EXTRAPOLATION COLLAPSE IS A READOUT FAILURE: LONG-CONTEXT MODELS RETRIEVE WHAT THEY CANNOT SAY

Anonymous authors
Paper under double-blind review

ABSTRACT

Interpolation methods such as YaRN let a pretrained model be evaluated far beyond its training window, and they are judged by end-task score. By that measure they collapse: YaRN-extended Qwen3-8B falls from 0.51 at 64K to 0.23 at 128K on a controlled retrieval task. We show that the retrieval computation largely survives this collapse and that the failure lies in the readout. Letting 20 frozen retrieval heads vote over the parsed candidate set, without gold labels and without changing any weight, recovers the correct answer in 71 to 96% of the samples the model gets wrong, at every extension depth we measure, and never overturns a correct one. We call this gap the *readout bottleneck*, and it breaks behavior-only evaluation: at 128K one checkpoint scores 0.00 while locating the answer in 92% of its failures, and the behavioral ranking of two checkpoints inverts relative to their retrieval ability. Cutting the output path into four stages of decreasing dependence on the token distribution places the damage after retrieval rather than in it, and the damage itself is confined to the frequency bands the extension method rewrites, which we predict analytically before measuring. Its mechanism is deanchoring to filler text rather than confusion among candidates. The bottleneck also bounds what interventions can do: steering the distribution with the same attention signal, including a published method run under our protocol, moves behavior by at most 0.06 where reading the trace moves it by 0.4 to 0.5. These results invite a reconsideration of what long-context benchmarks measure.

1 INTRODUCTION

A long-context language model is evaluated by what it emits. Needle probes (Kamradt, 2023), RULER (Hsieh et al., 2024), InfiniteBench (Zhang et al., 2024) and LongBench-v2 (Bai et al., 2025) all score the generated answer, and so do the methods that stretch a model's window past its training length (Chen et al., 2023; Peng et al., 2024; Ding et al., 2024). By that measure, interpolation-based extension fails in a well-documented way: accuracy decays as the extension ratio grows. YaRN-extended Qwen3-8B scores 0.51 at 64K and 0.23 at 128K on a controlled retrieval task, and Mistral-7B falls from 0.27 to 0.08.

This paper asks what the decay is made of, and finds that the emitted answer is the wrong place to look. Take the same model, the same prompt and the same decoding, and read the *attention trace* instead. We let 20 frozen retrieval heads (Wu et al., 2025) vote at every decoding step for the candidate that holds their maximal attention logit. On the samples the model answers incorrectly, that vote recovers the correct answer 71 to 96% of the time, at every extension depth we measure, and it overturns no correct answer. The retrieval computation is largely intact. What extrapolation breaks is the path from that computation to the tokens.

The consequence is not only mechanistic. Because behavior and retrieval come apart, a benchmark that scores only behavior mis-ranks models. Two of our checkpoints score identically at 64K (0.26) yet differ by 26 points of trace accuracy (0.98 against 0.72). At 128K the ranking inverts: a checkpoint that scores 0.00 locates the answer in 92% of its failures, while one that scores 0.08 locates it in 54%. Read behaviorally, the first model has no long-context ability. Read mechanistically, it is the better retriever.

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

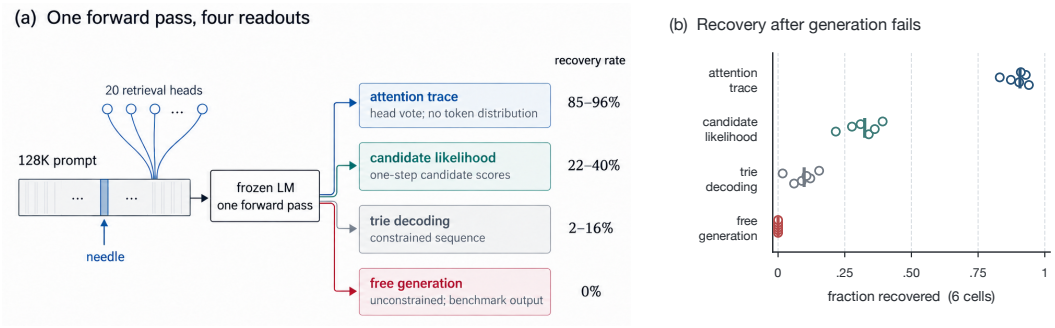

Figure 1: One forward pass, four places to read the answer from. Left: the retrieval heads concentrate on the needle, and four readouts take the answer from that same forward pass. Right: the fraction of behaviorally failed samples each readout recovers, one circle per measured cell and a tick at the median. Free generation recovers nothing, and it is what the benchmark scores. The relative order of the two middle readouts is not universal (§4).

To make "readout failure" measurable we cut the output path into four stages, each removing one degree of freedom from the previous one: free generation, generation constrained to a trie over the candidate strings, whole-string candidate likelihood, and the attention trace (Figure 1). On the same failed samples, recovery rises from nothing for free generation to 85 to 96% for the trace (Figure 1), and every step of the path is significant under a paired sign test. Every added reliance on the token distribution costs answers, and the largest single loss occurs between using that distribution and not using it.

Two further results locate the damage and bound the claim. The margin deficit is confined to the frequency bands the extension method rewrites, which we predict analytically before measuring, and the mechanism is de-anchoring to filler text: competing keys take under 1% of the drifted attention, and 82% of wrong answers are well-formed strings that appear nowhere in the document. Ordering three extension schedules by how much of the spectrum they rewrite also orders the damage to the trace, and under the most aggressive schedule every stage collapses together, so the readout is not an oracle.

Contributions.

- A four-stage decomposition of the output path that locates extrapolation damage in the readout, with a monotone recovery gradient reproduced in six measured cells (§4).
- A measurement consequence: behavior-only evaluation mis-ranks checkpoints, including a strict inversion at 128K (§3).
- A frequency-localized account of the damage, predicted before measurement, with four competing explanations eliminated (§5).
- The boundaries of the account: three applicability conditions, a published steering method run under our protocol, and four negative results on in-place weight repair (§6).

Section 7 positions this work against the position-scaling literature, the retrieval-head literature, and inference-time steering.

2 Setup

Rotary embeddings and bands. RoPE (Su et al., 2024) rotates queries and keys by an angle proportional to absolute position, with frequencies $\theta_j = b^{-2j/d_h}$ over rotary sub-spaces j and $b = 10^6$ for both models, so attention logits depend on position only through the offset. Let L be the training window, 32,768 here, and $\lambda_j = 2\pi/\theta_j$. We index sub-spaces by the rotations they complete inside that window, $r_j = L/\lambda_j$, and call a contiguous range of r_j a *band*.

checkpoint (s=4, 128K)	behavior	trace accuracy	trace recall on failures
Qwen3-8B	0.23	0.94	92.2%
Qwen3-8B-Base	**0.00**	**0.92**	92.0%
Qwen3-4B	0.08	0.58	54.3%
Mistral-7B	0.08	0.86	84.8%
Qwen3-14B	0.54	**1.00**	100%

Table 1: Behavior against trace at s=4, same samples, prompts and decoding. *Behavior*: RULER score of the generated answer. *Trace accuracy*: vote correctness over all samples. *Recall*: fraction of behaviorally wrong samples whose vote still names the needle. n=100 for the two 7–8B instruct models and n=50 for the others. At s=2 the same checkpoints score 0.51, 0.26, 0.26, 0.27 and 0.92 with trace accuracy 0.98, 0.98, 0.72, 0.92 and 1.00.

YaRN. For target length $L' = sL$, YaRN (Peng et al., 2024) interpolates each frequency by an amount depending on r_j,

$$\theta'_j = \frac{\theta_j}{(1-\gamma_j)s+\gamma_j}, \qquad \gamma_j = \mathrm{clip}\left(\frac{r_j-\alpha}{\beta-\alpha}, 0, 1\right), \tag{1}$$

leaving $r_j \geq \beta$ untouched and fully scaling $r_j \leq \alpha$, and divides logits by a temperature with $1/\sqrt{t} = 0.1 \ln s + 1$. We use the reference implementation with $\alpha = 1$ and $\beta = 32$, and $s = 1$ recovers unmodified RoPE, that is, direct extrapolation. Two properties matter below. The correction is a function of the sub-space index alone, shared by every head and layer, and once s is chosen it is constant in position.

Models, tasks, protocol. We study Qwen3-8B (Yang et al., 2025) and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), with Qwen3-8B-Base, Qwen3-4B and Qwen3-14B as additional checkpoints. All have a 32K training window. We evaluate YaRN at $s \in \{2, 3, 4\}$, matched to target lengths of 64K, 96K and 128K, and use unmodified RoPE (s=1) as a direct-extrapolation control. The controlled task is RULER (Hsieh et al., 2024) NIAH_MULTIKEY_3, which asks the model to retrieve a uuid associated with a query key among hundreds of distractor key-value lines. It is exact-match, its distractors make retrieval rather than length the difficulty, and its candidate set is enumerable, which is what makes trace-side measurement possible. Breadth is reported in the appendix: 13 RULER tasks, InfiniteBench (Zhang et al., 2024) KV and En.MC, LongBench-v2 (Bai et al., 2025), and two further schedules. All runs use greedy decoding, the model's chat template, official generation budgets, n=100 per cell unless stated, seed 42, and paired exact sign tests with Holm correction. The 20 retrieval heads per model are identified once on held-out data and frozen (§B.1).

3 Behavior and Retrieval Come Apart

Our sharpest case is the following. Extend Qwen3-8B to 64K with position interpolation and it answers none of our retrieval prompts correctly, a score of 0.000 over 50 samples. Read the attention trace of 20 frozen retrieval heads on those same prompts and it names the correct answer in 100% of them. Across four such cells, covering two models and two tasks, behavior is 0.000 throughout while the trace runs 0.80 to 1.00. Nothing about the model's retrieval is missing. What is missing is the path from retrieval to text.

This section makes the gap systematic. Table 1 reports, for five checkpoints, the score of the generated answer and what the model's attention was pointing at while it produced that answer. We obtain the second quantity by parsing the task's candidate set from the prompt and letting the 20 heads vote, at every decoding step, for the candidate that holds their maximal attention logit. The plurality winner is the trace's answer. The procedure never sees the gold label and changes no weight and no activation. Section 4 defines it in full.

