Lost in the Context: Probing Why Transformers Fail to Revisit the Right State

Anonymous authors
Paper under double-blind review

Abstract

Transformers can attend to their entire preceding context, yet they often fail on long-horizon reasoning tasks that require keeping and revisiting intermediate computational states. We distinguish two possible bottlenecks: state persistence (does a state written earlier still influence computation at later tokens and layers?) and referential persistence (can the model later retrieve it from among competing states?). Taking a dynamical-systems view in which token-wise transformations act as reaction terms and attention acts as a non-local transport and retrieval operator, we treat recall as depending on how much of a state reaches the query, how the query reads it out, and how it is selected among competitors. We measure the first as the linear response of the residual stream to a perturbation of one token's embedding; as earlier Green's-function analyses found, it decays roughly as a power law, with an exponent that falls with depth. Post-training can strengthen this long-range influence: in the Qwen pair, RL slows its decay on a reasoning-style probe through a broad shift of attention toward long range in middle and late layers (exponent change -0.021, replicated at -0.014, robust to a constant floor in the fit and about 75 times the change from random updates of the same size), but not on raw text. Stronger influence, however, does not mean better recall. An OLMo DPO step raises the state's influence 1,000 tokens away by 73% and the answer's sensitivity to it by 24%, yet lowers recall from 97.5% to 85% (15 prompts lost, none gained), and across prompts we detect no reliable link between the decay exponent and which recalls fail. Head ablation makes the dissociation causal: removing the heads that carry the most long-range influence steepens decay but leaves recall intact, whereas other head sets, without raising perplexity, abolish recall when the value must be selected among competing assignments. Long-range influence is therefore not a proxy for memory: methods meant to help models revisit earlier states should be judged by retrieval and readout, not by how far influence travels.

1 Introduction

Transformers have access to their entire preceding context, yet they often fail on long-horizon reasoning tasks that require maintaining and revisiting intermediate computational states (Liu et al., 2024; Lei et al., 2026; Mehta & Datta, 2026). Tree traversal makes the demand explicit: in a post-order traversal, the model must return to a parent after finishing its entire subtree, so it needs a pointer written many steps earlier. Such a failure can have two causes. *State persistence* fails when the state written earlier is no longer represented in the computation that needs it. *Referential persistence* fails when the state is still there but cannot be retrieved uniquely among competing states.

We study both through a dynamical-systems view of Transformer computation. Token-wise transformations act as a reaction term, and attention moves information across positions: local attention admits a transport–diffusion approximation, and long-range attention a non-local one. As a diagnostic, we perturb the embedding of one token and record the response at every later token and layer, a linear-response Green's function. Over the range we measure, the response decays roughly as a power law with distance, as Brändel et al. (2026) found for embedding-to-output sensitivity. We add a view resolved by layer and head, position controls, a link to the answer and, above all, how the response changes under post-training. Our central finding is that this influence and correct recall come apart: training can raise one while lowering the other, and the heads that carry influence are

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.

not the ones recall needs (Sections 5.3 and 5.4). In a phenomenological model, memory decay and the reliability of retrieval enter a depth law for tree traversal as separate terms (Prop. 4).

Training shapes both, so we use post-training as a controlled intervention. At the level of generation, KL-regularised RL adds a reaction term, the reward, to the transport learned by next-token prediction (Kappen, 2005; Todorov, 2006; Levine, 2018). Inside the network, a first-order identity bounds how much one update can change any internal property at a given context: the change is limited by how closely that context aligns, in the tangent kernel, with the states the objective trains on (Thm. 1). The identity helps organise our results, although we do not measure the alignment. We compare RL with supervised objectives on public checkpoints from two model families and by fine-tuning the same starting model.

Contributions.

