrm{sm}}(H) + \Delta_{\mathcal{O}}(H)) Y_{\mathrm{sm}}(H) > 0.$$

Since $Y_{\mathrm{sm}}(H)^2 > 0$, divide by $Y_{\mathrm{sm}}(H)^2$ to obtain

$$1 + \Delta_{\mathcal{O}}(H) / Y_{\mathrm{sm}}(H) > 0,$$

which is equivalent to

$$-\Delta_{\mathcal{O}}(H) / Y_{\mathrm{sm}}(H) < 1.$$

This is exactly $\Omega_{\mathcal{O}}(H) < 1$. Applying the same equivalence separately to $\mathcal{O}_1$ and $\mathcal{O}_2$ gives the two-operator statement in the main text. The proof is deliberately direct: the theorem is a statement about direct answer-direction contribution, not final model output.

For each head h and operator $\mathcal{O}$, define $Y_h = Y_{\mathrm{sm}}(\{h\})$, $\Delta_{\mathcal{O},h} = Y_{\mathcal{O}}(\{h\}) - Y_h$, and $\Omega_{\mathcal{O},h} = -\Delta_{\mathcal{O},h} / Y_h$. For any protected set with $Y_H = \sum_{h \in H} Y_h \neq 0$,

$$\Omega_{\mathcal{O}}(H) = \sum_{h \in H} \frac{Y_h}{Y_H} \, \Omega_{\mathcal{O},h}.$$

The weights sum to one but can be negative. Thus operator-specific compatibility cannot be inferred by simply sorting per-head scores: changing $\mathcal{O}$ changes the attention weights and hence the signed protected set drift, while changing H changes both the numerator and denominator of $\Omega_{\mathcal{O}}(H)$.

Empirical observation on Mistral (not a universal claim). On Mistral the discovered MSC contains one head with $\Omega_h < 1$ on a majority of prompts (an "H_+ rank-preserver") and one head with $\Omega_h > 1$ on a majority of prompts (an "H_- sign-canceller"). We observe this on Mistral but do not conjecture it as universal: the signed-weighted-average proposition allows MSCs composed entirely of H_+ heads, entirely of H_- heads, or mixed.

Cross-model mechanism scope. We test whether the Mistral mechanism (protected-set heads are BOS-argmax with anomalous polynomial amplification) generalizes to the MSCs of 8 other models. Per protected-set head and sampled non-protected-set peers in the same layer, we measure (i) the fraction of prompts where position 0 (BOS) is the softmax-attention argmax, (ii) mean attention mass at position 0, and (iii) the analytic ratio $\|y_{\mathrm{poly}}\| / \|y_{\mathrm{sm}}\|$ (head outputs computed from a single forward pass via captured pre-softmax scores and post-GQA value states), peer-normalized to a same-layer z-score. On Mistral-Inst $\{(0, 17), (0, 28)\}$ and Mistral-base $\{(1, 24)\}$ all protected-set heads are BOS-argmax on 100% of prompts with z-score ≥ 2 amplification; the BOS-amplifier characterization fits. On the remaining 7 models, the pattern fails: Granite-3.3-8B (layers 23–26, 80 heads) has BOS mass $\approx 0.05\%$ and amplification z-scores scattered around 0; Llama-3.1-8B (layers 0–6, 80 heads), Llama-3.2-3B (layers 0–4), Phi-4-mini, Qwen2.5-3B have mixed BOS-attendance and

uniformly low amplification (peer-relative z-scores typically $|z| < 2$); OLMo-2-0425-1B has 0 BOS-attending protected-set heads; RecurrentGemma-9B's singleton $(14, 14)$ is partially BOS-attending (84%) but not amplified ($z = +0.82$). The structural-functional dissociation (protected-set heads are bottom-rank under structural detectors yet causally important under polynomial substitution) generalizes across 7 of 9 standard transformer families; the BOS-argmax + anomalous amplification characterization of the protected set is Mistral-family-specific.

Cross-model head statistics. To probe whether any simple statistic universally distinguishes protected-set heads, we measure four per-head quantities on the final-query softmax-attention row at $n = 50$ NIAH pad=128 prompts, comparing each model's MSC heads to 5 sampled non-protected-set peers in the same layer(s): Shannon entropy of the softmax distribution (H_{sm}), Shannon entropy of the polynomial-renormalized distribution (H_{poly}), KL(sm $\|$ poly), and top-1 mass. Across 9 models, 7/9 have protected set H_{sm} at least as high as peer H_{sm} (range -0.9 to $+2.5$ nats; average $+0.95$ nats), suggesting protected-set heads tend to be more diffuse than their layer-peers under softmax. The KL pattern is mixed: protected set KL exceeds peer KL on Mistral-Inst (+20.8), Mistral-base (+2.0), Granite (+4.3), and RecurrentGemma-9B (+3.9); but is lower on Llama-3.1-8B, Llama-3.2-3B, Phi-4-mini, Qwen2.5-3B, and OLMo-2-1B (range -2.7 to -5.3). Classifying protected sets by these signatures: "diffuse + high-KL" (Mistral family + Granite + RG: 4/9); "diffuse-only" (Llama×2, Phi-4, Qwen: 3/9 – 4/9 if we include the Llama-3.2 edge case); "KL-only" (Granite if entropy is exactly tied: 1/9); "neither under these metrics" (Llama-3.2-3B, OLMo-2: 2/9). The weakest recurring signal is diffuse attention; the strongest cross-model signal is the structural-functional dissociation itself, not a per-head statistical signature. A unified cross-model mechanism characterization remains open and is left to follow-up work.

Layer concentration. Pairing the MSC search results with the existing layer-range protected-set intervention data exposes a second axis of cross-model variation: how concentrated the causally important role is at the layer granularity. On Mistral-7B-Instruct, keeping only layer 0 as softmax (rest polynomial) recovers 96% (100% of softmax baseline); the 2-head MSC retains 80% of softmax. On other models the picture differs sharply: Granite-3.3-8B's best single-layer keep recovers 36% of 66% softmax baseline (55%); Phi-4-mini's best L_0–L_7 range recovers 72% of softmax; Llama-3.2-3B, Pythia-1.4B, Qwen-2.5-3B have no single tested layer-range recovering above 2%. Their full MSCs (which span 3–8 layers in each case) do recover ($K \in [40, 80]$, multi-layer). This says: Mistral has a uniquely concentrated single-layer protected set; for the remaining 7 models the protected set spans multiple layers and the causally important role is layer-distributed. The early-layer dissociation we report in the main text is cross-model rather than universal: protected-set heads cluster in the early or mid stack on 7/9 models, whereas the layer-concentration property is not.

We report this honestly: the mechanism story is one model-family's story, and what distinguishes other models' protected sets (early-layer or, on Granite, mid-stack) is an open question left to follow-up work.

BOS-pathway null. A natural follow-up test on Mistral: if the protected-set heads' causally important role is mediated by the BOS token, then masking BOS attention in these heads while leaving softmax unchanged elsewhere should disrupt retrieval. We implement a per-head pre-softmax mask that sets the score at position 0 to $-\infty$ for specified (ℓ, h) pairs, leaving all other heads untouched (no polynomial substitution). On Mistral-7B-Instruct at $n = 100$ NIAH (substring scoring, $\max_new = 1$), the variants give: all-softmax baseline 95.0%; BOS-masked protected set $\{(0, 17), (0, 28)\}$ 95.0%; BOS-masked random layer-0 non-protected-set heads (three seeds) 95.0%; BOS-masked $(0, 19)$ (a non-MSC BOS-argmax head) 95.0%. Masking BOS attention on the protected-set heads has no measurable effect on retrieval. Combined with our amplification measurement showing both protected-set heads have very low BOS mass (0.42% on $(0, 17)$, 8.32% on $(0, 28)$; the remaining 99.6%/91.7% is spread thinly across $\sim$4000 non-BOS keys), the natural interpretation is: BOS-argmax is a property of where the protected-set heads' attention peaks but the causally important function is in the diffuse tail, not at BOS itself. Polynomial substitution amplifies the diffuse-tail attention pattern, and that amplification — not the BOS pathway specifically — is the causal mechanism. The "BOS-argmax diffuse-attention" label correctly describes the protected-set heads' attention geometry; we use "BOS-amplifier" here for shorthand but it should be read as "polynomial-amplifies-diffuse-attention-with-BOS-argmax-as-incidental-peak", not "causally depends on BOS."

Operator-specificity probes. Three further interventions on Mistral-7B-Instruct at $n = 100$ NIAH together show that the protected set's rescue is specific to the polynomial operator in this task, not to arbi-

trary attention degradation. (i) Inverse-BOS — keep only BOS for the protected-set heads (set every non-BOS pre-softmax score to $-\infty$ on $\{(0,17),(0,28)\}$), under softmax everywhere else: 95.0%, identical to baseline. The diffuse tail of the protected-set heads is not causally necessary under softmax. (ii) All-heads BOS-mask sanity — set the pre-softmax score at position 0 to $-\infty$ for all $32 \times 32 = 1024$ heads in every layer, no polynomial substitution: 96.0%, also identical to baseline. The model is globally robust to BOS removal under softmax via in-row renormalization, confirming that BOS attention is not a causal pathway for retrieval at all (and incidentally confirming the BOS-mask hook fires correctly on every layer/head). (iii) Uniform-attention substitution — replace softmax with a row-uniform-over-unmasked-keys distribution (mass spread evenly over all valid positions): all-uniform 0.0%, protected set-protected-uniform (protected set $\{(0,17),(0,28)\}$ kept softmax, remaining 1022 heads uniform) 0.0%, and five seeds of random-2-uniform controls also 0.0%. Protecting the protected-set heads from substitution recovers 74.2% retrieval against an all-polynomial null — but recovers nothing against an all-uniform null. The protected set is calibrated specifically for the polynomial operator's failure mode (heavy non-BOS amplification with rank-preservation), not for every non-softmax disruption. Equivalently, the protected-set heads are not "retrieval heads" in the Wu-protocol sense (their attention pattern is not where retrieval lives); they are operator-level compatibility patches whose softmax-shaped behavior is the specific signal that the surrounding polynomial-substituted heads can no longer produce.

Cross-model replication: property generalisation, not size generalisation. We replicate the protected set-protected uniform test on two further architectures using each model's own MSC protected set: Llama-3.1-8B-Instruct ($K = 80$, layers 0–6) and OLMo-2-0425-1B ($K = 40$, layers 0–6), at $n = 100$ NIAH each. Both replicate the Mistral pattern: all-softmax 100.0%/100.0%, all-uniform 0.0%/0.0%, protected set-protected-uniform 0.0%/0.0%, five random-K uniform controls 0.0% each. Across three families (Mistral 7B with GQA; Llama-3.1 8B with GQA; OLMo-2 1B with sliding-window attention) the MSC protected set recovers retrieval against polynomial substitution but not against uniform substitution. We emphasise this is a property-level generalisation (operator-specificity is observed across the tested architectures) rather than a size-level one: the required protected set size differs by model ($K = 2$ on Mistral, $K = 80$ on Llama-3.1, $K = 40$ on OLMo-2), and the size depends on architecture as well as the operator. The compact two-head case study is Mistral-specific; operator-specificity is observed in all three tested architectures.