The table supports two readings. First, the ranking induced by behavior is not the ranking induced by retrieval. At s=2, Qwen3-8B-Base and Qwen3-4B score identically (0.26) while their trace accuracies differ by 26 points (0.98 against 0.72). At s=4 the inversion is strict: behavior ranks Qwen3-4B (0.08) above Qwen3-8B-Base (0.00), while the trace ranks Qwen3-8B-Base (0.92) far

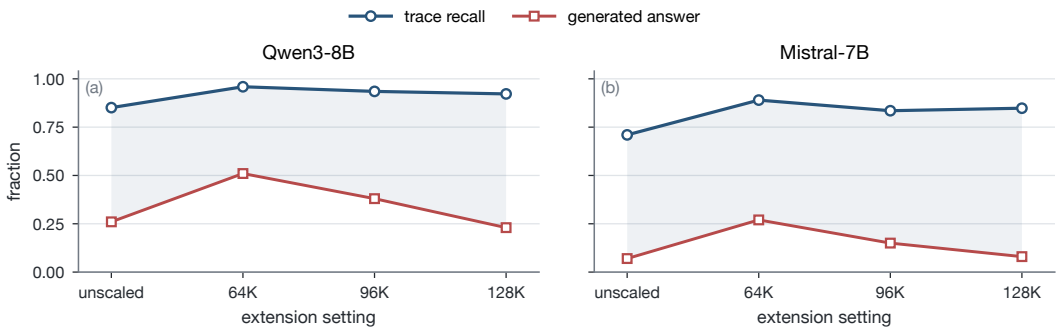

Figure 2: Behavior collapses with the extension ratio and the trace does not. *Trace recall* is measured on behaviorally failed samples. The shaded band is the gap between what the model locates and what it says. n=100 per point.

above Qwen3-4B (0.58). A leaderboard built on emitted answers would report that the base model has no long-context ability at 128K. It locates the needle in 92% of the samples it answers incorrectly, with a median vote margin of 0.64 and precision 1.00 on the samples the vote accepts at threshold 0.3.

Second, the gap widens with depth rather than closing (Figure 2). Behavior falls by roughly two thirds between s=2 and s=4 on both models, while trace recall stays within a few points of flat. Even at s=1, direct extrapolation past the window, the trace finds the needle in 85% and 71% of failures. The vote overturns no behaviorally correct sample in any of the eight cells, so the gap is not a trade.

A third dissociation appears when the design choice varies rather than the model or the depth, and it carries a direct engineering cost. Replacing YaRN with NTK-aware base scaling at 64K improves the generated score in all three cells we measure, by up to 14 points, while degrading the trace in two of them. Choosing a schedule on end-task score at 64K therefore selects the schedule with the weaker retrieval signal, and at 128K that schedule is the one that collapses further (Table 2). Across design choices, behavior and retrieval move in opposite directions.

4 Locating the Loss in the Output Path

"Readout failure" is a real claim only if the output path can be cut into stages and the loss attributed to one of them. We cut it into four. Each stage removes one degree of freedom from the previous one, and we measure all four on the same behaviorally failed samples, so the first stage scores 0 by construction and the other three are directly comparable.

The four stages are: *free generation*, greedy decoding scored by the benchmark metric, where the model chooses what to say and how to say it; *trie decoding*, greedy decoding restricted to a token trie over the enumerated candidates, where only the choice of candidate is free and it is still made one token at a time; *candidate likelihood*, where each candidate is scored in full under the same forward pass and the argmax is taken, so sequential commitment is gone; and the *attention trace*, which does not consult the token distribution at all and uses only the order of the retrieval heads' attention logits over the candidate set.

Figure 3 gives the result, and we call the pattern it shows the readout bottleneck. Trie decoding recovers 2 to 16% of the failures. Removing sequential commitment as well recovers 22 to 40%. Reading the trace recovers 85 to 96%. Each step is significant on the shared samples. Pooled over cells, the likelihood stage beats trie decoding on 109 samples and loses on 17 (p=1.3×10^{-17}), and the trace beats the likelihood stage on 255 and loses on 8 (p=7.1×10^{-65}, exact paired sign tests).

One part of this ordering is universal and one is not. The gap between the two middle stages and the trace holds in every cell we have measured, including checkpoints of different size and training recipe. The order of the two middle stages does not. Under position interpolation at 64K, trie decoding recovers 34 to 42% of failures while candidate likelihood recovers none (21↓/0↑ and

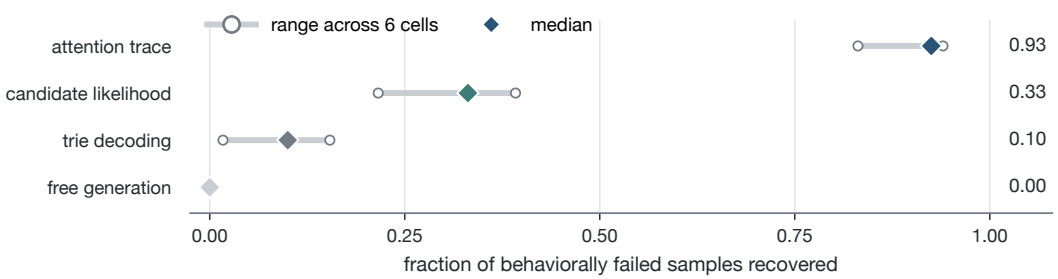

Figure 3: The four stages on behaviorally failed samples, six cells. Free generation is 0 by construction. n=100 per cell for stages 2 and 3, and trace recall is computed on the 49 to 92 failed samples of the matching probe. Diamonds are medians; intervals span the six cell values.

17↓/0↑, $p < 10^{-4}$; we verified the scores against plain teacher forcing to rule out an artifact). Scoring a whole 32-character answer requires the joint probability of a long span to be ranked correctly, whereas trie decoding needs only a locally correct next token given a feasible prefix. When position information is contracted uniformly, the long-span judgement breaks first. The output path is therefore not one ladder but several channels with different sensitivities to positional degradation, and the channel that consults the token distribution least degrades last.

The evidence is present and the loss is graded. If extrapolation had destroyed the retrieval computation, no stage downstream of it could recover the answer and the trace would sit at chance, about 3% for 32 candidates. If instead only token-level commitment were broken, trie decoding would recover most failures, since it removes exactly that freedom. It recovers under a sixth. What we see is a graded loss, and its largest single step falls between using the token distribution and not using it.

The damaged quantity is salience, not preference. The trace uses only the order of attention logits within the candidate set, and that order is intact. The downstream stages additionally depend on the magnitude of attention on content relative to everything else, because that magnitude determines the value vector the residual stream receives. Section 5 shows that magnitude is what extrapolation damages, and that the damage is localized in frequency.

When the trace fails. A readout that always worked would be suspicious, and the decomposition says where it must stop working. If a schedule damages the candidate ordering rather than only its salience, no stage can recover the answer. Position interpolation at s=4 is that case (§5.2). Every stage then sits on the floor together, with the trace at 0.12 to 0.16 in three of four cells and both middle stages below 0.06. Under NTK-aware scaling, which contracts the same bands less aggressively, the whole path sits just above that (§C.3). The trace is not an oracle that finds answers wherever attention exists. It fails where the mechanism says the information it reads has been destroyed.

Cost and calibration. The trace readout needs a parseable candidate set, and it inherits a confidence signal from the vote margin $m = (v_1 - v_2)/\sum_i v_i$. At $m \geq 0.3$ the accepted votes are essentially always correct, at 66 to 94% coverage, and rejected samples fall back to the generated text. The full sweep is in §C.3. Where the candidate set is not parseable, or the answer is not an extractive span, this instrument stops working, and §6 draws those boundaries by measurement.

5 What Interpolation Breaks

5.1 A temperature argument, and why it does not explain the collapse

The cheapest explanation for a retrieval failure at length N is a softmax budget problem, and it can be eliminated analytically. Let $\ell_\star$ be the needle's attention logit and $\mu = \ln \mathbb{E}[e^{\ell_i}]$ the competitors' log-mean-exponential. Then

$$\operatorname{logit}(p_\star) = \ell_\star - \ln\Big(\sum_i e^{\ell_i}\Big) \approx \ell_\star - \mu - \ln(N-1), \tag{2}$$

schedule	bands rewritten	behavior		trace recall on failures	
		MK3	KV	MK3	KV
YaRN	low only, high exempt	0.23 / 0.08	– / 0.06	**0.92 / 0.85**	– / **0.83**
NTK	all (base rescale)	0.09 / 0.02	– / 0.00	0.81 / 0.52	– / 0.64
PI	all (θ_j/s)	0.00 / 0.00	0.00 / 0.00	0.66 / **0.16**	**0.12 / 0.12**

Table 2: Three schedules at s=4 (128K), Qwen3-8B / Mistral-7B, n=50 to 100. Ordering schedules by how much of the spectrum they rewrite orders the damage to the trace, while behavior does not separate them. At 64K NTK is behaviorally better than YaRN yet worse in the trace (§3), and PI scores 0.000 at both depths while preserving the trace at s=2 (0.80 to 1.00).

so holding the needle's attention mass fixed as N grows requires the logit scale to grow like $\ln N$. From the training window N_0 to $N = sN_0$ the required amplification is

$$\rho(s) = \frac{\ln(sN_0)}{\ln N_0} = 1 + \frac{\ln s}{\ln N_0} = 1 + 0.0962\ \ln s \quad (N_0 = 32{,}768). \tag{3}$$

YaRN's attention temperature supplies $1/\sqrt{t} = 1 + 0.1 \ln s$ (§2), so its empirical constant matches $1/\ln N_0$ to within 4%: the $\ln N$ term is already paid at every s we study.

The collapse is therefore not a global temperature deficit, and no single global logit scale should recover it. None does. A single multiplier swept over $[0.70, 1.60]$ never beats the baseline significantly, and it collapses performance at both ends of the range (§C.3). Length-dependent softmax rescaling (Nakanishi, 2025) addresses the $\ln N$ term, and that term is not what is broken. The deficit is not expressible as one number, which sends us to the frequency axis.

5.2 The damage is confined to the bands the method rewrites

The prediction comes before the measurement and costs no fitted parameter. YaRN's interpolation strength per rotary dimension is the analytic γ_j of Equation (1). Dimensions completing more than 32 rotations inside the training window are untouched (γ_j=0), those completing fewer than one are fully interpolated (γ_j=1), and a ramp connects them. Grouped into eight bands of eight dimensions this gives $\bar{\gamma}_b = (0, 0, 0, 0.24, 0.74, 1, 1, 1)$. If the collapse were generic long-context degradation, the failure-versus-success gap in the retrieval heads' attention-logit margin would spread across the spectrum. If it is interpolation damage, the gap must live where $\bar{\gamma}_b > 0$.

The measurement is clean (Figure 5, appendix). In the three untouched bands the gap is at most 0.1 logit on both models. Every interpolated band shows a deficit, the bulk of it in the two lowest bands, and the deficit deepens monotonically from s=2 to s=4 while the untouched bands stay clean. A stronger reading is available in absolute terms. The untouched bands carry almost none of the retrieval margin, between -0.05 and $+0.18$ logits, against 10 to 57 for the interpolated ones. YaRN's high-frequency exemption therefore does not protect the margin, because the margin was never there. It protects the fine positional detail that keeps the candidate ordering stable, which is what the readout of §4 reads and what survives.

That gives a falsifiable prediction about other schedules, and it is a prediction about which component each one breaks. YaRN exempts the high-rotation bands. NTK-aware scaling rewrites the rotary base, which contracts every band. Position interpolation divides every frequency by s, the most aggressive of the three. If the exempted bands are what keeps the ordering stable, the trace should survive under YaRN, degrade under NTK and fail under PI, even though end-task score judges all three the same way.

The prediction holds (Table 2). Relative to YaRN at the same depth, NTK costs between 11 and 33 points of trace recall, and PI at s=4 costs almost all of it.