- **Influence is not recall.** Across training steps, prompts and interventions, long-range influence and recall come apart. A DPO step that raises long-range influence and answer sensitivity lowers recall; the decay exponent shows no reliable per-prompt link to recall failures (answer-level sensitivity does better, but no better than the answer's own confidence); and ablating the heads that carry the most long-range influence steepens decay without affecting recall, while other heads, whose removal does not raise perplexity, are necessary for recall among competing states (Sections 5.3 and 5.4).
- **Post-training strengthens long-range influence, depending on the probe.** In the Qwen pair, RL slows decay on a reasoning-style probe but not on raw text, robustly to the fitted form and about 75 times more than random updates of the same size, in the middle and late layers and with a broad shift of attention toward long range; the change reaches the answer, though not selectively. Cross-entropy fine-tuning of the same model steepens decay (Sections 5.2 to 5.5).
- **Diagnostics.** Layer-resolved linear-response Green's functions of the residual stream, with linearity, stationarity, position and floor controls; head-level attention to the state and head ablation with behavioural and perplexity readouts; and the answer's sensitivity to the state (Section 3).
- **Theory.** A read factors into transport, readout and selection among competitors (Prop. 5), which is why influence and recall can come apart. A residual stack keeps the tail exponent of a single layer (Thm. 2), and a tree depth law separates memory from pointer retrieval (Prop. 4). We treat the fractional transport model as phenomenological.
- **A controlled model** in which exposure to long gaps without restating raises persistence without any reward, and RL raises passive persistence only when restating is costly (Section 5.6).

We also report which hypotheses failed and which links, above all the link to reasoning success, remain untested (Section 6).

2 Related work

Generation as controlled transport. KL-regularised control has an exponentially tilted optimum and a linear Bellman or Hamilton–Jacobi–Bellman equation that Feynman–Kac solves (Todorov, 2006; Kappen, 2005; Levine, 2018). For language models, this view appears in RL framed as Bayesian inference (Korbak et al., 2022), in the model read as a soft Q-function (Rafailov et al., 2024), and in twisted sequential Monte Carlo (Zhao et al., 2024a). Feynman–Kac steering applies it to diffusion models (Singhal et al., 2025), and generator matching unifies diffusion, flow and jump processes (Holderrieth et al., 2025). We build on these results; our contribution is to connect them to the network's internal transport. **Transformers as dynamical systems.** Transformers have been modelled as particle systems and as PDEs with diffusion, reaction and non-local terms (Lu et al., 2019; Geshkovski et al., 2023; Zhang & Zhou, 2024). Attention has been analysed as a non-local integral operator (Yu et al., 2024) and as a low-pass filter over tokens (Wang et al., 2022), and fractional neural attention is *designed* around the fractional Laplacian (Qu et al., 2025). Brändel et al. (2026) measure embedding-to-output Jacobian sensitivity in trained language models, find power-law-type

decay with distance, and control for it with shuffled text and random initialisation; the power-law observation is theirs. Attention distance grows with depth (Vig & Belinkov, 2019), and natural language itself has power-law correlations (Lin & Tegmark, 2017). Retrieval heads explain long-context factuality in ablation studies (Wu et al., 2024), and synthetic context-extension data changes these heads in ways that track downstream performance (Zhao et al., 2024b); our head analysis follows this line. What we add is a residual-stream response resolved by layer and head and, mainly, how post-training changes it. **How post-training differs.** RL updates are small and avoid principal weight directions (Mukherjee et al., 2025; Zhu et al., 2025), forget less (Shenfeld et al., 2025), and raise activation intensity (Zhang et al., 2025). Nie (2026) argues from behaviour that post-training methods differ mainly in the states they train on; our bridge theorem and measurements give the internal counterpart. Chain-of-thought fine-tuning can hurt needle-in-a-haystack retrieval in hybrid linear-attention models (Zhou et al., 2026), and replaying earlier insights helps long reasoning (Lei et al., 2026).