Operator-graded transfer within the polynomial family. We further probe specificity by transferring the Mistral protected set across operator recipes, at $n = 100$ NIAH each (5 random-2 controls):

Operator	all-softmax	all-op	protected set-protected	random-2 max
poly $p=4, b=0$ (calibration)	91.6%	12.6%	**74.2**%	17%
poly $p=4, b=1$	95.0%	95.0%	95.0%	95%
poly $p=8, b=0$	95.0%	1.0%	**19.0**%	2%
sigmoid $b=0$	95.0%	4.0%	6.0%	5%
Wortsman ReLU$^1/L$	95.0%	0.0%	0.0%	0%

Three predictive observations follow. (i) The bias shift $p = 4, b = 1$ does not damage retrieval under all-polynomial substitution (95%), so the protected set effect collapses; the protected set matters only when the substitution itself damages retrieval. (ii) Higher polynomial degree ($p = 8$) is more damaging than $p = 4$, and the protected set partially transfers (protected set 19% vs random max 2%, +17pp dissociation) but the recovery is much weaker than at the calibration recipe. (iii) Both non-polynomial operators tested (sigmoid, Wortsman) show zero protected set advantage over random — the protected set offers no protection outside the polynomial family. Together these observations refute a "softmax-compatibility-general" interpretation of the protected set and pin it down to a polynomial-recipe-specific compatibility patch, with protected set efficacy graded by how much the operator damages retrieval and by distance from the calibration recipe within the polynomial family.

Rescue-ratio formalisation. The operator-transfer table is easier to read in a damage-normalised form. For substitution operator O and protected head set H, define the rescue ratio

$$\mathcal{R}(O,H) \;\equiv\; \frac{\mathrm{Acc}(O,H) - \mathrm{Acc}(O,\varnothing)}{\mathrm{Acc}_{\mathrm{sm}} - \mathrm{Acc}(O,\varnothing)},$$

which is the fraction of operator-induced damage recovered by protecting H, whenever $\mathrm{Acc}_{\mathrm{sm}} > \mathrm{Acc}(O,\varnothing)$ (the operator damages retrieval). Whenever $\mathrm{Acc}(O,\varnothing) \le \mathrm{Acc}(O,H) \le \mathrm{Acc}_{\mathrm{sm}}$, both numerator and denominator are nonnegative and the second is positive, so $0 \le \mathcal{R}(O,H) \le 1$. Applying to the five rows above: poly $p = 4, b = 0$ gives $\mathcal{R} = (74.2-12.6)/(91.6-12.6) \approx 0.78$; poly $p=8, b=0$ gives $\mathcal{R} \approx 0.19$; sigmoid gives $\mathcal{R} \approx 0.02$; Wortsman gives $\mathcal{R} = 0$; poly $p = 4, b = 1$ is undefined (denominator zero). The rescue ratio quantifies the operator-graded effect cleanly: protected set efficacy under the calibration recipe is $\sim 4\times$ stronger than at $p = 8$ and $\sim 40\times$ stronger than at sigmoid.

Constructive operator-specificity: no compact uniform protected set in the tested seed family. A natural follow-up: does any protected set recover retrieval under uniform substitution? We run the greedy MSC search procedure but seeded under uniform-attention instead of polynomial, sweeping seed sizes $K \in \{40, 80, 160, 320, 640\}$ ($K = 640$ is 62.5% of Mistral's 1024 heads). All five seed sizes report zero retrieval (0.0% on $n=20$ NIAH per cell), and the greedy descent does not even initiate because each seed protected set already falls below the 0.5× softmax-baseline floor. This is a striking asymmetry within the bottom-$|c_*|$ seed family: under polynomial substitution at $p=4, b=0$, a two-head protected set (0.2% of heads) recovers 74.2% retrieval; under uniform substitution, 640 heads (62.5% of heads) protected with the same bottom-K $|c_*|$ seed recover zero retrieval — a 312× protection-cost asymmetry. Polynomial substitution admits a tiny patch within this seed family; uniform substitution does not recover through any bottom-$|c_*|$ seed we tested. We do not claim a global impossibility over all possible uniform-attention protected sets.

Finite-family exclusion certificate. The zero-success result above can be tightened into a quantitative upper bound on true protected set accuracy. Let $\mathcal{F} = \{H_K : K \in \{40, 80, 160, 320, 640\}\}$ be the finite family of bottom-K $|c_*|$ uniform-protection protected sets tested. For each $H \in \mathcal{F}$, evaluating on n independent prompts and observing zero successes gives a one-sided Clopper–Pearson upper bound $U_H = 1-\delta^{1/n}$ on the true success probability q_H at tail probability δ. Setting $\delta = \beta/|\mathcal{F}|$ and applying a union bound, with family-wise confidence at least $1-\beta$,

$$q_H \;\le\; 1 - (\beta/|\mathcal{F}|)^{1/n} \quad \text{for every } H \in \mathcal{F}.$$

At $|\mathcal{F}|=5$, $\beta=0.05$, $n=20$: $q_H \le 1-(0.01)^{1/20} \approx 21\%$ family-wise. Tightening to $n=100$ would give $\sim 4.5\%$. The honest claim is therefore not "uniform attention has no compatible protected set," but "within the tested bottom-$|c_*|$ seed family for Mistral-7B-Instruct, the true accuracy of every protected set tested is at most 21% (95% family-wise confidence at $n = 20$); reaching the 45% floor for polynomial-recipe-style rescue requires a protected set discovered by a different search procedure than bottom-$|c_*|$ seeding."

Sigmoid does admit a compact compatible protected set — and it differs from polynomial's. A natural question after the uniform-vs-polynomial asymmetry: does every damaging non-softmax operator lack a compact protected set, or is uniform the special case? We run the same greedy MSC search procedure with the polynomial operator replaced by sigmoid-with-zero-bias on Mistral-7B-Instruct, seeded with the same bottom-40 $|c_*|$ heads. At $n=10$ the seed protected set achieves 80% retrieval (above the 45% floor) and the descent successfully shrinks the protected set all the way from $K = 40$ to $K = 1$ while maintaining $\geq 80\%$ accuracy at every accepted round. We reproduce at $n = 20$ (threshold 0.5× softmax) on two seed sizes: $K = 40$ descends to size 1 at 80% accuracy, and $K = 80$ descends to size 1 at 85%. Critically, during both descents the polynomial-MSC heads $(0, 17)$ and $(0, 28)$ are removed without breaking; the $K=40$ descent converges to a single head $(0, 12)$, identity-disjoint from the polynomial protected set but still in layer 0. Sigmoid and polynomial therefore share the layer-0 location but not the specific head identity — suggestive evidence, within this Mistral case study, that layer 0 may be a privileged location for compatibility heads while the specific head identity is operator-specific. (The $K = 160$ seed-size descent was started but stopped at size 154, 90% accuracy, for budget reasons; the partial descent is consistent with the K=40/K=80 results.) The picture across three damaging operators is therefore: polynomial has a compact $K=2$ protected set on Mistral; sigmoid has an even smaller protected set that is identity-disjoint from polynomial's; uniform has no compact protected set within the bottom-$|c_*|$ seed family up to $K = 640$. Sample sizes for sigmoid descents are smaller than the polynomial $n=500$ headline number; we therefore present sigmoid as a paired-operator probe, not as a comparable second case study. The qualitative finding (operator-specific protected sets, different identities across operators) appears in both sigmoid descents; because the sample sizes are small, the quantitative match to polynomial's case-study depth is left to follow-up. "Compatibility protected set" is a per-operator object, not a property of the architecture.

Bias-dependence of the collapse: the collapse claim is specific to $b = 0$. Stress-test scope: is the 12/14-checkpoint collapse claim a knife-edge artifact of $b=0$? At $n = 50$ NIAH pad=128 we compute b_{auto} via 5%-mass-budget calibration and evaluate all-polynomial retrieval at $\{b_{\text{auto}}/2,\, b_{\text{auto}},\, 2b_{\text{auto}}\}$. On Mistral-7B-Instruct ($b_{\text{auto}} = 3.31$): softmax baseline 96.0%, all three multipliers 96.0% each. On Gemma-2-2B-it ($b_{\text{auto}} = 1.71$): softmax baseline 100.0%, all three multipliers 100.0% each. These two checks show that the canonical zero-bias collapse is not an unavoidable consequence of polynomial attention on those checkpoints: moving the global level can restore literal NIAH. This is a scoped statement, not a general calibrated-bias rescue claim; elsewhere $P_{4,b_{\text{auto}}}$ still collapses OLMo/Qwen NIAH, NoLiMa on Mistral/Llama, and PG-19 on several families. We there-

fore keep the headline tied to the canonical $b = 0$ stress recipe, and treat b_{auto} as an adaptive diagnostic workaround rather than a deployable substitution (see Table 58).

Downstream attention-delta scan. The protected-set heads sit in layer 0; which downstream heads' attention patterns are most sensitive to whether the protected set is preserved? We run two conditions on the same $n = 100$ NIAH prompts: clean (protected set softmax, rest polynomial; the 74.2% paper baseline) and corrupted (every head polynomial; the 12.6% all-polynomial baseline). At each downstream (ℓ, h) we capture the softmax attention row at the final query position under both conditions and compute the L_2 distance between the two distributions, averaged across prompts. This is an attention-pattern divergence scan, not a full path-patching circuit trace in the sense of Wang et al. (2023); Conmy et al. (2023): we identify candidate downstream consumers by where protected set substitution changes attention; we do not yet establish that those consumers causally mediate retrieval. Per-layer maximum deltas:

layer	3	9	12	14	15	17	18	21	30
max $\|\Delta\|_2$	0.34	0.32	0.44	0.33	**0.69**	0.41	**0.65**	0.34	0.27

The downstream consumption is concentrated in layers 12–18 (peak at layer 15, $\|\Delta\|_2 = 0.69$ for head $(15, 24)$). The early layers 1–3 show only modest deltas (≤ 0.34), and very late layers 24–31 are essentially insensitive (≤ 0.27). The top-impact set is dominated by layer-12–18 heads; its first entries are $(15, 24)$, $(18, 1)$, $(18, 0)$, $(18, 3)$, and $(12, 6)$. The Wu-style attention-at-needle head $(12, 7)$ has L_2 delta 0.159 and ranks 191/1022 among non-protected-set heads — well below the top-100 cutoff of 0.198, and approximately the median (0.096) of the non-protected-set distribution. The candidate downstream consumers form a distinct head population from the structurally identified retrieval heads. This is consistent with the structural-functional dissociation: protected set removal changes middle-layer attention patterns different from the heads that structural detectors (Wu-style, $|c_*|$) flag. It does not by itself establish those heads as causal consumers.