The damage to the ordering is graded in the amount of contraction rather than fixed per schedule. The same PI runs at s=2 leave the ordering intact. Behavior is already 0.000 in all four cells, yet trace recall is 0.80 to 1.00. PI first removes the salience that generation needs, and only at s=4 removes the order that a trace-side readout needs. Ordering the three schedules by how much of the spectrum they contract therefore orders the damage, and increasing the contraction within one

schedule moves it along the same axis. The band structure is not a description of one schedule's symptoms. It says which interventions cost salience only, and which also cost the ordering that salience is applied to.

5.3 THE MECHANISM IS DE-ANCHORING TO FILLER, NOT CONFUSION AMONG KEYS

A band-localized margin deficit does not yet explain how a model that ranks the needle first still answers wrongly. Three mechanisms could, and each predicts a different destination for the attention mass in failures: signal destruction predicts the needle stops winning even head-internal comparisons, key confusion predicts migration to the competing keys, and sink collapse predicts pooling at the attention sinks. We bucket every (head $\times$ decoding step) global argmax into four destinations: the needle, the distractor keys, the sinks (first four positions), and filler haystack text.

All three fail (Figure 6, appendix). In successes the argmax sits on the needle in 53 to 81% of head-steps; in failures, in about 0.2%. The mass does not move to the competing keys (0.1 to 1%) and it does not pool in the sinks (9 to 18%). It drifts to filler: 76 to 90% of head-steps place their maximal attention on task-irrelevant haystack text. Three companion measurements close it. Restricted to candidate lines, the needle still wins 61 to 74% of head-step votes in failures, which rules out signal destruction. When a head does rank the needle first globally, its median softmax advantage is about 10^4. In successes all 20 heads reach rank-1, while in failures a median of 2 to 6 ever do.

Two further negatives rule out the benign readings. Failures are not near misses. Over the 169 failed samples of the four checkpoint cells, the character-level longest common prefix with the gold answer has median 0, and only 2 to 21% share eight leading characters. Nor are they retrievals of the wrong entry. Every failed output is a well-formed uuid, but only 30 of 169 (18%) occur anywhere in the haystack. The dominant failure is a format-conformant fabrication: the shape of the answer is reproduced from the prior, with no entry copied from the document. That is what de-anchoring predicts and what key confusion does not. Once the maximal attention sits on filler, there is no content string in the value stream to copy.

The chain has three links. Interpolation compresses the low-band components of the attention logits (§5.2). Content then loses its salience advantage over filler at most heads and steps, so attention de-anchors from the content region. The softmax-weighted value stream that generation reads is dominated by filler and the answer never reaches the output, even though the order of preferences among candidates still encodes it (§4). The failure is specific: extrapolation degrades content-versus-filler salience while sparing content-internal discrimination.

6 BOUNDARIES OF THE ACCOUNT

6.1 WHERE A TRACE-SIDE READOUT APPLIES

The instrument of §4 is not general, and its conditions are decidable before the model runs. Two axes matter: whether the answer is an extractive span of the context, and how large the parsed candidate set is. In the extractive and enumerable regime (KV, needles, about 32 candidates) the trace beats generation by $+0.4$ to $+0.5$ and the vote margin is calibrated, with $m \geq 0.3$ giving precision 1.00. With an open candidate set ($\text{QA}_{1,2}$, 2.1 to 4.7K document sentences) the advantage shrinks to $+0.06$, which is not significant, and only rule agreement remains usable as a confidence signal (0.75 against 0.43). In the non-extractive regime (En.MC, LongBench-v2, four options) the trace gains nothing at all, -0.09 to 0.00, with precision flat at 0.48 to 0.51 (§C.4).

The negative row is instructive. On multiple-choice benchmarks the option the retrieval heads attend to at the answer position is the option the model emits, in 301 of 334 samples (90%), so the trace echoes the model's own decision rather than the location of evidence. A third condition is a property of the model rather than the task. Trace accuracy at $s{=}4$ spans 0.58 to 1.00 across our checkpoints (Table 1), and margin calibration degrades in step. The 4B model loses both, and its heads have a median retrieval score of 0.525 against 0.760 on Qwen3-8B. That median is a cheap pre-check, computed once per model from the diagnostic that selects the heads.

6.2 READING THE TRACE AGAINST STEERING THE DISTRIBUTION

If the trace carries the answer, an alternative to reading it is to steer the model with it. DySCO (Ye et al., 2026) does that. At each decoding step it aggregates query-focused retrieval heads' attention, selects the top context positions, and adds $\log \beta$ to the attention logits of all heads there. It reports gains on reasoning-style long-context benchmarks, including at 128K with YaRN at ratio 4, which is our headline configuration. We ran the authors' implementation with their published head set and hyperparameters under our protocol, paired against greedy decoding in the same process (Table 7, appendix). On the tasks where our readout gains $+0.4$ to $+0.5$, steering moves behavior by -0.04 to $+0.06$ and no cell is significant at n=50. A stronger bias, β=5, is not better. Their release supports Qwen3 and Llama only, so this is a single-model comparison, and with few discordant pairs the test has power against large effects only. It supports the claim that steering leaves the collapse in place, not the claim that steering is inert.

The comparison of magnitudes is the argument rather than the null result. The same attention signal moves behavior by at most 0.06 when it biases the token distribution and by 0.4 to 0.5 when it replaces it, which is what §4 predicts, because steering still routes the answer through the lossiest stage. The two methods do not compete in the non-extractive regime. On En.MC our readout gains nothing while DySCO reports gains, and we read that as a division of labour rather than a contradiction.

6.3 REPAIRING THE WEIGHTS IN PLACE IS HARDER THAN IT LOOKS

If salience is what interpolation compresses, restoring salience should fix the collapse. We probed that space and report it as a cautionary result rather than a method. A 74K-parameter query-side gain trained once at s=2 raises MK3 from 0.51 to 0.98 at 64K and from 0.23 to 0.90 at 128K, but it damages what it never saw, costing up to 0.42 on aggregation tasks and 0.116 on variable tracking inside the training window (p=5×10^{-8}). Gating it on a gold-free trace statistic does not rescue it: the best detector we could build is indistinguishable from leaving the gain always on, and it still costs 0.38 in its worst cell. Four preregistered negatives on the four-site repair space tell the same story: every site beats baseline at the fitted ratio and their ranking inverts with depth. Salience is the right quantity, but a static rescaling of it is not a safe instrument (§B.3, §C.4).

7 RELATED WORK

Rewriting the position map. Most extension methods edit RoPE (Su et al., 2024) directly. Position interpolation (Chen et al., 2023) divides every frequency by the extension ratio, NTK-aware scaling rescales the base, YaRN (Peng et al., 2024) interpolates a middle band and adds a logit temperature, and LongRoPE (Ding et al., 2024) searches a per-dimension schedule. A parallel line changes which distances are visible at all (Han et al., 2024; Jin et al., 2024; Li et al., 2026). All are judged by end-task score. We contribute not another schedule but a measurement of what a schedule breaks: the damage is band-localized and it spares the ordering of candidates, so it cannot be read off the end-task score.

Heads as instruments. Functionally identifiable heads have served as interpretability handles since induction heads (Olsson et al., 2022). Retrieval heads (Wu et al., 2025) and their query-focused refinement (Zhang et al., 2025) are the specialization we rely on. That literature uses such heads to explain a behavior or to select context. We use them as a readout, and we treat the model's token distribution as the failing component rather than as ground truth.

Intervening on attention at inference time. The nearest work is DySCO (Ye et al., 2026), which aggregates query-focused retrieval heads at each decoding step and adds $\log \beta$ to the attention logits of all heads at the top-scoring context positions. Two points of contact deserve precision. First, DySCO independently reports the observation our analysis is built on, that retrieval heads stay aligned with the gold span while accuracy degrades, and it reports it in an overlapping regime, since its 128K experiments also apply YaRN at ratio 4 to Qwen3. We therefore claim that observation as concurrent and more sharply quantified, with zero flips, four extension settings, a calibrated abstention margin and a candidate-set condition, rather than as novel.

Second, the two papers act on it in opposite ways. DySCO steers the distribution and reports gains on reasoning-style benchmarks, while we argue the distribution is the damaged component and read around it. Section 6.2 runs their implementation under our protocol and §6.1 states where each approach wins. Global attention sharpening (Nakanishi, 2025) is a weaker form of the same intervention, and the attention-sink literature (Xiao et al., 2024) explains why much attention mass is uninformative, which our destination analysis treats as a control rather than as the explanation.

Evaluation. Long-context benchmarks (Kamradt, 2023; Hsieh et al., 2024; Zhang et al., 2024; Bai et al., 2025) and positional-bias studies (Liu et al., 2024) score the emitted answer, which on our tasks conflates retrieval with readout. The gain used as evidence in §6.3 is a diagonal restriction of a LoRA update (Hu et al., 2022) on the query projection.

8 Implications for Long-Context Evaluation

A long-context score is the endpoint of a pipeline, not a diagnosis of the pipeline. In our setting, the same failed forward pass contains a correct evidence location, a weaker candidate-likelihood signal, and an incorrect generated answer. Calling all three a retrieval failure collapses distinct capabilities into one number. This distinction matters only when the claim being evaluated is about retrieval: for an end-user task, the emitted answer remains the outcome that matters. But when a benchmark is used to argue that a model cannot find information at length, behavior alone is insufficient evidence for that attribution.

The results suggest a two-axis report for extractive, enumerable tasks. The first axis is the ordinary generated-answer score. The second is an evidence-location diagnostic computed on the same samples, accompanied by its precision–coverage curve and the size of the candidate set. The two axes answer different questions: whether the deployed system returns the answer, and whether its long-context computation still isolates the evidence. Their separation exposes failures that a single leaderboard hides, including the checkpoint-ranking inversion in Table 1. Reporting both also prevents a high trace score from being mistaken for a usable system: the behavioral score remains visible, and the diagnostic's applicability conditions are explicit.

The same separation changes how extension methods should be compared. A schedule can preserve candidate ordering while degrading the salience needed by the token distribution, so selecting it at one length by end-task score can discard a model with the stronger retrieval computation. Conversely, the repair experiments show that restoring salience on the diagnostic task is not enough: a valid repair must preserve both the trace and behavior across lengths and task families. This makes the readout bottleneck a design constraint rather than a substitute metric. It identifies where information is lost, rules out repairs that merely overfit the probe, and turns "long-context collapse" from a single score into a falsifiable account of which stage failed.

9 Conclusion

Interpolation-based context extension fails unevenly: it damages what a model says more than what it attends to. Our four-stage cut locates the loss after retrieval. Evidence remains in the attention trace of a handful of frozen heads across every extension depth, then is progressively lost as the answer passes through the token distribution. The band-localized margin deficit and shift of global attention from the needle to filler provide a mechanism, while the schedule and repair experiments bound the account.