3 Measuring state and referential persistence

Let e_x be the input embedding at position x and $h_\ell(y)$ the residual stream at layer ℓ and position y. We add ϵv to e_{x_0} and record the relative response

$$G_{\ell,v}(r) = \mathbb{E}\,\frac{\|h_\ell(x_0 + r; e_{x_0} + \epsilon v) - h_\ell(x_0 + r; e_{x_0})\|}{\|h_\ell(x_0 + r; e_{x_0})\|}, \qquad r \geq 0, \tag{1}$$

averaged over prompts. We use three perturbation directions of equal norm ($\epsilon\|v\| = 0.02\,\overline{\|e\|}$): **S** moves the first digit of a stored four-digit value toward another digit, **R** is a random direction at the same token, and **N** moves an irrelevant colour word. Each prompt writes the value, adds about $1{,}500$ tokens of filler and then asks for the value. The *raw-text* probe uses variable assignments. The *reasoning-style* probe ("own" in tables and figures) writes the value inside the model's chat template and reasoning turn, after `<think>` for R1-style models, with arithmetic filler. It is closer to the RL training distribution than raw text, but it is not sampled from the policy. It differs from raw text in template, in filler, and in competition: the raw-text probe contains 2–40 assignments of the same form before the target and many after it, whereas the reasoning-style probe has a single, uniquely cued value. All models in a comparison see identical tokens, with 120 prompts per run. Qwen's tokenizer splits numbers into single digits; OLMo's groups digits into chunks of up to three, so for OLMo the state token is the first three-digit chunk and S changes its first digit. Each checkpoint builds v from its own embeddings. For the Qwen pair, the digit-difference directions agree to a cosine of at least 0.99999 and the mean embedding norms to five digits, so the perturbation is effectively the same (we did not check this for OLMo).

For retrieval, we also record the attention from every later position back to the state token, the effect of ablating the heads that read it from far away, and the answer's sensitivity to the state: the change in the logit gap between the true and a counterfactual digit under the S perturbation. Note that G is the norm of a Jacobian–vector product. It measures how strongly the perturbation reaches later positions, not whether the value can be decoded there; a readout sees only the part of the response along its own direction (Section 5.3). At the last layer we use the hidden state after the final normalisation, which is what the unembedding reads; earlier layers are the residual stream. We summarise decay by the power-law exponent p of $G_{L,S}$ ($G \propto r^{-p}$; smaller p means slower decay), from an unweighted least-squares fit of $\log G$ on $\log r$ over integer r from 16 to the end of the window (about $1{,}500$ tokens), so long distances dominate the fit. Because a constant floor can absorb part of the long-range signal (Brändel et al., 2026), we also fit $G = g + b(r+1)^{-p}$ (Table 7). We report paired bootstrap 95% intervals over prompts. **Controls.** Doubling ϵ doubles the response (ratio 2.00–2.06 in every run), so we are measuring linearised dynamics. When the source lies at least ten lines into the prompt, the averaged response depends on distance rather than absolute position (stationarity gap 0.05–0.14). This fails near the start of the prompt, where attention sinks dominate (Xiao et al., 2024), which affects the Qwen raw-text runs (0.38–0.39). We therefore re-estimate every effect on the half of the prompts whose source lies later (the "position control"). All checks are in Table 4.

4 A dynamical-systems view: transport, retrieval and training

This section sets out our vocabulary and the consequences we test. Formal statements and proofs are in Appendix A.

Internal dynamics (a phenomenological model). Writing u for the propagated signal and τ for depth, we describe the data (Section 5.1) with a causal fractional transport model,

$$\partial_\tau u = \rho(\tau)\, u - c(\tau)\, \mathcal{D}_+^{\alpha(\tau)} u + s, \qquad \mathcal{D}_+^{\alpha} u(x) \propto \int_0^\infty \big(u(x) - u(x-r)\big)\, r^{-1-\alpha}\, dr, \tag{2}$$

where ρ is a per-token reaction (the MLP and attention to the token's own position), $\mathcal{D}_+^\alpha$ is a one-sided fractional derivative (one-sided because attention only looks back; Marchaud form for $0 < \alpha < 1$), and s is a source (writing or restating the state). At $\alpha = 1$ the operator reduces to advection; local diffusion belongs to the symmetric family $-c(-\Delta)^{\alpha/2}$, not to this one-sided family. The model is a hypothesis. Power laws also arise from mixtures of decay rates, from heterogeneous heads and from multi-hop paths, so a good fit does not identify the operator. Two properties guide our measurements. First, a completely monotone response is a superposition of exponential decays, and a power law r^{-p} corresponds to decay rates that accumulate at zero (Proposition 1); if our envelopes were completely monotone, which we do not verify, p would summarise the low-frequency end of the token spectrum. Second, the steady state of Equation (2) decays as $r^{\alpha-1}$, but a discrete residual stack whose layers have one-sided kernels with tail $r^{-1-\alpha}$ keeps that single-layer exponent at long range, and a finite stack has no poles in depth (Theorem 2). Depth alone therefore does not change the long-range exponent. Qwen's final-layer exponent is below 1 with or without a constant floor (0.35 and 0.73), which the asymptotic tail of such a stack ($1 + \alpha$) cannot produce, so its response is pre-asymptotic; for OLMo the answer depends on the fitted form (1.28 with a floor). Either way α is not identified by our data, and we use the model as vocabulary rather than as a fitted mechanism.