Causal consumer test: the top-impact consumers are correlated, not causally privileged. To check whether the top-impact downstream heads from the delta scan are causally privileged for retrieval, we ran a direct intervention. Comparing on $n = 100$ NIAH prompts ($p = 4, b = 0$, first-token scoring): protected set alone $\{(0, 17), (0, 28)\}$ recovers 81.0%; protected set plus the top-5 downstream consumers $\{(15, 24), (18, 1), (18, 0), (18, 3), (12, 6)\}$ recovers 82.0% (+1.0 pp); protected set plus 5 random non-protected-set non-layer-0 heads (five independent seeds) recovers $82.2 \pm 1.5\%$ on average, max 83.0%. Adding the top-impact consumers gives essentially the same effect as adding random downstream heads. This is a negative result and consistent with the limitation emphasized in the main text: the attention-delta scan identifies heads whose attention patterns change when the protected set is polynomialized (a correlate), but those heads are not causally privileged consumers of the protected-set output relative to random non-protected-set heads. A plausible interpretation is diffuse downstream use rather than a small consumer set, but this remains a hypothesis rather than a traced circuit. We therefore do not present the layer-12–18 candidates as a circuit; the scan establishes that the protected set changes downstream attention, but not that any particular downstream head causally mediates the rescue.

Differentiable cross-check. As an orthogonal test of the structural-functional dissociation, we train a learnable per-head Hard-Concrete gate (Louizos et al., 2018) that mixes softmax and polynomial attention, optimising language-modelling loss on $n = 500$ PG-19 samples at seq_len = 2048 for 1000 steps with sparsity penalty $\lambda = 10^{-2}$. At threshold 0.5, the converged mask retains $K = 14$ heads as softmax on Mistral-7B-Instruct (1024 total) and $K = 23$ on Granite-3.3-8B-Instruct (1280 total). Both learned sets have Jaccard overlap 0 with the top-K by $|c_*|$ (intersection 0/14 on Mistral, 0/23 on Granite), consistent with the random null ($\mathbb{E}[\text{Jaccard}] \approx 0.007$ on Mistral, 0.009 on Granite for size-K sets drawn uniformly). On Mistral the learned set is:

$$(0, 7), (0, 12), (1, 6), (1, 7), (1, 25)$$
$$(3, 0), (3, 19), (4, 8), (4, 9), (4, 25)$$
$$(5, 12), (6, 11), (7, 17), (12, 28)$$

Thus 13/14 heads concentrate in layers 0–7, with $(12, 28)$ the sole exception. A completely differentiable optimisation, with no access to our greedy MSC search or structural detectors, independently locates a causally important softmax-protected head set that (a) is concentrated in early layers — corroborating the direct layer-range test that recovers 94 % NIAH retrieval from layer 0 alone — and (b) is disjoint from the structural top-K by $|c_*|$. The two protected sets (greedy MSC of size 2 and Hard-Concrete of size 14) are different (only 1 of the 14 Hard-Concrete heads, $(12, 28)$, sits outside the early-layer cluster), consistent with the existence of multiple sufficient protected sets; both confirm the same negative result against high-$|c_*|$ retrieval-head protection.

Counterexample to a universal $|H_-| \geq 1$ claim. Consider a two-head toy with fixed negative background $B = -1$ (from substituted heads outside the protected set) and two protected-set members h_1, h_2 each contributing $Y_{\text{sm}}(\{h_i\}) = Y_{\text{poly}}(\{h_i\}) = 0.6$. Each head has $\Delta_i = 0$ and hence $\Omega_i = 0 < 1$ — both are H_+ rank-preservers. Singletons $\{h_1\}$ and $\{h_2\}$ give total logits $B + 0.6 = -0.4 < 0$ (not θ-sufficient), but $\{h_1, h_2\}$ gives $B + 1.2 = +0.2 > 0$ (sufficient). The minimum sufficient protected set is $\{h_1, h_2\}$ with $H_- = \emptyset$. The universal claim "every minimum protected set contains at least one sign-canceller" is therefore false, even under the strict Ω_h-based definition of H_+/H_-. This counterexample motivates the proposition-plus-empirical-pattern framing above.

Empirical descent profile on Mistral. The greedy MSC trace nonetheless gives a striking empirical pattern: at every accepted protected set along the 38-round descent from bottom-40 to size 2, the protected set partitions into H_+ (per-head $\Omega_p < 1$ on a majority of prompts) and H_- (per-head $\Omega_p > 1$ on a majority of prompts). The seed contains $|H_+| = 32$ rank-preservers and $|H_-| = 8$ sign-cancellers; the greedy algorithm monotonically removes H_+ heads first and only begins shedding H_- heads at round 7. By the converged protected set at round 38, the partition is $|H_+| = 1$, $|H_-| = 1$. Across all 39 accepted rounds, $|H_-|$ never reaches 0 on Mistral. We report this as a Mistral-specific descent profile, not as a structural prediction — consistent with the counterexample above, an MSC search on a different model could traverse a descent path that empties H_- entirely.

The empirically observed descent profile and the leakage-matrix spectrum (Mistral bottom-40: stable rank ≈ 3 at relative threshold 0.5, $\sigma_2/\sigma_1 = 0.93$, $\sigma_3/\sigma_1 = 0.52$) are both consistent with $|H_{\text{suff}}| = 2$ on Mistral to within an order-one factor; leakage-matrix details are in the supplementary artifact (leakage_matrix_analysis.md). Any protected set search procedure (greedy, Ω_p-seeded, differentiable) is expected to converge to a protected set satisfying Proposition AX.5, but not necessarily to one with both H_+ and H_- heads present.

AY. Protected-set size saturation curve $R(K)$

The protected-set search halts at $K = 2$ on Mistral, but a reader may reasonably ask whether the rescue ratio grows sharply for $K > 2$ or saturates near the $K = 2$ value. We sweep $K \in \{2, 3, 4, 5, 8\}$ by extracting the first K greedy-trace accepted protected sets and evaluating each on the $n = 500$ NIAH held-out set ($p = 4$, $b = 0$). Table 55 reports the result.

Table 55. $R(K)$ on Mistral-7B-Instruct, NIAH $n = 500$, $p = 4$, $b = 0$. Protected sets are nested initial segments of the greedy MSC descent trace. Random-K max is the maximum over 5 seeds at $n = 500$. Span = $\text{Acc}_{\text{sm}} - \text{Acc}_{\text{poly}} = 91.6 - 12.6 = 79.0$ pp.

K	protected set acc	$R(K)$	rand-K max
2	74.2%	0.78	17.0%
3	78.8%	0.84	15.2%
4	75.6%	0.80	14.2%
8	80.8%	0.86	25.0%

The curve saturates: doubling K from 2 to 4 leaves R unchanged within sample noise, and doubling again to 8 adds only +0.07 to R. Random-K controls never reach the protected set across the sweep ($K = 8$ random max 25.0% vs protected set 80.8%). This is consistent with the $K = 2$ saturation reading: $\sim 92\%$ of the available K=8 rescue is already captured by the first two greedy picks, and incremental heads provide diminishing returns.

Operator transfer ($p = 6$ details). Table 3 shows $R_{p=6} = 0.24$ for the Mistral K=2 protected set; this paragraph gives the random-control breakdown. At $p = 6$, $b = 0$ on Mistral NIAH $n = 500$: softmax 91.6%, all-polynomial 3.0%, K=2 protected set 24.2%, random-2 over 5 seeds mean 3.5%, max 4.8%, 0/5 at or above protected set. The protected set still beats every random-2 draw, but rescue magnitude collapses from 0.78 at the calibration recipe to 0.24 at $p = 6$ and 0.19 at $p = 8$. The same K=2 identity that gives near-saturation rescue at $p = 4$ provides only one-third of that rescue at $p = 6$, quantifying the operator-specificity claim along the polynomial-degree axis.

AZ. QK/OV single-head and structural protected-set controls

The QK/OV decomposition below reports two protected sets. Single-head and structural protected-set controls at $n = 100$ NIAH (Mistral-7B-Instruct, $p = 4$, $b = 0$) localise the mechanism within each protected set and rule out the structural top-2 heads as alternative explanations. All values are accuracies; R is the rescue ratio under the $n = 100$ span ($\text{Acc}_{\text{sm}} = 0.96$, $\text{Acc}_{\text{poly}} = 0.13$, span $= 0.83$). In this appendix, a head's "write" means its value-weighted contribution to the residual stream or, after multiplication by the unembedding, to logit space.

OV-circuit structural correlate. For each layer-0 head we compute the per-head OV map $W_O[:, h]\, W_V[h] \in$

Table 56. Per-head and structural protected-set QK/OV controls at $n = 100$ NIAH. "softmax" protects the listed heads with softmax inside an otherwise-polynomial network; "uniform" replaces their attention pattern with uniform; "zero" silences them entirely. The greedy MSC and alt-1 protected sets at $n=500$ are defined in Appendix AX.

protected set	softmax	uniform	zero
$\{(0,17)\}$ alone	62.0 (R=0.59)	74.0 (R=0.74)	70.0 (R=0.69)
$\{(0,28)\}$ alone	29.0 (R=0.19)	47.0 (R=0.41)	49.0 (R=0.43)
$\{(0,19)\}$ alone	38.0 (R=0.30)	38.0 (R=0.30)	26.0 (R=0.15)
$\{(0,22)\}$ alone	46.0 (R=0.40)	40.0 (R=0.32)	30.0 (R=0.20)
greedy $\{(0,17),(0,28)\}$ [n=500]	74.2	85.2	86.2
alt-1 $\{(0,17),(0,22)\}$ [n=500]	79.4	76.8	57.2
alt-2 $\{(0,19),(0,22)\}$ [n=500]	67.6	58.6	21.0
random $\{(2,27),(24,24)\}$ seed=42	10.0 (R=−0.05)	6.0 (R=−0.10)	11.0 (R=−0.04)
top-Wu $\{(12,7),(18,0)\}$	12.0 (R=−0.02)	0.0 (R=−0.17)	1.0 (R=−0.16)
top-$\lvert c_*\rvert$ $\{(30,5),(30,6)\}$	13.0 (R=−0.01)	14.0 (R=0.00)	14.0 (R=0.00)
all layer-0 (32 heads)	**93.0** (R=0.97)	0.0 (R=−0.17)	0.0 (R=−0.17)
greedy K=8 trace protected set	89.0 (R=0.93)	**94.0** (R=0.99)	82.0 (R=0.84)

$\mathbb{R}^{d\times d}$ (rank $\leq d_{\text{head}}$) and its singular value spectrum on Mistral-7B-Instruct. Effective rank — defined as $\sum_i \sigma_i^2/\sigma_1^2$, which is small when the OV write concentrates in a single direction and large when it spreads over many directions — decisively separates the two mechanism types:

head	operator norm σ_1	effective rank	protected set role
$(0,17)$	0.040	8.46	shared (greedy + alt-1)
$(0,19)$	0.033	12.38	high-amplification, not in protected set
$(0,22)$	0.050	**3.74**	causally important under softmax (alt-1, alt-2)
$(0,28)$	0.053	**18.09**	diffuse-write head (greedy)

$(0,22)$ has the most concentrated OV write of any tested protected-set head (low eff-rank, write in ~ 4 directions); $(0,28)$ has the most diffuse (eff-rank ~ 18). Both share the same order of operator norm. This is the predicted structural signature: causally important heads write through a narrow direction (consistent with a retrieval-direction write); polynomial-incompatible diffuse-write heads spread their write across many directions, giving the substituted operator more directions in which to inject off-manifold energy. Operator norm alone does not discriminate; effective rank does.