The central consequence is one of measurement. Two checkpoints with the same benchmark score can differ by 26 points in their ability to locate evidence; at 128K their ranking can invert. Behavior remains the correct measure of whether a deployed system returns an answer, but does not by itself establish retrieval failure. For extractive, enumerable tasks, we recommend pairing the emitted answer with a trace-side statistic, its candidate-set condition, and abstention calibration. Extension methods should analogously preserve both retrieval and readout across lengths rather than optimize either diagnostic in isolation. We release the instrument and protocol needed to make this distinction. A long-context model may still know where the answer is after it has lost the ability to say it; evaluation should make that boundary visible.

AI USE STATEMENT

Generative AI tools were used to edit the manuscript for readability and to check formatting. The authors reviewed and verified all AI-assisted revisions and take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The experimental protocol is summarized in §2; the appendices specify the head-selection and trace-readout procedures, report all measured cells and ablations, and document the negative results. Code accompanying the submission implements the trace readout and the reported evaluation protocol.

REFERENCES

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of ACL*, 2025.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. LongRoPE: Extending LLM context window beyond 2 million tokens. In *International Conference on Machine Learning (ICML)*, 2024.

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*, pp. 3991–4008, 2024.

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? *arXiv preprint arXiv:2404.06654*, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, Zirui Liu, Chia-Yuan Chang, Huiyuan Chen, and Xia Hu. LLM maybe LongLM: Self-extend LLM context window without tuning. In *International Conference on Machine Learning (ICML)*, 2024.

Greg Kamradt. Needle in a haystack: Pressure testing LLMs. `https://github.com/gkamradt/LLMTest_NeedleInAHaystack`, 2023.

Haoran Li, Sucheng Ren, Alan Yuille, and Feng Wang. CoPE: Clipped RoPE as a scalable free lunch for long context LLMs. *arXiv preprint arXiv:2602.05258*, 2026.

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

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

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 31932–31951, 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. In *International Conference on Learning Representations (ICLR)*, 2025.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Xi Ye, Wuwei Zhang, Fangcong Yin, Howard Yen, and Danqi Chen. DySCO: Dynamic attention-scaling decoding for long-context LMs. *arXiv preprint arXiv:2602.22175*, 2026.

Wuwei Zhang, Fangcong Yin, Howard Yen, Danqi Chen, and Xi Ye. Query-focused retrieval heads improve long-context reasoning and re-ranking. In *Proceedings of EMNLP*, 2025.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. ∞bench: Extending long context evaluation beyond 100K tokens. In *Proceedings of ACL*, 2024.

A The Diagnosis in Full

This appendix is the long form of §3–§5: every cell, every table, and the two findings the main text compresses into single paragraphs.

This section establishes one claim. When a RoPE model is pushed past its training window under interpolation-based extension, what breaks is not its ability to *find* the answer. What breaks is its ability to *act* on the answer it has already found. We call this a readout failure, and we establish it in four steps. Finding 1 localizes the damage: the discrimination margin collapses exactly in the frequency bands the extension method rewrites, a partition computable before any measurement. Finding 2 delivers the surprise: the retrieval signal survives at every depth, with or without the extension method installed. Finding 3 identifies the mechanism: attention de-anchors from content to filler, while preferences among candidate answers stay intact. Finding 4 maps the repair space: no structural prior survives contact with the data. The deciding experiment for each finding was preregistered with explicit decision branches before it ran.

Setup. We diagnose the failure on one controlled task before proposing any fix. The testbed is RULER `niah_multikey_3` (MK3). The model must retrieve a UUID associated with a query key hidden among hundreds of distractor key–value pairs. We choose MK3 because it isolates exact-match retrieval under distraction, and because it has the largest dynamic range of the 13 RULER tasks on both models. All numbers are scores over $n{=}100$ samples per cell, with greedy decoding, the model's own chat template, and RULER's official per-task generation budget. Paired comparisons use exact sign tests on shared samples, with Holm correction within each test family. At $n{=}100$ the detectable effect size is about 0.10; we never interpret a non-significant cell as evidence of no effect. Diagnosis instruments a fixed set of 20 retrieval heads per model. These heads are identified once on held-out short-context data (§B) and no gold information is used at decision time.

The problem: YaRN is necessary, insufficient, and not free. Figure 4 fixes the starting point with three facts. First, direct extrapolation dies behaviorally. Without YaRN, Qwen3-8B falls from ≥ 0.99 inside the window to 0.26 at 64K and to 0.00 beyond; Mistral-7B behaves alike (0.08 at 64K). Second, YaRN recovers only part of the loss. With the matching extension ratio, Qwen reaches $0.51/0.36/0.25$ at 64/96/128K and Mistral reaches $0.28/0.17/0.10$. Third, YaRN costs something inside the window. Qwen pays up to -0.10 in-window; Mistral pays heavily (32K: $0.71 \rightarrow 0.47$). Two controls sharpen what "collapse" means. Holding everything else fixed, the failure scales with the number of competing keys (MK1 0.89 vs. MK3 0.25 at 128K on Qwen). A non-retrieval long-range task survives at the same length (variable tracking: $0.94/0.90$). The received interpretation, ours included until the measurements below, is that extrapolation *damages retrieval*. Everything that follows shows this interpretation is wrong, in a specific and consequential way.

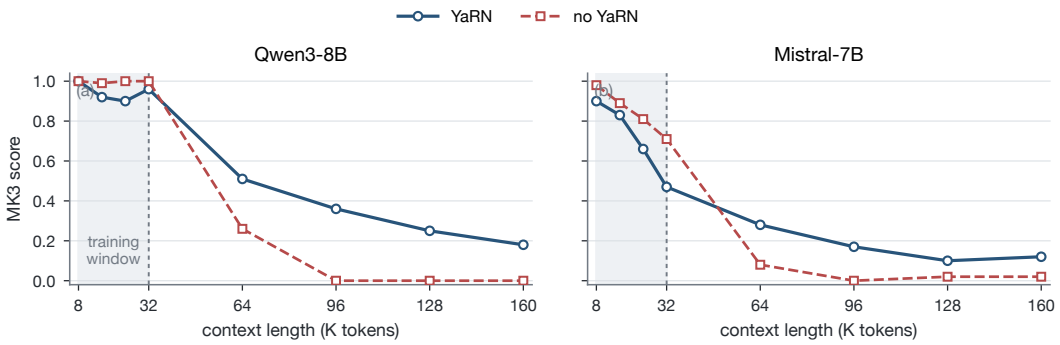

Figure 4: MK3 score vs. context length. The training window ends at 32K (vertical line; shaded region is in-window). Beyond the window, "no YaRN" is direct extrapolation (s=1); YaRN uses the extension ratio matching the target length. Within the window, the YaRN curve uses s=2, i.e. the deployment configuration of a 64K-extended model. n=100 per point.

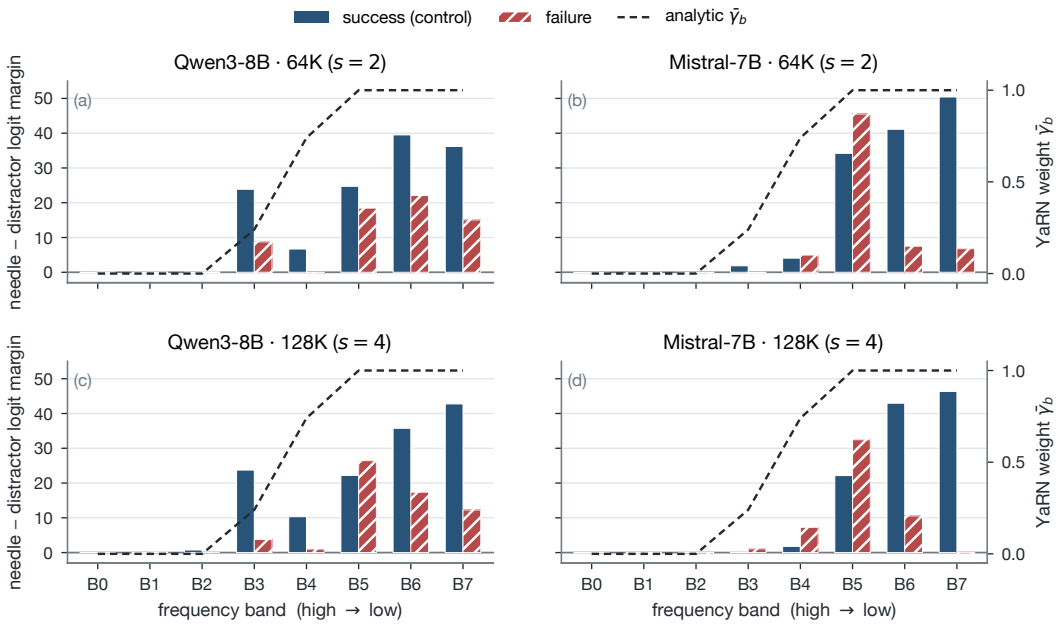

Figure 5: Per-band needle-versus-distractor attention-logit margin of the 20 retrieval heads, success versus failure, at both depths. The dashed line is $\bar{\gamma}_b$, computed analytically from the RoPE frequencies and the YaRN ramp with no fitted quantity.

A.1 Band Localization

The analysis comes before the measurement. YaRN's interpolation strength per rotary dimension is an analytic function of the RoPE frequencies. Write r_j for the number of rotations dimension j completes inside the training window. Dimensions with r_j>32 are untouched (γ_j=0); dimensions with r_j<1 are fully interpolated (γ_j=1); a ramp connects the two regimes. Per 8-dimension band this gives $\bar{\gamma}_b = (0, 0, 0, 0.24, 0.74, 1, 1, 1)$. This yields a falsifiable prediction. If the failure were generic long-context degradation, the failure–success margin gap should spread over the spectrum. If the failure is interpolation damage, the gap should be confined to the bands with $\bar{\gamma}_b$>0.

The measurement decides cleanly (Figure 5). We decompose the retrieval heads' attention logits by band, separately on the failure and success pools. In the three untouched bands, the gap is at most 0.1 logit on both models. Every interpolated band shows a deficit, and the bulk sits in the two lowest, fully interpolated bands. The same morphology deepens monotonically from s=2 to s=4,

model	extension	behavior	trace recall (failures)	trace accuracy
Qwen3-8B	none (s=1)	0.26	85.1%	0.89
	YaRN s=2	0.51	95.9%	0.98
	YaRN s=3 (96K)	0.38	93.5%	0.96
	YaRN s=4 (128K)	0.23	92.2%	0.94
Mistral-7B	none (s=1)	0.07	71.0%	0.73
	YaRN s=2	0.27	89.0%	0.92
	YaRN s=3 (96K)	0.15	83.5%	0.86
	YaRN s=4 (128K)	0.08	84.8%	0.86

Table 3: Generated-answer score (behavior) vs. the attention-trace vote. *Trace recall*: fraction of behaviorally failed samples in which the plurality vote identifies the needle. *Trace accuracy*: vote correctness over all samples. 64K except where noted; n=100 per row, behavior re-measured on the probe run. The vote never overturns a behaviorally correct sample (flip rate 0% in all eight cells).

while the untouched bands stay clean. The deficit is therefore not merely correlated with frequency. It is confined to the bands the extension method rewrites, exactly as the zero-parameter prediction demands. Finding 1 localizes what interpolation perturbs. The next findings measure what that perturbation actually breaks, and what it leaves intact.