Transport, readout and selection. Let z be the answer's logit gap at the query position q, w its readout direction, and $J = \partial h_L(q)/\partial e_{x_0}$, with h_L taken after the final normalisation. To first order, a perturbation ϵv changes the answer by $\delta z = \epsilon \, \|Jv\| \, \|w\| \cos\theta_v$, that is, transport (what reaches q, which G measures) times readout coupling. Whether the answer prefers the true digit to this counterfactual depends on the sign of the unperturbed margin, which neither factor determines, and producing the true digit requires beating every other token as well. If the read attends over the correct state and m competitors with score margins Δ_j, the weight on the correct state is $\pi_0 = (1 + \sum_j e^{-\Delta_j})^{-1}$, however well each value was transported, so $\pi_0 \geq 1/2$ requires margins of order $\ln m$ (Proposition 5). A read can therefore fail through transport (the state's age), through readout coupling, or through selection among competitors, and G measures only the first.

Two failure modes. A post-order traversal of a dense binary tree returns to each node only after visiting its whole subtree, so the pointer it needs has age $|T(v)| - 1$. Suppose a read of age a succeeds from memory with probability $\Phi((a/a^*)^{-p})$, where $a^* = (C/\sigma)^{1/p}$ is the horizon of a signal Ca^{-p} against noise σ (Proposition 2), and otherwise succeeds through retrieval with probability q_{ret}. Then success falls to $1/2$ at depth $d_{1/2} = \frac{1}{p}\log_2 \frac{C}{\sigma} + \log_2 \frac{1}{1-q_{\text{ret}}} + O(1)$ (Proposition 4). Memory decay (state persistence) and pointer retrieval (referential persistence) contribute separate terms; restating caps the age of each read but does not remove the limit. We have not tested the law on real traversals. Its role here is to motivate measuring influence and retrieval separately, which Section 5.3 shows is necessary.

Training as an intervention. At the level of generation (established theory), the KL-regularised optimum is $\pi^* \propto \pi_{\text{ref}}\, e^{R/\beta}$, and its desirability $\psi = e^{V/\beta}$ solves a linear advection–diffusion–reaction equation, $\partial_t \psi + b \cdot \nabla\psi + \frac{1}{2}\sigma^2 \Delta\psi + (r/\beta)\psi = 0$, in which next-token prediction learns the drift b and the reward r is the reaction term (Kappen, 2005; Todorov, 2006). In this sense RL adds the reaction term that next-token prediction lacks. Inside the network, a first-order identity links an update to the transport: after one policy-gradient step, the change in any differentiable internal property F at a context x is an advantage-weighted average, over the policy's own states, of their tangent-kernel alignment with x; for supervised learning, the average is over the dataset (Theorem 1). The identity bounds how far weakly aligned probes can move, but it does not tell us which probes are weakly aligned, and we do not measure the alignment. In a scalar model, short-range targets push p up and long-range targets push it down (Proposition 3). A loss dominated by local targets, such as cross-entropy on reasoning text, should therefore steepen decay, although the crossover fraction (about 37% for illustrative parameters) is not calibrated for LLM training. Restating laws and an exposure theorem for a persistence–refresh model are in Appendices A and B.