Cross-model QK/OV: OLMo diagnostic pair. The cross-model protected-set audit above uses OLMo-2-0425-1B's $K = 40$ MSC. For the narrower QK/OV mechanism probe, we also tested an OLMo $K = 2$ diagnostic pair $\{(4,1),(11,0)\}$ at $n = 100$ NIAH ($p = 4$, $b = 0$). This pair is not the OLMo MSC: it is a compact probe used to test whether the Mistral kill-QK/kill-OV decomposition qualitatively repeats in a different architecture. Softmax baseline is 100% and all-polynomial is 1%:

intervention	accuracy	R
pair=softmax (control)	22%	0.21
pair=uniform (kill QK)	**29**%	**0.28**
pair=zero (kill OV)	3%	0.02

The OLMo K=2 diagnostic pair is a mixed-mechanism probe: uniform attention recovers more than softmax (+7 pp, kill-QK helps), but zero attention collapses retrieval to all-polynomial (kill-OV hurts). The kill-QK > softmax-control direction therefore repeats across two model families, while the kill-OV behaviour is Mistral-greedy-specific. Reading: the selected pair's attention pattern is dispensable in both models; what differs is whether the selected pair's attention output is also dispensable (Mistral greedy) or causally important (OLMo).

The single-head rows establish the per-head story for the greedy MSC: both $(0,17)$ and $(0,28)$ are individually destructive under polynomial. Silencing $(0,17)$ alone gives $R = 0.69$ (vs $R = 0.59$ when softmax-protected), and silencing $(0,28)$ alone gives $R = 0.43$ (vs $R = 0.19$). The two heads' dispensability effects roughly add to the K=2 greedy protected set's $R = 0.93$, confirming that both greedy MSC members are operator-incompatible diffuse-write heads individually and jointly.

The structural protected-set negative controls confirm the dissociation. The top-2 Wu heads $\{(12,7),(18,0)\}$ and the top-2 $|c_*|$ heads $\{(30,5),(30,6)\}$ both give $\sim 13\%$ retrieval under softmax-protection (no rescue, identical to all-polynomial); silencing them either leaves retrieval unchanged or pushes it slightly below all-polynomial. These structural top-K heads are neither incompatible diffuse-write heads nor causally important for polynomial substitution — consistent with the structural–functional dissociation.

The K=32 all-layer-0 row pins the layer-0 collective: protecting all 32 layer-0 heads with softmax recovers 93% ($R = 0.97$), nearly the softmax baseline, but silencing all 32 collapses retrieval to 0%. This is decisive evidence that layer-0 as a whole is causally important, while only $(0,17)$ and $(0,28)$ within layer-0 are net destructive under polynomial.

The single-head rows for $(0,19)$ and $(0,22)$ confirm the head-by-head dissociation: silencing $(0,19)$ alone drops R from 0.30 to 0.15 (-15 pp), and silencing $(0,22)$ alone drops R from 0.40 to 0.20 (-20 pp). Both are causally important under polynomial. The four-head per-head map is then: operator-incompatible diffuse-write heads $\{(0,17),(0,28)\}$ where silencing improves R; causally important under polynomial $\{(0,19),(0,22)\}$ where silencing harms R. This aligns with the OV-circuit effective-rank table above: causally important heads have low eff-rank (concentrated writes); incompatible heads have higher eff-rank (diffuse writes).

The K=8 greedy trace protected set row makes the silencing-saturation story complete: protecting all 8 with softmax gives $R = 0.93$, replacing their attention with uniform gives $R = 0.99$ (nearly the softmax baseline), and silencing all 8 gives $R = 0.84$. The pattern is consistent with the per-head map: the K=8 protected set includes two incompatible diffuse-write heads and six causally important heads, so silencing harms the causally important six but helps with the incompatible two, netting a slightly lower R than softmax-protect.

OV→Vocab projection. For each protected-set head h at layer ℓ we compute the head's logit-write matrix $M_{\ell,h} = W_U \cdot W_O[:, h] \cdot W_V[h, :]$, shape (vocab, d_{model}). M has rank at most $d_{\text{head}} = 128$. A weight-only OV→Vocab projection of the four layer-0 candidates has low effective rank (2.67–7.54) and does not align with answer-token directions, consistent with structural-detector failure. But it is only a weak guide to the heads' behavior on real inputs: an empirical SVD of the protected set's actual attention output to logit space ($W_U \cdot o_proj \cdot attn_output$ across 100 prompts) gives a top-1 direction whose cosine with the weight-only OV→Vocab top-1 is 0.47, with 0/30 token overlap in the top-promoted set. The empirical write is determined by the V_{eff} the head computes on real inputs, differs from the head's weight-preferred direction, and is rank-1 across prompts without a clean semantic decoding. We therefore rely on the cross-prompt SVD and mean-patching results below, not on the OV→Vocab projection, for the mechanism claim.

Cross-prompt SVD. We capture each protected-set head's attn-output at the final query position under softmax for 100 NIAH prompts, stack into a $100 \times (K \cdot d_{\text{head}})$ matrix W, and SVD. Effective rank $= \sum_i \sigma_i^2 / \sigma_1^2$ measures how many directions the head writes across prompts:

head set	task	σ_1	eff-rank	var in σ_1	var in top-5
greedy $\{(0,17),(0,28)\}$	NIAH	0.74	**1.04**	**95.9**%	97.7%
greedy $\{(0,17),(0,28)\}$	NoLiMa	1.13	**1.00**	**99.8**%	99.9%
alt-1 $\{(0,17),(0,22)\}$	NIAH	0.92	**1.00**	**99.8**%	100.0%
OLMo K=2 diagnostic pair $\{(4,1),(11,0)\}$	NIAH	29.2	**1.06**	**94.6**%	97.2%
top-Wu $\{(12,7),(18,0)\}$	NIAH	55.1	1.75	57.0%	71.0%

The rank-1 result is therefore unified across both K=2 candidate protected sets on Mistral, across the NoLiMa task, and across the OLMo K=2 diagnostic pair: these selected low-dimensional head sets write a near-constant rank-1 vector in the residual stream regardless of prompt. The top-Wu retrieval-head control has eff-rank 1.75 with only 57% variance in σ_1, confirming that genuine retrieval heads vary their write across prompts while the selected compatible/probe head sets do not.

Across the Mistral MSC/candidate sets and the OLMo diagnostic pair, the selected head-set output across 100 prompts has effective rank essentially equal to 1: a single direction in the concatenated head-output space captures 94–99.8% of the across-prompt spectral energy $\sigma_1^2 / \sum_k \sigma_k^2$. Because the per-prompt write matrix M is stacked without column-centering, this is energy, not centered variance: a near-constant DC component contributes to σ_1. This is what we want here, since the near-rank-1 near-constant write is the headline mechanism. The top-Wu retrieval-head control has eff-rank 1.75 and only 57% energy in σ_1, confirming that retrieval heads vary their write across prompts (as expected for actual retrieval) while the selected compatible/probe head sets output near-constants.

Mean-direction patching. The direct test of "the protected-set outputs a constant" is to replace the protected-set heads' attention output with the mean of all donor captures (a fixed per-head vector), and patch this single constant into every test prompt's polynomial forward pass. At $n = 100$ NIAH on Mistral greedy MSC, mean-patching gives match-A **93**% versus 82% under different-target donor patching, ~ 81% under protected-set=softmax, and 86% under protected-set=zero. Mean-patching gives the highest rescue we have found, beating every per-prompt intervention. The result generalises to alt-1 (mean-patch 91%).

Cross-degree confirmation. We compute the cross-prompt SVD of the protected-set attn-output under polynomial substitution at $p \in \{2, 4, 6, 8\}$ ($b = 0$, $n = 100$ NIAH):

operator	σ_1	eff-rank	var in σ_1	rescue R (from Table 3)
softmax	0.74	1.04	95.9%	1.00
poly $p=2$	3.07	1.01	99.4%	0.98
poly $p=4$	3.02	1.00	99.8%	0.78
poly $p=6$	2.94	1.00	100.0%	0.24
poly $p=8$	2.91	1.00	100.0%	0.19

σ_1 is essentially constant across the polynomial degree axis (all ~ 3.0), yet rescue R varies fivefold ($0.98 \to 0.19$). Magnitude alone cannot explain the failure spectrum; the polynomial direction of the rank-1 write must shift across degree, moving farther from the softmax-tolerated direction as p increases. At $p = 2$ the polynomial-amplified write is close enough to the softmax-mean direction that downstream layers tolerate it ($R = 0.98$); at $p = 8$ the direction has rotated far enough that the same magnitude becomes destructive ($R = 0.19$). This rules out the simple "magnitude inflation" story and confirms the direction-×-magnitude reading: the polynomial mechanism's failure is an angular drift away from the original-softmax direction, not a magnitude blow-up.

Direction × magnitude: a clean interaction. A direction-vs-magnitude sweep clarifies the mechanism. Patching the protected-set heads with either the empirical mean direction or a random Gaussian, at 0.5×, 1×, 2×, and 4× the natural softmax magnitude:

direction	0.5×	1×	2×	4×
mean (softmax-empirical)	—	**93**%	92%	**89**%
random Gaussian	91%	84%	—	**9**%

The mean direction tolerates any magnitude up to 4× (89–93% across the range, $\sigma_1 = 3.02$ approximates the polynomial-amplified magnitude). A random direction tolerates only low magnitudes: at 0.5–1× it gives 84–91%, but at 4× retrieval collapses to 9% ($\leq$ all-polynomial floor 14%). The destructive factor is therefore neither pure magnitude nor pure direction but their interaction: downstream layers tolerate any low-magnitude write, or any-magnitude write along the softmax mean direction, but they cannot tolerate a high-magnitude write in a non-mean direction. Polynomial substitution puts the protected set in exactly that state: cross-prompt SVD under polynomial confirms rank-1 with $\sigma_1 = 3.02$ (4× amplification), but the polynomial write is in a different rank-1 direction than the softmax mean (otherwise polynomial would be tolerated by the same mechanism that tolerates mean-patch at 4×). The amplification mechanism is direction-conditional: polynomial substitution amplifies protected-set outputs orthogonal to the original-softmax direction, and that orthogonal high-magnitude write is what downstream layers cannot absorb.

Logit lens: the answer signal trajectory through layers. Following tuned-lens and logit-lens style probes (Belrose et al., 2023), we trace the layer-by-layer rank of the correct answer token in the residual stream under three operators on $n = 20$ NIAH prompts:

operator	emergence (median rank ≤ 10)	rank-0 at final layer ($n = 20$)
softmax	layer 18	20/20
all-polynomial	never (median rank 16 at layer 31)	2/20
protected-set=zero (silenced)	layer 22	20/20

Softmax produces the answer-token signal at layer 18 and refines it to rank-0 by layer 25. All-polynomial never produces the signal: median rank at the final layer is 16 (∼2/20 prompts solved). Silencing the protected set at layer 0 delays answer emergence to layer 22 (four layers later than softmax), and the model still reaches rank-0 at the final layer on 20/20 prompts. The destruction therefore localises: it is something the polynomial protected-set outputs at layer 0 that propagates through the residual stream and prevents the answer signal from ever crystallising; removing that write delays but does not prevent retrieval.