A.2 The Surviving Trace

If interpolation destroyed retrieval, the needle's attention signal should be gone in failed samples. It is not. In 69% of Qwen failures (62% on Mistral, s=2), some retrieval head ranks the needle *first* among all ∼65K keys at some decoding step. What distinguishes failure from success is *holding* that rank: needle argmax occupancy is 0.12 in failures vs. 0.74 in successes (Qwen).

To quantify how much retrieval survives, we read the attention trace itself, without gold information. We parse the candidate key–value lines from the prompt. At every decoding step, each retrieval head casts a vote for the candidate holding its maximal attention logit. The trace's answer is the plurality winner.

Table 3 states the central fact, and it is a scissors (Figure 2). As depth grows from s=2 through s=3 to s=4, behavior collapses monotonically by two thirds ($0.51 \to 0.38 \to 0.23$ and $0.27 \to 0.15 \to 0.08$), while trace recall stays within a few points of flat ($95.9 \to 93.5 \to 92.2$ and $89.0 \to 83.5 \to 84.8$). Measured at every extension depth we run, the two curves separate; the thing being lost is not the thing being measured by behavior. Across two architectures and four extension settings, including no extension method at all, the trace identifies the correct needle in 71–96% of the samples the model answers incorrectly. It never overturns a correct one.

Retrieval, as performed by the attention mechanism, is largely *natively extrapolatable*. Even at s=1, where the model scores $0.26/0.07$, the trace still finds the needle in $85\%/71\%$ of failures. What the extension method mainly buys is not the ability to find the needle (recall gains 11–18 points). It buys the ability to *act* on the needle (behavior gains 25 points on Qwen). The collapse is a failure of the path from attention to generated text. It is a *readout* failure, not a retrieval failure.

The word "readout" can be made precise by asking how much of the answer each successive stage of the output path retains. Figure 3 does this on behaviorally failed samples only, so plain generation scores 0 by construction. Constraining generation to the enumerated candidate strings with a token trie, which removes every degree of freedom except which candidate to emit, recovers only 2–16%. Scoring each whole candidate string by its likelihood under the same forward pass, which removes the sequential commitment as well, recovers 22–40%. Reading the attention trace recovers 85–96%. The ordering holds in all six cells and it locates the damage: the further a readout stands from the token distribution, the more of the answer it retains, and the model's own distribution is the lossiest stage of all.

The condition on this finding: the candidate set. What the trace supplies is a *ranking within a candidate set*, so how far the finding carries depends on how large that set is. Table 3 uses the task's own enumerated candidates ($\approx$ 32 key–value lines), and there the scissors is flat in depth.

checkpoint	length	behavior	recall (fail)	trace acc	flips	n_{corr}
Qwen3-4B	64K	0.26	62.2%	0.72	0	16
	128K	0.08	54.3%	0.58	0	2
Qwen3-8B-Base	64K	0.26	97.3%	0.98	0	13
	128K	0.00	92.0%	0.92	–	0

Table 4: NIAH_MULTIKEY_3, n=50 per row. n_{corr} is the number of behaviorally correct samples, i.e. the denominator of the flip count; at 0.00 behavior the zero-flip statement is vacuous and we mark it. Behavior is re-measured in the probe run and differs from the generation run by 2–3 samples on Qwen3-4B (0.32/0.04 there), a bf16 kernel effect; the Qwen3-8B-Base rows agree sample for sample. Qwen3-4B is the weakest cell in the paper and the reason §B.2 states a capacity condition.

Repeating the measurement with an open candidate set, every sentence of the document, parsed from punctuation alone, 2.1–4.7K candidates per prompt, keeps the qualitative result and weakens the quantitative one: the sentence holding the answer is ranked first in 0.40–0.74 of samples at s=2 (chance $\approx 4 \times 10^{-4}$), and on the same task that figure falls from 0.74 at s=2 to 0.52 at s=4 (§C). Retrieval is natively extrapolatable in the sense that matters, the evidence is still locatable from the trace when the generated answer is wrong, but the margin by which it survives narrows as the candidate list grows, and we state F2 with that condition attached rather than as a claim about arbitrary readout targets.

A second condition is worth stating here because it is easy to over-read the scissors: the trace and the behavior are distinct variables only when the answer is an *attended span*. On multiple-choice benchmarks the option the retrieval heads attend to at the answer position is the option the model emits, in 90% of samples, so there is no gap between trace and behavior to measure (§C). F2 is a statement about extractive readout, not about long-context accuracy in general.

The collapse reproduces on two further checkpoints, and the sharpest case is the one with no instruction tuning. The two models used above are both instruction tuned, which leaves open the reading that the readout gap is an artifact of chat formatting or alignment. It is not. Table 4 repeats the measurement on Qwen3-4B (same family, smaller scale, matched to the same absolute lengths through its own 40960-token native window) and on Qwen3-8B-Base, which is not instruction tuned and is evaluated in a completion protocol with no chat template.

Qwen3-8B-Base at s=4 is the cleanest instance of the finding in this paper: the model scores 0.00, it answers not a single one of the 50 samples correctly, while the trace identifies the needle in 92% of them, with a median vote margin of 0.64 for correct votes against 0.03 for wrong ones and τ=0.3 precision of 1.00. The answer is present and separable in the attention trace of a model whose task accuracy is zero. Inspection of its outputs confirms the F3 pattern rather than degeneration: it emits well-formed uuids drawn from the wrong entry.

The same table carries the honest counterweight. On Qwen3-4B the trace is much weaker, recall 62.2%/54.3% and trace accuracy 0.72/0.58, and its vote margin barely separates correct from wrong votes (0.67 vs. 0.58), so τ=0.3 precision is only 0.72/0.60 instead of 1.00. This is not a head-selection failure: the median retrieval score of the 20 selected heads is 0.525 on Qwen3-4B against 0.760 on Qwen3-8B and 0.596 on Qwen3-8B-Base. Smaller models have weaker retrieval heads, so both the finding's magnitude and the readout built on it scale with capacity.

A.3 De-anchoring to Filler

Findings 1 and 2 leave a puzzle. The margin is damaged, yet the signal is readable. Three mechanisms could reconcile the two, and each makes a different prediction about where the attention mass goes in failed samples. Under *signal destruction*, the needle should stop winning even head-internal comparisons. Under *key confusion*, mass should migrate to the competing keys. Under *sink collapse*, mass should pool at the attention sinks. We bucket every (head $\times$ decoding step) global argmax into four destinations: the needle, the distractor keys, the sinks (first 4 positions), and filler text.

All three candidate mechanisms fail (Figure 6). In successes, the global argmax sits on the needle in 53–81% of head-steps; in failures, in about 0.2%.

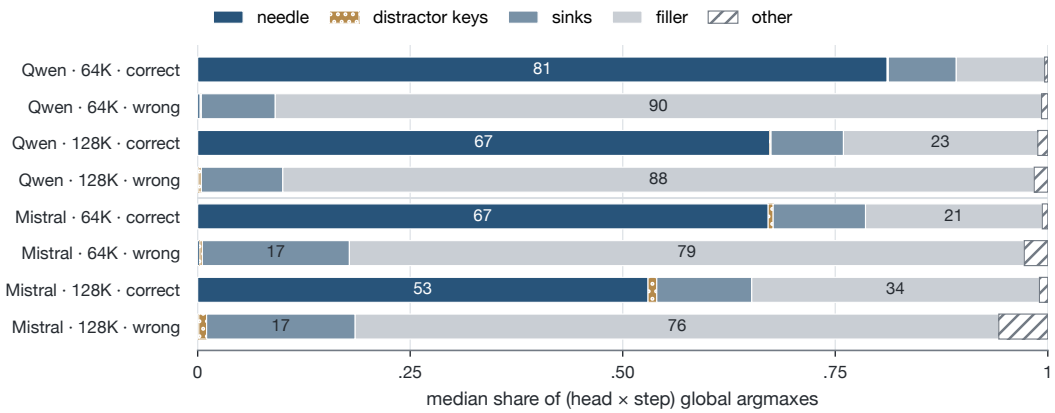

Figure 6: Destination of the global argmax, success against failure. Widths are the median over samples of the per-head share of (head $\times$ step) argmaxes. Numbers label segments of at least 15%; *other* is the small residual outside the four named destinations.

The mass does not move to the competing keys (0.1–1%), and it does not pool in the sinks (9–18%). It drifts to filler: 76–90% of head-steps place their maximal attention on task-irrelevant haystack text. The failure is *de-anchoring*. Three companion measurements complete the picture. First, relative preference among the actual candidates is intact. Restricting the same argmax to candidate lines, the needle still wins 61–74% of head-step votes in failures. This is exactly why the plurality readout of Finding 2 works, and it rules out signal destruction directly. Second, when a head does put the needle on top globally, it does so decisively. The median softmax advantage over the best distractor is about 10^4; failure is not a photo-finish among keys. Third, the deficit is one of coverage and persistence. In successes, all 20 instrumented heads reach rank-1 and hold it. In failures, a median of 2–6 heads ever do, and only transiently.

Two further negatives rule out the benign readings of these numbers. The first is that the readout is not broken but merely imprecise – that failures are approximate hits, a partially transcribed needle. On every behaviorally failed sample of the four checkpoint cells of Table 4 ($n = 169$), we measured the character-level longest common prefix between the emitted uuid and the gold uuid. The median is 0 characters in all four cells, and only 2–21% of misses share as many as 8 leading characters. Failures are not degraded copies of the target. The second is that failures are simply retrieval of the wrong document entry – a key-confusion story at the level of whole strings rather than of attention mass. Every failed output ($169/169$) is a well-formed uuid, but only 30 of them (18%; $2/50$ to $13/37$ across cells) occur anywhere in the haystack.

The dominant failure is a format-conformant fabrication: the model reproduces the *shape* of the answer from its prior while copying no entry from the document at all. This is what de-anchoring predicts and what key confusion does not. Once the maximal attention sits on filler, there is no content string in the value stream to copy, correctly or incorrectly.

Assembled with Finding 1, the causal chain has three links. Interpolation compresses the low-band components of the attention logits (F1). Content lines then lose their salience advantage over filler at most heads and steps, so attention de-anchors from the content region (this finding). The softmax-weighted value stream that generation reads is therefore dominated by filler, and the answer never reaches the output, even though the *order* of preferences among candidates still encodes it (F2). The failure is specific: extrapolation degrades content-vs-filler *salience* while sparing content-internal *discrimination*.

A.4 The Repair Space

Before building a fix, we probed the space of minimal weight-space repairs. The probes are per-head positional gains fitted by short-context self-distillation at each of the four sites of the attention path (query, key, value, output), and per-head-per-dimension gains over the rotary spectrum. Four preregistered structural results follow; full protocols and tables are in §C.

(i) Every site helps; no site is the mechanism. Fitted gains at all four sites significantly repair MK3 at the fitted depth. This includes the two sites that cannot change the attention distribution. The damage is multi-component. This is consistent with F3's mixture of salience loss (distributional) and the amplitude consequences of band compression.

(ii) Fitted repairs do not transfer. The best site at the fitted depth is architecture-dependent. On Qwen the site ranking *inverts* with depth: the axis best at $s{=}2$ is the worst at $s{=}4$. Cap and clamp interventions localize the cause. It is the fitted in-window gain profile interacting with deeper interpolation, not gain extrapolation.

(iii) The repair is not spectrally principled. Gains trained with access to the full rotary spectrum place 89–92% as much mass on the untouched dimensions ($\bar{\gamma}_j{=}0$) as on the interpolated ones. Their within-band shape does not follow $\bar{\gamma}_j$ (ramp-dim correlation null). The constraint is also costly in the other direction: restricting a trained gate to the interpolated band loses 0.25 (Qwen) and 0.22 (Mistral) at 128K relative to the same gate trained over the full spectrum. The *damage* is schedule-localized (F1), but the repair cannot be derived from, or even confined to, the schedule.

(iv) The repair is not head-principled. Fitted gains show no significant concentration on the independently identified retrieval heads. Restricting a trained gate to those heads forfeits most of its effect. The heterogeneity a weight-space repair must express is distributed across the head population.

The summary judgment: weight-space repair works, but it is a distributed patch without a structural principle. It re-injects content salience wherever the optimizer finds purchase. It cannot be read off the extension schedule. It must be re-fitted per architecture and validated per depth. This is precisely what one expects if the damage it compensates is F3's diffuse de-anchoring rather than a localized broken circuit.

A.5 Summary

Extrapolation under interpolation-based extension does not break retrieval. It breaks the readout of retrieval. Each finding converts into a design constraint, and together they leave little freedom:

- F1 (damage is schedule-localized in frequency) rules out treating the collapse as generic degradation; any fix must confront the interpolated spectrum.
- F2 (the order among candidates survives at every depth) makes a training-free repair possible in principle: the answer can be read from the trace instead of the generated text.
- F3 (the specific deficit is content-vs-filler salience, with transient, scattered rank-1 events) dictates the form of that readout: restrict comparisons to candidates, and count votes across heads and steps rather than trusting any single one.
- F4 (no structural prior survives) dictates the form of the weight-space complement: impose no site, head, or band structure; learn a tiny unconstrained gain field and validate it under depth transfer.

§B builds exactly these two instruments and nothing else.

B Instruments

This appendix specifies completely the trace readout of §4 and the salience gain of §6.3.

The method of this paper is derived, not designed. The diagnosis of §A separates what is broken from what is intact. Broken: the salience of content over filler, hence the softmax mass that generation reads (F1, F3). Intact: the order of preferences among candidate answers, at every depth and even without an extension method (F2). Unavailable: any structural shortcut for a weight-space fix (F4). Two instruments follow. The trace readout reads the intact order and bypasses the broken mass readout entirely; it is training-free but requires a parseable candidate structure (§B.2). The salience gain repairs the mass itself with a tiny, deliberately unconstrained learned gate; it does not require a parseable candidate set, but it requires one fitting run and task-level safety validation (§B.3). The two address different layers of the stack (§B.4).

B.1 The Fixed Head Set

Both instruments, and every diagnostic, use the same fixed set $\mathcal{H}$ of 20 retrieval heads per model. We identify them once, inside the training window, on held-out short-context retrieval data: rank all heads by the fraction of decoding steps at which their attention argmax falls on the gold span, and keep the top 20. The set is then frozen. It is never re-identified per length, per extension ratio, or per task, and identification never touches evaluation data.

Two properties make this a safe foundation. First, the set is cheap: one short-context pass per model, no gradients. Second, and critically, Finding 2 shows that the signal these heads carry is natively extrapolatable. Heads identified at 32K still rank the correct candidate first in 71–96% of failed samples at 64–128K, with or without YaRN. The instrument does not decay with the very extrapolation it is used to repair. This is the empirical license for everything below.

B.2 The Trace Readout

Rule. Parse from the prompt a candidate set $\mathcal{C}$ of answer-bearing spans, using only the task's surface pattern (here: key–value lines; no gold information). Run the model unchanged. At each decoding step t, each head $h \in \mathcal{H}$ casts one vote for the candidate containing its maximal attention logit, restricted to $\mathcal{C}$:

$$v(c) \;=\; \sum_{t} \sum_{h \in \mathcal{H}} \mathbf{1}\left[c = \arg\max_{c' \in \mathcal{C}} \max_{p \in c'} \; q_{h,t}^{\top} k_{h,p} \right],$$

and the readout returns the value field of $\hat{c} = \arg\max_c v(c)$.

Procedure. The end-to-end pipeline has four steps. (1) *Parse*: extract $\mathcal{C}$ with the task's candidate pattern; $|\mathcal{C}|$ is approximately 32 in the headline task, while the open-set scope experiments use up to 4.7K candidates. (2) *Prefill*: run the standard prefill; cache the key vectors of the heads in $\mathcal{H}$ at candidate positions only. This adds $O(|\mathcal{H}| \cdot |\mathcal{C}| \cdot \ell \cdot d_h)$ memory, where ℓ is the candidate span length; everything else in the KV cache is untouched. (3) *Decode*: generate as usual. At each step, one small matrix product against the cached candidate keys yields the per-head candidate argmax; accumulate votes. (4) *Read out*: return the value field of the plurality winner $\hat{c}$, together with a confidence score.

Why votes over candidates, not mass over everything. Each half of the rule neutralizes one measured failure mode. Restriction to $\mathcal{C}$ neutralizes de-anchoring: the global argmax drifts to filler in 76–90% of failed head-steps (F3), so unrestricted votes lose 30–40 recall points, but preference *within* the candidates is intact (61–74% needle share in failures). Plurality across head-steps neutralizes transience: in failures only a median of 2–6 heads ever reach rank-1 and only briefly (F3), so no single step is trustworthy, but counting over the whole trace is. The rule reads exactly the quantity Finding 2 showed to survive, and ignores exactly the quantity Finding 3 showed to break.

Confidence and abstention. The vote margin $m = (v(\hat{c}) - v(c_2))/\sum_c v(c)$, with c_2 the runner-up, is a natural confidence signal. The trace readout abstains and falls back to the generated text when $m < \tau$. Empirically the margin separates cleanly *when the candidate set is small*: across two models and four settings (no extension, s=2, s=3, s=4) on candidate sets of ≈ 32 enumerated values, the median margin of correct votes is 0.61–0.89, against 0.02–0.16 for wrong votes. At τ=0.2 the trace readout accepts 66–94% of samples with 97–100% precision on the accepted set; at τ=0.3 precision is 100% in all eight model–setting cells.

This calibration is a property of small candidate sets, not of the margin as such, and we state the boundary rather than extrapolate it. On sentence-level candidate sets (2.1–4.7K candidates parsed from punctuation alone) the votes spread out and the margin collapses to the 0.01 range: median 0.012–0.034 for correct picks against 0.008–0.015 for wrong ones, and thresholding buys no precision (§C). The residual signal in that regime is *agreement between aggregation rules* rather than the margin: when plurality, per-candidate maximum, and per-candidate mass all select the same span, localization accuracy is 0.52–0.91; when they disagree it is 0.38–0.61. That is a usable abstention signal but a much weaker one than the margin on enumerable sets, and we do not claim more.

Calibration also requires condition (i) of the scope below, not just a short candidate list. On the two MC suites, whose candidate sets are four options, the margin carries no information about

correctness: sweeping τ from 0 to 0.5 moves the accept rate from 1.00 to 0.41 while precision stays at 0.48–0.51 throughout. A short candidate list makes the vote concentrate; it does not make the vote informative when the answer was never an attended span.

Empirical safety. The readout with fallback can, in the observed regime, only add correct answers on top of behavior. In all eight measured model–setting cells the plurality vote overturns zero behaviorally correct samples, so the composed system's score is bounded below by the base model's score in every measured cell. We state this as an empirical property of the regime, not a guarantee; the margin threshold exists precisely to keep it true under distribution shift.

Cost. No training, no extra forward passes, no change to the model or its outputs elsewhere. The overhead is the candidate-key cache and one $|\mathcal{C}|$-sized argmax per head per step, which is negligible against the base cost of long-context decoding.

Scope: two conditions, not one. The trace readout replaces a readout, so it helps exactly when the readout is what fails. That requires two things of the task, and both are checkable before the model runs.

(i) The answer must be an extractive span. The trace tells us *where* the frozen heads are reading; that is the answer only when the answer is the attended text. Key–value retrieval, needle tasks and extractive QA satisfy this. Multiple choice does not: the correct option is established by comparing the document against the option semantics, not by copying an attended span. Measured on two real MC suites (InfiniteBench En.MC, LongBench-v2), the trace readout neither helps nor hurts, Δ between -0.09 and 0.00 over eight cells, and its letter coincides with the generated letter in 90% of samples (100% in four of eight cells). On such tasks the trace and the behavior are the same variable, so there is nothing for a readout to recover (§C).

(ii) The candidate set must be small enough for the vote to concentrate. With ≈ 32 enumerated candidates the trace keeps 92–96% of its localization at $s{=}4$ and the margin is calibrated. With ≈ 4.7K sentence candidates top-1 localization falls from 0.74 at $s{=}2$ to 0.52 at $s{=}4$ and the margin degrades to a weak signal.

The two conditions are independent, and the three measured regimes separate cleanly along them: extractive and enumerable (KV, needles) gives $+0.4$ to $+0.5$; extractive with an open candidate set (QA) gives reliable evidence localization but only a directional end-to-end effect; non-extractive (MC) gives nothing. We state the method's domain as this conjunction rather than as a task list.

(iii) A capacity condition on the model, not the task. The trace readout inherits the quality of the heads it reads. Across four checkpoints the trace accuracy at $s{=}4$ is 0.94 (Qwen3-8B), 0.92 (Qwen3-8B-Base), 0.86 (Mistral-7B) and 0.58 (Qwen3-4B), and the margin calibration degrades in step: $\tau{=}0.3$ gives precision 1.00 on the three 7–8B checkpoints and 0.60 on Qwen3-4B, whose 20 selected heads have a median retrieval score of 0.525 against 0.760 on Qwen3-8B. We therefore claim the trace readout for models whose retrieval heads are strong enough to be read, report the 4B numbers as the measured lower end, and treat the head-score median as the cheap pre-check, it is computed once per model, before deployment, from the same diagnostic that selects the heads.

Positioning. Prior uses of retrieval heads at inference time *steer*: they intervene so that the softmax mass moves toward the tokens those heads favor, and the generation pathway is trusted to pick the answer up. The trace readout instead *reads*: the model's computation is untouched, and the answer is taken from the attention trace directly. Steering inherits the readout pathway's fragility under extrapolation (F3) and perturbs in-window behavior by construction; reading does neither. We test this head-to-head in §C.1: under an identical protocol, on the same samples and heads, a decode-time steering baseline moves accuracy by -0.06 to $+0.05$ with no comparison significant, while the trace readout reaches 0.92–0.98 on the same cells. The difference is not the amount of signal extracted from the heads, both read the same attention, but where the answer is required to appear. Steering still has to route the answer through the token distribution that F3 shows is de-anchored; reading does not.