Layer-0 rank-1 control. Applying the same cross-prompt SVD to all 32 layer-0 heads jointly (4096-dim joint write across 100 NIAH prompts) gives eff-rank 1.02 and 98.5% energy in σ_1 (same uncentered convention as above). Layer-0 as a whole writes a near-rank-1 vector to the residual stream, of which the K=2 MSC protected set is a particular sub-decomposition that is uniquely amplified by polynomial substitution. The rank-1 structure is therefore necessary but not sufficient to identify the destructive protected set; the additional identifier is the per-head OV-amplification factor under polynomial (Section 4, 42× and 9×).

Cross-prompt path patching. The strongest test of whether the protected set is doing actual retrieval is to substitute the protected set's clean softmax output from one prompt into another prompt's polynomial forward pass. If the protected set encodes prompt-specific retrieval, then patching a donor prompt B's protected-set output into test prompt A's polynomial run should break A's retrieval (because B's protected-set output encodes B's needle, not A's). We run this experiment on 100 NIAH pad-128 test prompts paired with random donor prompts B whose target token $\neq A$'s target token (donor seed 42). Per-test-prompt: capture protected-set attn-output at layer 0 under softmax with prompt B; run polynomial substitution forward on A with the protected-set heads' output overwritten by B's captured tensor; record whether the predicted first token matches A's target, B's target, or neither.

Table 57. Cross-prompt path-patching of protected set attention output. A = test prompt, B = donor (different target). 100 test prompts at $p = 4, b = 0$, NIAH pad=128. Donor-target rate 0% across the board: protected set never carries the donor's answer.

head set	match A (test target) ± 95% CI	match B	neither
greedy $\{(0,17),(0,28)\}$ (L0)	**82**% [74, 89]	0%	18%
alt-1 $\{(0,17),(0,22)\}$ (L0)	**85**% [78, 92]	0%	15%
alt-2 $\{(0,19),(0,22)\}$ (L0)	61% [51, 70]	0%	39%
single $\{(0,17)\}$ (L0)	62% [52, 71]	0%	38%
top-Wu $\{(12,7),(18,0)\}$ (L12, L18)	14% [8, 21]	0%	86%
random $\{(2,27),(24,24)\}$ (L2, L24)	10% [5, 16]	0%	90%

The match-A rate for the four layer-0 rows (61–85%) is comparable to or higher than the protected-set=softmax intervention at $n = 100$ (∼81% for greedy, ∼87% for alt-1, ∼62% for single-$(0, 17)$) and far above the all-polynomial baseline (14%). Even the single shared head $(0, 17)$ alone, patched from a different prompt, gives match-A 62% — identical to its softmax-protected single-head score. Donor-target match is 0% across all six rows: the protected set never causes the model to retrieve the donor's needle.

Two negative controls drop to the all-polynomial floor:

top-Wu (middle-layer structural retrieval heads) gives match-A 14% $[8, 21]$ and random-$K = 2$ (layers 2, 24) gives 10% $[5, 16]$. Patching from these heads' donor outputs is no better than running all-polynomial. Paired bootstrap (10K replicates) gives greedy − top-Wu = +68 pp $[+58, +77]$ and greedy − random = +72 pp $[+63, +81]$, with permutation $p < 10^{-4}$ in both cases. Prompt-independence is therefore a property specific to the layer-0 amplifier/causally important heads, not a generic feature of any protected-set patching.

Same-target sanity check. The strongest direct test of whether the protected set is doing actual retrieval is a same-target donor: if the protected set carried the answer, patching a donor with the same target as the test prompt should boost match-A relative to a different-target donor. We re-ran greedy protected-set path patching with donors picked from the same-target pool (when one exists in the $n = 100$ sample; otherwise a different-target fallback). Same-target match-A is 79% $[71, 87]$ vs. different-target 82% $[74, 89]$, a paired-bootstrap difference of -3 pp $[-11, +5]$ (not significant). The protected-set signal is therefore answer-independent: knowing the test's target token does not improve patching efficacy.

The conclusion is decisive: the layer-0 protected set's softmax output is essentially prompt-independent within this task. Cross-prompt patching preserves the test prompt's retrieval, so the retrieval signal that downstream layers use does not flow through the protected set.

BA. A constructive proposal: row-centered polynomial attention

The non-equivariance theorem (Theorem N.11) and the cross-degree output-rank analysis (above) together identify the central failure mode of fixed-bias polynomial substitution: the operator's behavior depends on the absolute score level r_*, and the resulting protected-set output under polynomial rotates away from the original-softmax direction as p grows. We propose a designed operator that removes the level dependence directly, and prove two properties. We do not claim it solves the angular-drift problem identified empirically across $p \in \{2, 4, 6, 8\}$ — that remains to be tested — but it removes one identifiable contributor.

Definition. For score row $z \in \mathbb{R}^n$ and degree $p > 0$ and offset $\tau > 0$, the row-centered polynomial is

$$m(z) := \max_i z_i, \qquad C_{p,\tau}(z)_i := \frac{\big[\tau + z_i - m(z)\big]_+^p}{\sum_k \big[\tau + z_k - m(z)\big]_+^p}. \tag{12}$$

Proposition BA.1 (Shift-invariance of $C_{p,\tau}$). For all $c \in \mathbb{R}$ and all $z \in \mathbb{R}^n$, $C_{p,\tau}(z + c\mathbf{1}) = C_{p,\tau}(z)$.

Proof. $m(z + c\mathbf{1}) = m(z) + c$. For each coordinate i, the numerator argument $\tau + (z_i + c) - m(z + c\mathbf{1}) = \tau + z_i + c - m(z) - c = \tau + z_i - m(z)$ is unchanged by the shift, so $[\cdot]_+^p$ in the numerator is unchanged. The denominator is the sum of the same unchanged terms. Hence $C_{p,\tau}(z + c\mathbf{1}) = C_{p,\tau}(z)$. □ □

Proposition BA.2 (Level-free retrieval bound). Assume the target index $*$ is score-maximal so $z_* = m(z)$. Let each distractor gap be $\delta_j := z_* - z_j \geq 0$, and suppose at most A distractors have $\delta_j \leq \tau$ (the others contribute zero through the ReLU). If every active distractor satisfies $\delta_j \geq \Delta \in [0, \tau]$, then

$$C_{p,\tau}(z)_* \geq \frac{1}{1 + A\,(1 - \Delta/\tau)^p}. \tag{13}$$

Proof. Since $z_* = m(z)$, $[\tau + z_* - m(z)]_+^p = \tau^p$. For each distractor j, $[\tau + z_j - m(z)]_+^p = [\tau - \delta_j]_+^p$, which equals $\tau^p(1 - \delta_j/\tau)_+^p$. Therefore

$$C_{p,\tau}(z)_* = \frac{\tau^p}{\tau^p + \sum_{j \neq *} \tau^p (1 - \delta_j/\tau)_+^p} = \frac{1}{1 + \sum_{j \neq *} (1 - \delta_j/\tau)_+^p}.$$

With at most A active distractors and $\delta_j \geq \Delta$ on all active distractors, $(1 - \delta_j/\tau)_+^p \leq (1 - \Delta/\tau)^p$, so $\sum_{j \neq *}(1 - \delta_j/\tau)_+^p \leq A(1 - \Delta/\tau)^p$. Substituting gives the bound. □ □

Proposition BA.2 assumes the target is the score-argmax. On the empirical Mistral NIAH rows the task target is often not the score-argmax (BOS-sink rows place mass on position 0; see §4 and Theorem N.5), so we record the exact target-mass identity which is the relevant primitive for non-argmax targets.

Proposition BA.3 (Exact target-mass identity for $C_{p,\tau}$). Let $m(z) = \max_i z_i$ and $\eta_i \equiv (m(z) - z_i)/\tau$ be the normalised gap of key i to the row maximum. Then for any target index $*$,

$$C_{p,\tau}(z)_* \;=\; \frac{[1 - \eta_*]_+^p}{\sum_k [1 - \eta_k]_+^p}.$$

If $\eta_* \geq 1$ (target outside the active set) then $C_{p,\tau}(z)_* = 0$. If $\eta_* < 1$ then

$$C_{p,\tau}(z)_* \;=\; \frac{1}{1 + \sum_{j \neq *} \big(\frac{[1 - \eta_j]_+}{1 - \eta_*}\big)^p},$$

and $C_{p,\tau}(z)_* > 1/2$ iff $\sum_{j \neq *} \big([1 - \eta_j]_+/(1 - \eta_*)\big)^p < 1$.

Proof. By definition $\tau + z_i - m(z) = \tau(1 - \eta_i)$ and $[\tau(1 - \eta_i)]_+^p = \tau^p [1 - \eta_i]_+^p$. Substituting into the numerator and denominator of $C_{p,\tau}(z)_*$ and cancelling τ^p gives the first identity. If $\eta_* \geq 1$ then $[1 - \eta_*]_+ = 0$ and the numerator vanishes. Otherwise dividing numerator and denominator by $(1 - \eta_*)^p > 0$ gives the second form. The $> 1/2$ condition is algebraically equivalent to $1 + \sum_{j \neq *} \big([1 - \eta_j]_+/(1 - \eta_*)\big)^p < 2$. □ □

Proposition BA.3 explains the $\tau = 1$ collapse observed in the empirical probe below: when τ is small enough that $\eta_{\text{target}} \geq 1$ on most rows, the operator assigns zero mass to the target by construction. On Mistral, $\tau \geq 5$ keeps the task target inside the active set in essentially every measured row; for non-argmax targets this invokes the exact identity above rather than the argmax-only bound in Proposition BA.2.

The exact identity also pairs with a regime where $C_{p,\tau}$ is provably close to a temperature-scaled softmax, which we record next.

Theorem BA.4 (Asymptotic softmax-approximation of $C_{p,\tau}$). Let $\beta \equiv p/\tau$ and $\hat{z}_i \equiv z_i - m(z) \leq 0$. Suppose for some $\rho \in (0,1)$, $\max_i |\hat{z}_i| \leq \rho\tau$ (every entry is active). Define

$$\epsilon_i \;\equiv\; p\big[\log(1+\hat{z}_i/\tau) - \hat{z}_i/\tau\big], \qquad E \;\equiv\; \frac{p}{2(1-\rho)^2} \cdot \frac{M^2}{\tau^2}, \qquad M \equiv \max_i |\hat{z}_i|.$$

Then $|\epsilon_i| \leq E$ for every i, and

$$e^{-2E}\, s_i \;\leq\; C_{p,\tau}(z)_i \;\leq\; e^{2E}\, s_i, \quad \text{where} \quad s_i = \text{softmax}(\beta z)_i.$$

Hence $\|C_{p,\tau}(z) - \text{softmax}(\beta z)\|_1 \leq e^{2E} - 1 = O(\beta M^2/\tau)$ as $E \to 0$.

Proof. With every entry active, $(\tau + \hat{z}_i)^p = \tau^p (1 + \hat{z}_i/\tau)^p$. Writing $x_i = \hat{z}_i/\tau$, $|x_i| \leq \rho$, we have $p\log(1+x_i) = \beta\hat{z}_i + \epsilon_i$ where ϵ_i is defined above. Therefore $(\tau + \hat{z}_i)^p = \tau^p \exp(\beta\hat{z}_i + \epsilon_i)$, and after the τ^p cancellation $C_{p,\tau}(z)_i = s_i e^{\epsilon_i}/(\sum_j s_j e^{\epsilon_j})$. Taylor's theorem on $\log(1+x)$ with $|x| \leq \rho$ gives $|\log(1+x) - x| \leq x^2/(2(1-\rho)^2)$, hence $|\epsilon_i| \leq E$. Since each $e^{\epsilon_i} \in [e^{-E}, e^{E}]$, the denominator is bounded in $[e^{-E}, e^{E}]$ (its weights s_j sum to one), giving $e^{\epsilon_i}/(\sum_j s_j e^{\epsilon_j}) \in [e^{-2E}, e^{2E}]$. Multiplying by s_i gives the two-sided sandwich; summing $|C_{p,\tau}(z)_i - s_i| \leq s_i(e^{2E} - 1)$ over i yields the ℓ_1 bound. □ □

Theorem BA.4 explains why $C_{p,\tau}$ approaches a temperature-scaled softmax when the row lies inside the active set and the score spread is small relative to τ. Note $\beta = p/\tau$: $p = 4, \tau = 5$ gives $\beta = 0.8$, i.e. effective temperature $T = 1/\beta = 1.25$ (slightly hotter / flatter than unit softmax); $p = 4, \tau = 10$ gives $\beta = 0.4$ ($T = 2.5$, much hotter / flatter). Whether this approximation governs the empirical $C_{4,5}$ result on Mistral requires measuring active-row rates and M/τ quantiles on the actual NIAH rows; we therefore report those diagnostics alongside the τ-grid result below.

The active-set regime. Theorem BA.4 assumes every entry is in the active set ($\rho < 1$). Empirical PG-19 rows often have $M > \tau$, so some entries are inactive. We therefore record both an active-set restriction identity and a stronger truncated-softmax approximation bound that holds without the all-entries-active assumption.

Proposition BA.5 (Active-set restriction identity). Let $m(z) = \max_k z_k$ and $A_\tau(z) = \{i : z_i > m(z) - \tau\}$. Then

$$C_{p,\tau}(z)_i = \begin{cases} \dfrac{(\tau + z_i - m(z))^p}{\sum_{j \in A_\tau(z)} (\tau + z_j - m(z))^p}, & i \in A_\tau(z), \\ 0, & i \notin A_\tau(z). \end{cases}$$

Proof. By definition, $C_{p,\tau}(z)_i = [\tau + z_i - m(z)]_+^p / \sum_j [\tau + z_j - m(z)]_+^p$. The numerator $[\tau + z_i - m(z)]_+$ is positive iff $z_i > m(z) - \tau$, i.e. $i \in A_\tau$. For inactive i, the numerator is zero, so $C_{p,\tau}(z)_i = 0$. For active i, $[\tau + z_i - m(z)]_+ = \tau + z_i - m(z)$, and the denominator can be restricted to the active set, giving the claimed expression. □ □

Theorem BA.6 (Active-set truncated softmax approximation). Let $\beta = p/\tau$, $\hat{z}_i = z_i - m(z) \le 0$, and $A = A_\tau(z)$. Assume $\rho_A \equiv \max_{i \in A} |\hat{z}_i|/\tau < 1$ (the all-active condition of Theorem BA.4 need not hold for $i \notin A$). Let $s_i = \mathrm{softmax}(\beta z)_i$, $Z_A = \sum_{i \in A} s_i$, and $s_i^A = s_i / Z_A$ for $i \in A$, $s_i^A = 0$ otherwise. Let $M_A = \max_{i \in A} |\hat{z}_i|$ and

$$E_A = \frac{p\, M_A^2}{2(1 - \rho_A)^2 \tau^2}.$$

Then

$$\|C_{p,\tau}(z) - s^A\|_1 \le e^{2E_A} - 1, \qquad \|C_{p,\tau}(z) - \mathrm{softmax}(\beta z)\|_1 \le e^{2E_A} - 1 + 2(1 - Z_A).$$

Proof. For inactive i, $C_{p,\tau}(z)_i = s_i^A = 0$ by Proposition BA.5 and the definition of s^A. For active i, write $x_i = \hat{z}_i/\tau$, $|x_i| \le \rho_A$, and $\epsilon_i = p[\log(1 + x_i) - x_i]$. Then $(\tau + \hat{z}_i)^p = \tau^p \exp(\beta \hat{z}_i + \epsilon_i)$, so after the τ^p cancellation and restriction to A,

$$C_{p,\tau}(z)_i \;=\; \frac{s_i^A\, e^{\epsilon_i}}{\sum_{j \in A} s_j^A\, e^{\epsilon_j}}, \qquad i \in A.$$

By Taylor on $\log(1 + x)$ with $|x| \le \rho_A$, $|\epsilon_i| \le E_A$. The denominator $\sum_{j \in A} s_j^A e^{\epsilon_j} \in [e^{-E_A}, e^{E_A}]$ (its s^A-weights sum to one), so $e^{\epsilon_i} / \sum_j s_j^A e^{\epsilon_j} \in [e^{-2E_A}, e^{2E_A}]$. Multiplying by s_i^A and summing gives $\|C_{p,\tau}(z) - s^A\|_1 \le e^{2E_A} - 1$.

For the comparison to full softmax s, note $s_i^A - s_i = s_i(1/Z_A - 1)$ for $i \in A$ and $s_i^A - s_i = -s_i$ for $i \notin A$. Therefore $\|s^A - s\|_1 = (1/Z_A - 1) \sum_{i \in A} s_i + \sum_{i \notin A} s_i = (1 - Z_A) + (1 - Z_A) = 2(1 - Z_A)$. Triangle inequality gives the second bound. □ □

The truncated-softmax bound separates two distinct error sources: $e^{2E_A} - 1$ is the polynomial-vs-exponential curvature error inside the active set, and $2(1 - Z_A)$ is the truncation error from inactive entries that $C_{p,\tau}$ assigns zero mass but full softmax does not. This matches the empirical regime where many rows have $M > \tau$ (so the all-active condition of Theorem BA.4 fails), but the active set still captures most of the softmax mass. These row-level quantities make the truncation source visible, but they do not by themselves predict behavioral valley width; Appendix BF adds value-scale diagnostics and treats the behavioral ranking as multi-factorial. We measure $|A_\tau|$, Z_A, E_A, and the empirical ℓ_1 distances in the active-set diagnostic table below.

Proposition BA.7 (Answer-direction active-set decomposition). Use the notation of Theorem BA.6: $a = C_{p,\tau}(z)$, $s = \mathrm{softmax}(\beta z)$ with $\beta = p/\tau$, active set $A = A_\tau(z)$, softmax active mass $Z_A = \sum_{i \in A} s_i$, and active-renormalized softmax $s_i^A = s_i / Z_A$ on A and 0 outside. Let v_i be the value vector for key i in a fixed head, let u be a downstream or answer readout direction, and set $g_i = u^\top v_i$. Define

$$\gamma_{\mathrm{sm}} = \sum_i s_i g_i, \qquad \gamma_C = \sum_i a_i g_i, \qquad T_u = \sum_{i \notin A} s_i g_i, \qquad G_A = \max_{i \in A} |g_i|.$$

Then

$$\gamma_C = \frac{\gamma_{\mathrm{sm}} - T_u}{Z_A} + R, \qquad |R| \le \big(e^{2E_A} - 1\big) G_A.$$

Proof. Since a is supported on A,

$$\gamma_C = \sum_{i\in A} a_i g_i = \sum_{i\in A} s_i^A g_i + \sum_{i\in A}(a_i - s_i^A) g_i.$$

The first term equals $Z_A^{-1}\sum_{i\in A} s_i g_i = Z_A^{-1}(\gamma_{\mathrm{sm}} - T_u)$. For the remainder,

$$|R| \le \sum_{i\in A} |a_i - s_i^A|\,|g_i| \le G_A \|a - s^A\|_1 \le \big(e^{2E_A} - 1\big) G_A,$$

where the final inequality is Theorem BA.6. □ □

Proposition BA.7 is the signed analogue of the value-tail norm bound. It says that ROWAN-(p,τ) keeps the active-set part of the temperature-softmax answer signal, renormalizes it by $1/Z_A$, and omits the signed tail contribution T_u, up to curvature error. If T_u is small and the curvature term is small, the answer-direction signal is preserved (or mildly amplified by active-set renormalization). If the omitted tail carries most of the positive answer-direction evidence, finite support fails even when the active-set approximation itself is tight.

Corollary BA.8 (Finite-support answer-direction obstruction). In the setting of Proposition BA.7, suppose $\gamma_{\mathrm{sm}} > 0$ but $\max_{i\in A} g_i \le 0$. Then the corresponding ROWAN answer signal, and indeed any nonnegative reweighting supported on A, satisfies

$$\gamma_C = \sum_{i\in A} a_i g_i \le 0 < \gamma_{\mathrm{sm}}.$$

Proof. The weights a_i are nonnegative and sum to one over A, so γ_C is a convex combination of $\{g_i : i \in A\}$ and is at most their maximum. The assumed maximum is nonpositive, while the full-support temperature-softmax signal is positive. □ □

This corollary makes the NoLiMa-style boundary structural: if answer-positive evidence lives outside the score-gap active set, changing the polynomial degree inside the fixed active set cannot recover it. The relevant diagnostic is therefore not only inactive-tail norm, but the signed answer-direction tail T_u when an answer or downstream readout direction is available.

Observation BA.9 (Threshold-retrieval equivalence). Consider a row-level retrieval proxy in which a normalizer π is counted as successful on prompt r if the target index j_r receives mass at least γ: $\pi(z^{(r)})_{j_r} \ge \gamma$. If two normalizers π_1, π_2 both satisfy this inequality for every prompt r in a suite, they are indistinguishable under that row-level threshold proxy regardless of the rest of the mass distribution.

The observation is essentially tautological but pins down why NIAH cannot separate $C_{4,5}$ from softmax$(0.8z)$: both assign above-threshold mass to the retrieval-relevant key on the diagnostic rows that drive answer recovery. (We do not claim the retrieval-relevant key is the score argmax in every head-row triple; the appendix shows the answer key is often not the score argmax on Mistral BOS-sink rows.) PG-19 perplexity and NoLiMa retrieval do not reduce to a single-key threshold and can therefore separate the two operators, as observed empirically (§5).