B.3 THE SALIENCE GAIN

Parameterization. A static multiplicative gain on the query vectors of every attention head, resolved per rotary dimension pair:

$$\tilde{q}_h \;=\; q_h \odot (1 + g_h), \qquad g_h \in \mathbb{R}^{d_h/2},$$

broadcast over the two components of each rotary pair, applied at all positions, backbone frozen. Nothing else changes. The full parameter count is $|\mathcal{H}_{\text{all}}| \times d_h/2$: about 74K on Qwen3-8B and 66K on Mistral-7B, below 10^{-5} of the backbone. Acting on the query side rescales how strongly content is attended (the salience deficit of F3) while leaving the key geometry, and hence the order information that the trace readout reads, untouched.

In the rotary basis the gain is exactly a *diagonal* adapter on the query projection: the diagonal restriction of a LoRA update, with the rank freedom removed. The restriction is deliberate. The gain has just enough capacity to rescale per-head, per-frequency salience, the quantity the diagnosis identifies as broken, and not enough to memorize the fitting task or to rotate the key geometry. We did not run a LoRA capacity control at equal data and objective; what we have is the weaker observation that 32 and 640 gate parameters perform indistinguishably (p=0.533), which suggests capacity is not the binding constraint (§C).

Training. The gain is fitted once, at the shallowest extension ratio (s=2), by short-context self-distillation. Each training sample pairs a long prompt the extended model fails with a short in-window prompt containing the same needle. The frozen, ungated model on the short prompt is the teacher; the gated model on the long prompt is the student; the loss matches the teacher's gold-span token log-probabilities. An anchor term computed on in-window batches penalizes any drift of short-context behavior:

$$\mathcal{L} \;=\; \mathcal{L}_{\text{distill}} \;+\; \lambda\, \mathcal{L}_{\text{anchor}}.$$

No gold labels beyond the teacher's own outputs are required, and no long-context gradient data beyond the failed pool. Hyperparameters and the data construction are in the appendix. The gain trained at s=2 is deployed unchanged at s=3 and s=4; transfer is evaluated, not assumed (§C).

Design by elimination. The gain's form is not a guess. Each choice is the surviving branch of a preregistered structural test from Finding 4, and every structural prior we tried to impose was rejected by the data:

candidate prior	test outcome	resulting choice
one privileged site (q/k/v/o)	every site helps; ranking flips across depth and architecture (F4-i, ii)	fixed query-side site, chosen for order preservation, not repair monopoly
restrict to retrieval heads	head-restricted gate forfeits most of the effect (F4-iv)	all heads
restrict support to the interpolated band	costs 0.25/0.22 at 128K vs. the unconstrained gate (F4-iii)	full rotary spectrum
derive shape from the schedule ($\alpha \cdot \bar{\gamma}$)	learned profile flat; ramp correlation null; signs mixed (F4-iii)	shape learned, not derived

Table 5: Every structural prior tested against the fitted probe. The damage is schedule-localized (F1); the repair is not. The surviving parameterization imposes no structure: maximal freedom at minimal capacity.

The evidence does not support a principled low-dimensional repair: because the failure is diffuse de-anchoring rather than a broken circuit (F3), the fitted gain is best viewed as a task-specific probe. It repairs the fitted retrieval task with few parameters, but its regressions outside that task prevent safe general deployment.

regime	tasks	candidates	trace − generation	confidence signal
extractive, enumerable	KV, needles	$\sim$32	$+0.4$ to $+0.5$	margin: $m \geq 0.3 \Rightarrow$ prec. 1.00
extractive, open	$\text{QA}_{1,2}$	2.1–4.7K	$+0.06$ (n.s.)	rule agreement: 0.75 vs. 0.43
non-extractive	En.MC, LB-v2	4	-0.09 to 0.00	none: prec. flat at 0.48–0.51

Table 6: Both conditions are decidable from the task format. The non-extractive row is a genuine negative, not a weaker positive.

task	extension	greedy	DySCO (β=2)	Δ	sign test
KV	s=2 (64K)	0.280	0.320	$+0.040$	2↑/0↓, p=0.50
KV	s=4 (128K)	0.240	0.300	$+0.060$	5↑/2↓, p=0.45
MK3	s=2 (64K)	0.540	0.500	-0.040	1↑/3↓, p=0.63
MK3	s=4 (128K)	0.240	0.240	0.000	1↑/1↓, p=1.0

Table 7: The authors' DySCO implementation under our protocol, Qwen3-8B, n=50 per cell, official generation budgets. Their release supports Qwen3 and Llama only, so this comparison is single-model. β=5 is not better (-0.04 on both MK3 cells). Discordant pairs are few, so this test has power only against large effects: the claim it supports is that steering leaves the collapse in place, not that steering is inert. Full sweep and our own weaker vocabulary-space variant in §C.1.

B.4 Applicability

The decision procedure at inference time is one branch. If the task exposes a parseable candidate structure, apply the trace readout; the gain may be present or absent underneath, since the trace readout reads order and the gain moves mass, and the two act at different layers. If it does not, the branch is a scope decision rather than an algorithm: the gain is enabled only for task families whose measured envelope is non-negative, and the ungated model runs otherwise.

The asymmetry between the branches is deliberate and we state it plainly.

The trace-readout branch is self-protecting: it falls back to generation below the confidence threshold and overturns no correct sample in any measured cell. The gain branch is not. It costs 0.42 on CWE and 0.20 on VT at s=2, and 0.22/0.20 on CWE at s=4, so a system that enables it everywhere is worse than the baseline on those families.

We looked for a runtime detector that would close this branch automatically, gold-free concentration statistics of the same frozen heads, over 24 task–model cells, and it does not exist at the level of concentration: the statistics predict *that* the gain will change an outcome (AUC 0.66 for breaking a correct sample, 0.71 for fixing a wrong one) without predicting the sign, and a leave-one-cell-out threshold underperforms both the ungated and the always-gated system (0.741 vs. 0.777 vs. 0.793 micro-average). We report this negative result in full (§C) because it bounds what the salience gain contributes: a demonstration that the deficit is repairable with $\approx$ 70K parameters, not a component one can switch on blindly. Separate results for the gain and the trace readout are reported in §C.

C All Measurements

Protocol. As in §A: RULER, n=100 per cell, greedy decoding, chat template, official budgets, paired exact sign tests with Holm correction. Both models are extended with YaRN at the ratio matching each target length; all trained modules are fitted once at s=2 and evaluated frozen at $s \in \{2, 3, 4\}$.

C.1 Main Measurements

Head-to-head against steering. The natural competitor to a readout is a *steering* method that uses the same retrieval heads to move the model's output distribution. Our first baseline steers in *vocabulary* space: the 20 heads' attention over prompt positions is aggregated into a per-vocabulary score, max-normalized so the most-attended item receives exactly $+\lambda$, and added to the LM logits at every decoding step including the first; the prefill is untouched, so a single prefill serves all

model	condition	64K	96K	128K
Qwen3-8B	YaRN baseline	0.51	0.38	0.23
	+ gate (the salience gain)	**0.98**	0.92	0.90
	the trace readout	**0.98**	**0.96**	**0.94**
Mistral-7B	YaRN baseline	0.27	0.15	0.08
	+ gate (the salience gain)	**0.99**	**0.97**	**0.94**
	the trace readout	0.92	0.86	0.86

Table 8: MK3 across extension depths. The gain is the fitted probe: all-head, full-spectrum query gain, trained once at s=2 (§B.3). The trace readout numbers are the trace-vote readout of §B.2 on the baseline model (no training). Earlier band-restricted and gate+axis variants are reported as ablations (§C.3). The trace readout cells are the no-abstention readout (τ=0); the τ sweep is in Table 11. The trace readout is run on the baseline (ungated) model; the two instruments are alternatives rather than a pipeline, so no readout-on-gated-model row is reported.

λ. Everything else— model, YaRN factor, prompts, samples, head set, answer prefix, generation budget, is identical to our own runs. This is deliberately the cheapest form of steering; the stronger attention-space form of Ye et al. (2026) is reported separately below.

task	model	λ=0	λ=2	λ=5	λ=20
MK3	Qwen3-8B	0.50	0.49	0.49	0.52
	Mistral-7B	0.29	0.30	0.32	0.29
KV	Qwen3-8B	0.43	0.43	0.44	0.48
	Mistral-7B	0.33	0.33	0.34	0.27

Table 9: Vocabulary-space logit-bias steering at 64K (s=2), $n = 100$ per cell, official generation budget. No cell moves significantly under an exact paired sign test: the largest gain is $+0.05$ (KV/Qwen, $5\uparrow/0\downarrow$, $p = 0.062$ uncorrected, which does not survive Holm over the 12 comparisons) and the largest regression is -0.06 (KV/Mistral, $3\uparrow/9\downarrow$, $p = 0.146$). The trace readout on the same samples and the same heads reaches 0.92–0.98 (Table 8). The bias is weak-by-measurement rather than vacuous-by-construction: an earlier version had no effect at all because the per-sub-token attention mass was ≈ 0.01 and the first decoded token was unbiased; both defects were fixed before these runs. KV accuracy is identical at generation budgets 16 and 64 (byte-identical per-sample outputs), so the budget is not the binding constraint.

C.2 The Four-Site Probe

model	length (s)	no gate	Q-SIDE	K-SIDE	V-SIDE	O-OUT
Qwen3-8B	64K (2)	0.51	0.80	**0.90**	0.66	0.83
	96K (3)	0.36	0.45	**0.51**	0.39	0.46
	128K (4)	0.25	**0.39**	0.12	0.31	0.34
Mistral-7B	64K (2)	0.28	0.62	**0.93**	0.71	0.61
	96K (3)	0.17	0.22	**0.62**	0.36	0.21
	128K (4)	0.10	0.12	**0.55**	0.34	0.20

Table 10: Per-head positional gains fitted independently at four sites (trained at s=2, frozen). All eight 64K cells beat baseline ($p_{\text{Holm}}<0.01$), multi-component damage (F4-i); ranking inverts with depth on Qwen (F4-ii). 13-task breadth: Q-SIDE clean on Qwen @128K (0 Holm-significant regressions, MK3 $+0.14$); K-SIDE non-positive on 13/13. Cap (freeze tail) moves nothing; clamp (remove in-window profile) destroys Mistral K-SIDE $0.55 \rightarrow 0.12$, in-window attribution.