Proposition BA.10 (Value-weighted active-set perturbation). Let $a = C_{p,\tau}(z)$, $s = \mathrm{softmax}(\beta z)$ with $\beta = p/\tau$, and use the notation of Theorem BA.6: $A = A_\tau(z)$, $Z_A = \sum_{i\in A} s_i$, $s_i^A = s_i/Z_A$ for $i \in A$ and $s_i^A = 0$ otherwise, $M_A = \max_{i\in A} |\hat{z}_i|$, and $E_A = pM_A^2/(2(1-\rho_A)^2\tau^2)$. Let $v_i \in \mathbb{R}^d$ be the per-key value vectors at the head under consideration, and write $y_C = \sum_i a_i v_i$ and $y_s = \sum_i s_i v_i$ for the head outputs under the two normalizers. If $\|v_i\|_2 \le B$ for every i, then

$$\|y_C - y_s\|_2 \;\le\; B\big(e^{2E_A} - 1\big) + 2B(1 - Z_A).$$

Proof. Decompose $y_C - y_s = \sum_i (a_i - s_i^A) v_i + \sum_i (s_i^A - s_i) v_i$. By the triangle inequality $\|y_C - y_s\|_2 \le \sum_i |a_i - s_i^A|\,\|v_i\|_2 + \sum_i |s_i^A - s_i|\,\|v_i\|_2 \le B\,\|a - s^A\|_1 + B\,\|s^A - s\|_1$. The first norm is bounded by $e^{2E_A} - 1$ and the second by $2(1 - Z_A)$ via Theorem BA.6. □ □

Proposition BA.10 sharpens the row-level L^1 bound to the actual head output, which is what downstream layers consume. The truncation term $2B(1 - Z_A)$ is a worst-case bound on the full difference between the active-set-renormalized softmax s^A and the full-support softmax s. One causally important component of that difference is

the inactive softmax tail $t_{\bar{A}} = \sum_{i \notin A} s_i v_i$, whose crude bound $B(1 - Z_A)$ can be much smaller when the inactive values cancel in direction. The second-moment result below diagnoses this omitted-tail component only; it does not replace the full $2B(1 - Z_A)$ term because active-set renormalization also moves the attention output. We measure the inactive-tail norm in Appendix BF for the four models in the PG-19 panel; it provides the per-model leverage missing from the M/τ-only test of the active-set theorem (§5 Testing the active-set prediction.).

Corollary BA.11 (Second-moment inactive-tail bound). Adopt the notation of Proposition BA.10. Let $t_{\bar{A}} \equiv \sum_{i \notin A} s_i v_i$ be the inactive-tail value output, and define the inactive softmax-weighted second-moment matrix

$$\Sigma_{\bar{A}} \;\equiv\; \sum_{i \notin A} s_i \, v_i v_i^{\top}.$$

Then

$$\|t_{\bar{A}}\|_2 \;\le\; \sqrt{(1 - Z_A)\, \sigma_{\max}(\Sigma_{\bar{A}})}\,,$$

where $\sigma_{\max}(\Sigma_{\bar{A}})$ is the largest eigenvalue of $\Sigma_{\bar{A}}$. Moreover this bound is always at least as tight as the crude inactive-tail bound $B(1 - Z_A)$ implied by $\|v_i\|_2 \le B$.

Proof. If $1{-}Z_A = 0$ the result is trivial. Otherwise let $w = 1{-}Z_A > 0$ and define the inactive probability distribution $\tilde{s}_i = s_i/w$ for $i \notin A$. Then $t_{\bar{A}} = w\, \mathbb{E}_{\tilde{s}}[v]$. By the Cauchy–Schwarz inequality applied to the expectation,

$$\|\mathbb{E}_{\tilde{s}}[v]\|_2^2 \;=\; \sup_{\|u\|_2=1} \big(u^{\top} \mathbb{E}_{\tilde{s}}[v]\big)^2 \;\le\; \sup_{\|u\|_2=1} \mathbb{E}_{\tilde{s}}\big[(u^{\top} v)^2\big] \;=\; \sigma_{\max}\big(\mathbb{E}_{\tilde{s}}[vv^{\top}]\big).$$

Since $\mathbb{E}_{\tilde{s}}[vv^{\top}] = \Sigma_{\bar{A}}/w$, we obtain $\|\mathbb{E}_{\tilde{s}}[v]\|_2 \le \sqrt{\sigma_{\max}(\Sigma_{\bar{A}})/w}$, and hence $\|t_{\bar{A}}\|_2 = w\, \|\mathbb{E}_{\tilde{s}}[v]\|_2 \le \sqrt{w\, \sigma_{\max}(\Sigma_{\bar{A}})} = \sqrt{(1{-}Z_A)\, \sigma_{\max}(\Sigma_{\bar{A}})}$. For the dominance claim, $\sigma_{\max}(\Sigma_{\bar{A}}) \le \mathrm{tr}(\Sigma_{\bar{A}}) = \sum_{i \notin A} s_i \|v_i\|_2^2 \le (1 - Z_A)B^2$, so $\sqrt{(1{-}Z_A)\sigma_{\max}(\Sigma_{\bar{A}})} \le B(1{-}Z_A)$. □ □

The corollary explains why the empirical $\|t_{\bar{A}}\|_2$ in Appendix BF is much smaller than its crude inactive-tail bound: when the inactive v_i span many directions, $\sigma_{\max}(\Sigma_{\bar{A}}) \ll \mathrm{tr}(\Sigma_{\bar{A}})$ and the bound shrinks by up to a $\sqrt{d}$ factor relative to $B(1 - Z_A)$. We report the second-moment inactive-tail bound side-by-side with the full first-moment truncation bound $2B(1 - Z_A)$ in the omitted-tail value diagnostic table, because the latter still controls active-set renormalization as well as tail omission.

Comparison with fixed-bias polynomial. The recipe-matched concentration bound for fixed-bias $P_{p,b}$ in Proposition 2.2 reads

$$P_{p,b}(z)_* \ge \frac{1}{1 + A(1 - \Delta/r_*)^p},$$

where $r_* = [z_* + b]_+$ is the post-ReLU absolute level of the target score. The structural difference is that r_* varies prompt-by-prompt with the absolute score level (this is the level dependence that drives the non-equivariance gap of Theorem N.11), whereas τ in Proposition BA.2 is an operator hyperparameter fixed once at calibration. $C_{p,\tau}$ therefore has a structural advantage on the level-dependence axis: the bound on target-token concentration depends only on Δ/τ, a ratio entirely controllable by the operator designer. Whether $C_{p,\tau}$ also avoids the angular drift across p that drives the protected set-rescue curve (Appendix AY, R from 0.98 at $p = 2$ to 0.19 at $p = 8$ with $\sigma_1 \approx 3.0$ constant) is an empirical question not settled by these propositions; we leave the comparison test to follow-up.

Calibration sketch. A natural calibration for τ is to match the softmax's median target-token concentration on a held-out NIAH set: choose τ such that $\mathbb{E}_z[C_{p,\tau}(z)_*]$ matches $\mathbb{E}_z[\mathrm{softmax}(z)_*]$ on the calibration set. Alternatively, choose τ to match softmax's entropy on the same set. In practice a small τ-grid suffices to bracket the working regime on Mistral; the empirical probe below uses such a grid.

Empirical probe on Mistral NIAH ($n = 500$). We evaluate $C_{p,\tau}$ at $p = 4$ on the same Mistral-7B-Instruct-v0.3 NIAH suite used throughout the paper ($n = 500$, pad 128), without any protected-set intervention, against softmax and the canonical fixed-bias polynomial $P_{4,0}$.

operator	NIAH acc. ($n=500$)	R vs. canonical
softmax	91.6%	—
$P_{4,0}$ (canonical poly)	12.6%	—
$C_{4,1}$	3.6%	-0.11
$C_{4,3}$	88.4%	0.96
$C_{4,5}$	**91.6%**	**1.00**
$C_{4,10}$	**91.6%**	**1.00**

The rescue ratio is computed against $P_{4,0}$ damage as $R=(\mathrm{Acc}_C-\mathrm{Acc}_{P_{4,0}})/(\mathrm{Acc}_{\mathrm{sm}}-\mathrm{Acc}_{P_{4,0}})$. At $\tau\in\{5,10\}$, $C_{4,\tau}$ matches the softmax NIAH accuracy exactly at $n=500$; at $\tau=3$ the operator recovers 96% of softmax retrieval. At $\tau=1$, by contrast, the active set is too narrow: on many diagnostic rows the retrieval-relevant key is outside the score-gap active set and therefore receives zero mass by Proposition BA.3. We read this as a calibration boundary, not a refutation of the proposition or the centered construction.

The empirical evidence supporting $C_{p,\tau}$ as a level-dependence-removing operator beyond this single τ-grid (cross-model NIAH on twelve checkpoints, multi-seed PG-19 preservation on five families, NoLiMa partial preservation on Mistral and Llama-3.1-8B, MMLU-STEM downstream rescue, softmax(βz) temperature-baseline comparisons, active-set diagnostics, and the one-pass amplifier screen) is given in §5. This appendix retains only the formal apparatus (proofs and identities used by the main text).

BB. Theory verification: detailed tables

This appendix collects the three detailed tables summarized in the body §5 Theory verification paragraph: the full τ-grid PG-19 sweep on five families, the $\beta = 0.8$ invariance test on Mistral, and the active-set M-prediction side table.

Full PG-19 τ grid.

	softmax PPL	$C_{4,3}$ PPL (ratio)	$C_{4,5}$ PPL (ratio)	$C_{4,10}$ PPL (ratio)
Mistral-7B-Inst	6.67	10.48 (1.57×)	7.04 (1.06×)	7.70 (1.15×)
OLMo-2-0425-1B	9.64	18.65 (1.93×)	10.79 (1.12×)	53.18 (5.51×)
Llama-3.1-8B-Inst	8.14	14.30 (1.76×)	8.86 (1.09×)	10.53 (1.29×)
Qwen3-1.7B	21.44	47.81 (2.21×)	22.79 (1.06×)	35.58 (1.66×)
Granite-3.3-8B-Inst	7.65	12.36 (1.60×)	7.92 (1.04×)	11.81 (1.54×)

All five models share their optimum at $\tau = 5$ consistent with the active-set theorem (Theorem BA.4): τ too small leaves many rows outside the active set (target receives zero mass per Proposition BA.3) and τ too large flattens the distribution. Preservation valley widths vary by model (Llama broad, OLMo-2 narrow).

$\beta = 0.8$ invariance on Mistral, NIAH and PG-19. The three centered operators $(p, \tau) \in \{(2, 2.5), (4, 5), (8, 10)\}$ all have $\beta = p/\tau = 0.8$.

operator	softmax / softmax$_{0.8}$	$C_{2,2.5}$	$C_{4,5}$	$C_{8,10}$
NIAH acc. ($n=500$)	91.6% / —	90.4%	91.6%	91.6%
PG-19 ratio (5 seeds, mean±std)	1.000 / 1.018±0.001	1.392±0.065	1.068±0.009	**1.012±0.002**

NIAH is β-invariant (Observation BA.9); PG-19 is not but converges to softmax$_{0.8}$ as p grows, matching the theoretical prediction that both the curvature error $E_A = M^2\beta^2/(2p(1-\rho_A)^2)$ and the truncation error $2(1-Z_A)$ shrink as p increases at fixed β (since $\tau = p/\beta$ grows with p).

Active-set PG-19 prediction. The active-set theorem (Theorem BA.4) predicts the approximation error to softmax(βz) scales as $\propto p\, M^2/\tau^2$ with $M = \max_i |z_i - m(z)|$ on a row.

model	M_{med}	M_{p95}	$C_{4,10}$ valley
Mistral-7B-Inst	13.09	18.62	1.15×
OLMo-2-0425-1B	12.00	18.44	5.51×
Llama-3.1-8B-Inst	13.62	19.16	1.29×
Qwen3-1.7B	14.59	21.08	—

Score magnitudes are tightly clustered ($M_{p95} \in [18.4, 21.1]$, $\sim$14% spread); the M^2/τ^2 scaling predicts $\sim$1.3× spread in valley widths but the observed spread is 4.8×. The active-set bound is necessary but not sufficient. Appendix BF reports a per-model V-scale diagnostic: it tightens the inactive-tail bound substantially but does not by itself predict the behavioral valley-width ranking, which remains multi-factorial.

BC. NoLiMa-main and MMLU-STEM support tables

Table 58. NoLiMa-main operator table behind the main-text finite-support boundary discussion. Values are substring accuracy on $n = 200$ prompts. The Mistral row uses a shared split for all displayed operators; the Llama row uses its own $n = 200$ draw.

Model	softmax	$P_{4,0}$	$P_{4,b_{\text{auto}}}$	$C_{4,5}$	softmax$_{0.8}$	softmax$_{1.2}$	sparsemax	entmax-1.5
Mistral-7B-Inst	74.5	0.5	8.0	55.0	65.5	71.5	0.0	0.0
Llama-3.1-8B-Inst	83.0	0.0	3.5	22.0	59.0	60.0	0.0	2.0

The decision-relevant NoLiMa comparisons are directional but not single-prompt claims. At $n = 200$, Wilson 95% intervals for the key cells are: Mistral ROWAN-$(4, 5)$ 55.0% $[48.1, 61.7]$, softmax$_{0.8}$ 65.5% $[58.7, 71.7]$, softmax$_{1.2}$ 71.5% $[64.9, 77.3]$; Llama ROWAN-$(4, 5)$ 22.0% $[16.8, 28.2]$, softmax$_{0.8}$ 59.0% $[52.1, 65.6]$, softmax$_{1.2}$ 60.0% $[53.1, 66.5]$.

Table 59. MMLU-STEM operator table behind the main-text downstream-preservation discussion. Accuracies are over 18 STEM subjects × 50 questions. Rescue ratios use $R = (\text{Acc}_{\text{op}} - \text{Acc}_{P_{4,0}})/(\text{Acc}_{\text{sm}} - \text{Acc}_{P_{4,0}})$.

Model	softmax	$P_{4,0}$	$C_{4,5}$	softmax$_{0.8}$	sparsemax	entmax-1.5	R_C	$R_{\text{softmax}_{0.8}}$
Mistral-7B-Inst	52.6	34.3	48.4	51.3	25.4	28.9	0.77	0.93
OLMo-2-0425-1B	38.2	23.9	36.2	31.8	24.4	25.0	0.86	0.55
Llama-3.1-8B-Inst	59.7	24.3	55.3	59.3	23.1	30.2	0.88	0.99
Qwen3-1.7B	60.7	28.0	56.8	59.8	25.4	31.7	0.88	0.97
Granite-3.3-8B-Inst	58.2	26.7	54.3	54.8	25.8	32.9	0.88	0.89

Table 60. Transfer boundary for the Mistral protected pair $\{(0, 17), (0, 28)\}$. "Pair protected" keeps exactly this pair under softmax while all other heads use $P_{4,0}$. The pair is a NIAH-specific compatible protected set, not a general downstream repair.

Transfer test	softmax	all-$P_{4,0}$	pair protected	reading
NoLiMa-main ($n = 200$)	74.5	0.5	0.0	no transfer
MMLU-STEM ($n = 900$)	52.6	34.3	32.8	no transfer

BD. NoLiMa-Hard full table

NoLiMa-Hard subset, $n = 200$, substring scoring on 4 operators (cohort_softmax not run on NoLiMa-Hard).

model (NoLiMa-Hard $n=200$)	softmax	$P_{4,0}$	$C_{4,5}$	softmax$_{0.8}$
Mistral-7B-Inst	45.0%	0.0%	34.0%	**41.5**%
Llama-3.1-8B-Inst	58.5%	0.0%	7.0%	**30.0**%

Rescue ratios: Mistral $R(C_{4,5}) = 0.76$, $R(\text{softmax}_{0.8}) = 0.92$; Llama-3.1-8B $R(C_{4,5}) = 0.12$ (near-collapse), $R(\text{softmax}_{0.8}) = 0.51$.

BE. Amplifier-screen robustness

Appendix BE ranks Mistral-7B-Inst heads by $\text{AmpScore}(h) = \text{median}_{\text{prompt}}[(1 - \delta^{(h)})/(\varepsilon^{(h)} + \varepsilon_0)]$ with $\varepsilon_0 = 10^{-3}$. The score formula and $\varepsilon_0 = 10^{-3}$

were fixed before inspecting the head ranks; we then re-aggregated the same n=50 NIAH parquet under four ablations: replace median by mean, drop the curvature term (low-ε only), drop the diffuseness term (high-(1–δ) only), and sweep $\varepsilon_0 \in \{0, 10^{-4}, 10^{-3}, 10^{-2}\}$. Each row reports the rank of the five layer-0 heads of interest (MSC $\{(0,17),(0,28)\}$, alt-1 $(0,22)$, same-KV-group $(0,19)$, distractor $(0,12)$) and Recall@K for $K \in \{5, 10, 50\}$ on the MSC pair.

score variant	rank $(0,17)$	rank $(0,28)$	rank $(0,22)$	rank $(0,19)$	R@5	R@10	R@50
median AmpScore (fixed screen)	**1**	**6**	14	3	1/2	**2/2**	**2/2**
mean AmpScore	**1**	**6**	13	2	1/2	**2/2**	**2/2**
low-ε-only ($1/\varepsilon_{\text{med}}$)	2	33	85	1	1/2	1/2	2/2
high-$(1-\delta)$-only ($1-\delta_{\text{med}}$)	247	456	249	974	0/2	0/2	0/2
$\varepsilon_0 = 0$	**1**	7	18	2	1/2	**2/2**	**2/2**
$\varepsilon_0 = 10^{-4}$	**1**	7	18	2	1/2	**2/2**	**2/2**
$\varepsilon_0 = 10^{-3}$ (fixed screen)	**1**	7	18	3	1/2	**2/2**	**2/2**
$\varepsilon_0 = 10^{-2}$	**1**	7	18	4	1/2	**2/2**	**2/2**

Three observations. First, median vs mean is irrelevant: both surface $(0,17)$ at rank 1 and $(0,28)$ at rank 6. Second, both ingredients are necessary: dropping the diffuseness term (high-$(1-\delta)$-only) destroys the screen entirely (both MSC heads beyond rank 200), while dropping the curvature term (low-ε-only) keeps $(0,17)$ at rank 2 but pushes $(0,28)$ to rank 33. The ratio formula is what makes the screen converge on amplifier-class heads rather than retrieval heads or low-mass sinks. Third, the ε_0 choice is non-critical: ranks shift by at most ± 1 across four decades of ε_0. The fixed (median, $\varepsilon_0 = 10^{-3}$) choice is therefore not a chance hit.

BF. Omitted-tail value diagnostic

Proposition BA.10 bounds $\|y_C - y_s\|_2 \leq B(e^{2E_A} - 1) + 2B(1 - Z_A)$ with $B = \max_i \|v_i\|_2$. The truncation term $2B(1 - Z_A)$ is a worst-case estimate for the full active-set-renormalization difference. Corollary BA.11 instead refines the inactive-tail component $t_{\bar{A}} = \sum_{i \notin A} s_i v_i$ to $\sqrt{(1-Z_A)\sigma_{\max}(\Sigma_{\bar{A}})}$, where $\Sigma_{\bar{A}} = \sum_{i \notin A} s_i v_i v_i^\top$ is the inactive softmax-weighted second-moment matrix. We extend the active-set diagnostic to capture $\|t_{\bar{A}}\|_2$, B, and $\sigma_{\max}(\Sigma_{\bar{A}})$ per row (the last via 3-step power iteration), then aggregate medians across the same $N = 8$ PG-19 batches at seq_len = 1024:

model	τ	B	empirical $\|t_{\bar{A}}\|_2$	inactive 1st $B(1-Z_A)$	full 1st $2B(1-Z_A)$	inactive 2nd moment
Mistral-7B-Inst	3	7.73	0.079	0.33	0.65	**0.09**
	5	7.73	0.249	1.49	2.98	**0.29**
	10	7.73	0.072	0.41	0.81	**0.08**
OLMo-2-0425-1B	3	18.35	0.119	1.28	2.56	**0.22**
	5	18.35	0.238	3.64	7.27	**0.35**
	10	18.35	0.026	0.28	0.56	**0.03**
Llama-3.1-8B-Inst	3	5.33	0.055	0.19	0.38	**0.06**
	5	5.33	0.235	1.14	2.27	**0.26**
	10	5.33	0.096	0.48	0.96	**0.10**
Qwen3-1.7B	3	56.31	0.502	2.71	5.41	**0.62**
	5	56.31	1.458	10.04	20.08	**1.92**
	10	56.31	0.431	4.02	8.03	**0.45**

Three findings. First, the empirical inactive-tail norm $\|t_{\bar{A}}\|_2$ is roughly 2.5–15× smaller than the matched inactive-tail first-moment bound $B(1 - Z_A)$ (equivalently 5–31× smaller than the full truncation term $2B(1 - Z_A)$, which also bounds active-set renormalization); the inactive v_i partially cancel rather than aligning to a single direction. Second, the inactive-tail second-moment bound $\sqrt{(1-Z_A)\sigma_{\max}(\Sigma_{\bar{A}})}$ tracks the empirical inactive-tail norm within 1.04–1.85× across all cells; the largest slack occurs on OLMo-2 at $\tau \in \{3, 5\}$ (ratios ≈ 1.85, 1.47). Third, B varies by $\sim 10\times$ across the panel (5.3 to 56.3) and dominates the cross-model spread of the analytic bound far more than $1 - Z_A$ does. This is a statement about the bound, not about behavioral valley width: B does not predict the empirical PG-19 valley ranking (Qwen3 has the largest $B = 56.3$ but a smaller $\tau = 10$ valley than OLMo-2, while OLMo-2 has the worst valley with $B = 18.4$). Behavioral valley width is therefore multi-factorial—set jointly by V-scale, the curvature term $e^{2E_A} - 1$, active-set renormalization, and downstream tolerance—and is not predicted by any single measured row- or value-statistic in isolation. We report the second-moment bound as a tight characterization of the inactive-tail component only, and leave a predictive per-model valley-width model to follow-up.