C.3 Ablations and Negatives

Head set (F4-iv). Gate on all heads vs. the 20 retrieval heads: 0.94 vs. 0.71 on Qwen @64K (24/1, $p_{\text{Holm}}<10^{-4}$).

model	setting	med. margin (corr./wrong)	τ=0.2		τ=0.3
			accept	prec.	prec.
Qwen3-8B	none (s=1)	0.73 / 0.07	80%	100%	100%
	YaRN s=2 (64K)	0.89 / 0.16	94%	100%	100%
	YaRN s=3 (96K)	0.81 / 0.11	88%	100%	100%
	YaRN s=4 (128K)	0.72 / 0.02	89%	100%	100%
Mistral-7B	none (s=1)	0.70 / 0.05	66%	98.5%	100%
	YaRN s=2 (64K)	0.88 / 0.03	83%	100%	100%
	YaRN s=3 (96K)	0.67 / 0.04	72%	100%	100%
	YaRN s=4 (128K)	0.61 / 0.10	76%	97.4%	100%

Table 11: Trace-readout abstention calibration. *Med. margin*: median vote margin of correct and of wrong plurality winners. *Accept*: fraction of samples with $m \geq \tau$; *prec.*: vote precision on that accepted set. n=100 per row. The vote overturns zero behaviorally correct samples in all eight cells.

Band support (F4-iii). Full-spectrum training places 89–92% relative mass on untouched dims; shape uncorrelated with $\bar{\gamma}$ in ramp dims. Behavioral cost of restricting support to the interpolated band (32–64): none at 64K (Qwen 0.98 vs. 0.94, Mistral 0.99 vs. 0.96, both n.s.), severe at 128K (Qwen 0.90 vs. 0.65, 27/2, $p<10^{-4}$; Mistral 0.94 vs. 0.72, 27/5, $p=10^{-4}$). The restriction looks free at the fitted depth and fails under transfer, the same trap as F4-ii.

Composition across weight-space components. With the band-restricted gate, adding the per-head axis helped on Qwen (96K: 0.72 $\rightarrow$ 0.82, 12/2, p=0.013) and was redundant on Mistral. The full-spectrum gate absorbs this margin (0.90–0.99 alone), so the final method carries a single weight-space component.

Vote scheme. Candidate-restricted plurality vs. global argmax votes: recall 95.9% vs. 65.3% (Qwen 64K), the filler drift of Finding 3 is what the restriction removes.

Abstention threshold. Table 11 sweeps τ over all eight model–setting cells. Precision on the accepted set is 100% at τ=0.3 everywhere, and the price is coverage, not correctness: rejected samples fall back to the generated text, so the composed score degrades toward the baseline rather than below it.

Two controls we did not run. Two ablations that a reader may expect are absent. We did not fit a rank-8/16 LoRA under the same objective; the 32-vs-640-parameter comparison (p=0.533) is only weak evidence that capacity is not binding. We also did not sweep the number of heads or decoding steps. Neither control is load-bearing for the claims here.

C.4 Scope Measurements

The trace readout's domain is a conjunction of two conditions. Table 6 places the three measured regimes on the two axes of §B.2: whether the answer is an extractive span, and how large the parsed candidate set is. The extractive-and-enumerable regime is where the headline numbers come from; the other two are measured here so that the domain is drawn by evidence rather than by assertion.

The non-extractive boundary, measured. On InfiniteBench En.MC and LongBench-v2 at s=3 and s=4 (n=16–50 per cell, set by the natural-length buckets), the trace readout moves accuracy by -0.09 to 0.00: En.MC 0.79 $\rightarrow$ 0.70 and 0.69 $\rightarrow$ 0.66 on Qwen, 0.75 $\rightarrow$ 0.75 and 0.72 $\rightarrow$ 0.72 on Mistral; LongBench-v2 unchanged in all four cells (0.34, 0.21, 0.44, 0.28). The diagnostic that explains why is the agreement rate: the option the trace readout selects is the option the model generates in 301/334 samples (90%; 100% in four cells). At the answer position the frozen heads attend to the option being emitted, so on a task whose answer is inferred rather than copied the trace and the behavior are one variable. This is the honest boundary of the method: a readout replacement recovers information only where the generation pathway was failing to transcribe something the attention had already found.

task	model	s=2 (64K)		s=4 (128K)	
		nogate→gate	Δ	nogate→gate	Δ
KV	Qwen	$0.28 \to 1.00$	$+0.72$	–	
	Mistral	$0.52 \to 0.96$	$+0.44$	–	
NIAH_MK2	Qwen	$0.84 \to 0.98$	$+0.14$	–	
	Mistral	$0.72 \to 0.92$	$+0.20$	–	
NIAH_SINGLE1	Qwen	$1.00 \to 1.00$	0.00	$1.00 \to 1.00$	0.00
	Mistral	$1.00 \to 1.00$	0.00	$0.92 \to 1.00$	$+0.08$
QA_1	Qwen	$0.66 \to 0.64$	n.s.	–	
	Mistral	$0.50 \to 0.46$	n.s.	–	
NIAH_MVALUE	Qwen	$0.98 \to 0.94$	-0.05	$0.91 \to 0.81$	-0.11
	Mistral	$0.93 \to 0.84$	-0.09	$0.83 \to 0.69$	n.s.
VT	Qwen	$1.00 \to 1.00$	0.00	$0.93 \to 0.92$	n.s.
	Mistral	$0.96 \to 0.77$	-0.20	$0.90 \to 0.49$	-0.41
CWE	Qwen	$0.71 \to 0.46$	-0.25	$0.40 \to 0.20$	-0.20
	Mistral	$0.49 \to 0.07$	-0.42	$0.27 \to 0.05$	-0.22

Table 12: The gain's damage envelope, n=50 per cell, paired sign tests on shared samples. The gain was fitted on NIAH_MULTIKEY_3 only, so every row is out of distribution for it. Rows ordered by effect. The losses on CWE/VT are significant at both depths measured. CWE, VT and NIAH_MVALUE are scored with partial credit, so their s=4 column is not directly comparable to the binary s=2 column; the sign of the effect remains comparable. Here, "n.s." marks $p > 0.05$ (VT/Qwen -0.00, $4\uparrow/5\downarrow$; NIAH_MVALUE/Mistral -0.14, $12\uparrow/24\downarrow$, $p = 0.065$).

Two caveats on this table's protocol. Prompt lengths in these suites follow the documents' natural distribution rather than being padded to the bucket name, so the median prompt is 84–118K tokens where the RULER cells sit within 2% of their target; the extension pressure in the MC cells is therefore somewhat lower than in the synthetic cells at the same nominal length. And the per-cell n ranges from 16 to 50 for the same reason, so single cells, En.MC/Mistral at s=3 in particular (n=16), carry little power on their own; the claim rests on the pooled 334 samples and on the agreement diagnostic, not on any one cell.

Two further checkpoints. Extending the trace readout to Qwen3-4B and Qwen3-8B-Base (§A.2, Table 4) gives the trace readout at τ=0 of 0.98/0.92 (Qwen3-8B-Base, 64K/128K) and 0.72/0.58 (Qwen3-4B) against generation of 0.26/0.00 and 0.26/0.08. The 8B-Base cells match the instruction-tuned models; the 4B cells are the measured lower end of the capacity condition of §B.2.

The salience gain helps single-span retrieval and damages aggregation, at both depths. Table 12 is the gain's measured envelope. It is not a uniform improvement and the losses are not noise: CWE loses 0.42 on Mistral at s=2 (p=1.4e−14) and 0.22 at s=4 ($p\approx$1e−8), so greater depth does not absolve it.

An open candidate set. To test the trace readout outside enumerable candidates we run it on RULER QA_1 (SQuAD) and QA_2 (HotpotQA) with the document's own sentences as candidates, parsed from punctuation alone: 2.1–4.7K candidates per prompt. The trace locates the sentence holding the answer first in 0.40–0.74 of samples at s=2 and 0.52 at s=4, against a chance rate of $\approx 4 \times 10^{-4}$. Feeding the selected sentence window back with the question (retrieve-then-read) scores 0.42–0.72 against 0.46–0.66 for long-context generation: positive in three of four cells, 26 wins to 15 losses pooled (p=0.117), and no cell survives Holm correction, so we report the end-to-end effect as directional only.

A random-window control scores 0.04–0.22, which bounds how much of it is answered from parameters rather than from the retrieved evidence. On the samples where generation fails, retrieve-then-read recovers 0.20–0.41 against 0.04–0.19 for the random window. The aggregation rule matters here in a way it cannot on single-token candidates: on QA_1/Qwen, localization is 0.64 (plurality), 0.80 (per-candidate maximum) and 0.76 (per-candidate mass) at s=2, and 0.40/0.52/0.62 at s=4.

No runtime detector for the salience gain. If the gate's envelope were predictable from the trace, the salience gain could be gated per sample. It is not, at least not from concentration. We measured gold-free concentration statistics of the same 20 frozen heads, top-k mass, entropy, participation ratio, blocks needed to cover 90% of the mass, on 24 task–model cells at s=2, n=50 each. Task-level rank correlation with the gain's effect is null (Spearman $|\rho| \leq 0.06$). At sample level the statistics are informative about *whether* the gain changes an outcome but not about the direction: participation ratio gives AUC 0.659 for breaking a correct sample and 0.711 for fixing a wrong one, in the same direction. A conditional gate with a leave-one-cell-out threshold reaches 0.741 micro-average against 0.777 ungated and 0.793 always-gated, with a worst-cell regression of -0.42. The most direct counterexample to the intuition: KV, where the gain gains most ($+0.72$), has the most diffuse trace (participation ratio 28.2), while VT, which it damages, has the sharpest (4.8).

Other interpolation schedules. Position interpolation scores 0.000 at every depth we test (degenerate repetition rather than wrong retrieval). NTK-aware base scaling beats YaRN at s=2 (NIAH_MK3/Qwen 0.65 vs. 0.51; KV/Mistral 0.40 vs. 0.33) and collapses at s=4 (0.09/0.02/0.00). F2 holds under NTK, trace recall 52–99% across 14 cells, zero flips, so the readout gap is not an artifact of the YaRN ramp, though its size is schedule- and model-dependent.

D Limitations and External Validity

The mechanism is established on synthetic retrieval tasks with enumerable candidates. The trace-side instrument does not transfer to non-extractive tasks, where the trace echoes the model's own decision in 90% of samples, and it does not survive an open candidate set undamaged, falling from 0.74 to 0.52 between s=2 and s=4 with document sentences as candidates. It also requires retrieval heads strong enough to read, which Qwen3-4B's are not.

Two attempts to make the analysis predictive failed, and we report them as bounds on it. Ranking extension schedules by their trace at 64K does not predict which of them collapses at 128K. On both models the trace order and the deep-behavior order disagree, because PI is trace-perfect at 64K and behaviorally dead at both depths. The instrument diagnoses whether evidence is present, not whether a configuration will survive further extension. Second, no gold-free trace statistic we tried predicts, per sample, when an in-place repair should be enabled. The best detector is indistinguishable from always enabling it and still costs 0.38 in its worst cell.

Our comparison against steering runs a single model, because the released implementation supports Qwen3 and Llama only, and its sign test has power against large effects only. The largest checkpoint we measure with the full path is 8B, with 14B behavior reported without the trace, so whether the ordering persists at 30B and above is untested. Our attempt at a quantitative law for the band deficit was falsified by our own data and is left open. Every claim above concerns YaRN-style interpolation and two corroborating schedules. We do not claim it for continued pretraining at length, which changes the weights and therefore the salience the analysis is about.