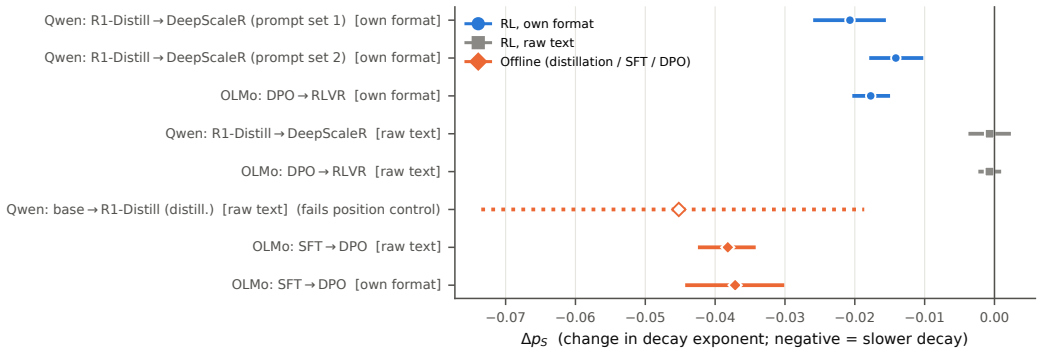

Figure 1: Change in the decay exponent of the state, Δp_S (negative = slower decay), with paired bootstrap 95% CIs over 120 prompts. RL steps (blue) are significant on the reasoning-style (own-format) probe and null on raw text; OLMo's single exponent is a window-averaged summary that reverses sign when a constant floor is fitted (Table 7). The Qwen distillation step (hollow) loses significance under the position control and is inconclusive.

5 Results

Checkpoints. Qwen 1.5B: Qwen2.5-Math-1.5B → DeepSeek-R1-Distill-Qwen-1.5B (distillation; Yang et al., 2024; DeepSeek-AI, 2025) → DeepScaleR-1.5B-Preview (GRPO; Luo et al., 2025). OLMo-2 1B: SFT → DPO (Rafailov et al., 2023) → RLVR1 (Team OLMo, 2025; Lambert et al., 2024). All forward passes run in fp32. The two RL steps, R1-Distill→DeepScaleR and DPO→RLVR, form our two independent RL pairs.

5.1 State persistence: the response decays as a power law over the measured range

On the reasoning-style probe, a power law fits better than a single exponential, a weak baseline, for *every* one of the 480 prompts across four checkpoints (Table 4). Because the fit range (16 to about 1,500 tokens) spans less than two decades, this is a description rather than an asymptotic law, and a non-negative mixture of 3–8 decay rates fits better still (4–6% median error, against about 12% for the power law and 32–34% for a single exponential in Qwen). The functional form therefore does not identify a mechanism. Measured at every layer of the Qwen pair (24 prompts), the response is a power law from the first layer on, and its exponent falls with depth (Figure 2b). A tempered fit, $\log G = \log c - p \log r - \lambda r$, selects $\lambda \approx 0$ at every Qwen layer after the first, so there is no cutoff within 1,400 tokens, although one beyond the window remains possible. We ran this fit only for Qwen; OLMo's local exponent rises at long range (below), which is consistent with a cutoff. The remaining oscillation in G has a period equal to the number of tokens per filler line in 7 of 8 runs (17.2 vs. 17.3, 11.0 vs. 10.9, 8.1 vs. 8.2, 6.1 vs. 6.2): these are spectral lines at the line frequency. Because OLMo's local exponent varies with distance, its single exponent is only a window-averaged summary and depends on the fitted form (Table 7); we report it for comparability and rely for OLMo on band-resolved exponents (Table 5) and fit-free response levels.

Coefficients (Table 6). The exponent falls from 1.44 [1.38, 1.47] at layer 1 to 0.73 at layer 8, 0.45 at layer 16 and 0.35 [0.32, 0.38] at layer 28. Layer 1's exponent is what one hop through a kernel with tail $r^{-1-\alpha}$, $\alpha \approx 0.44$, would give (Theorem 2(i)). By (iii), a stack of such layers keeps that exponent asymptotically, so within our window the fall reflects either later layers whose own attention reaches further (far-reading heads concentrate in layers 14–17; Section 5.4) or multi-hop paths at intermediate range; our data do not separate the two. With a floor, the exponent still falls from 1.26 at layer 1 to 0.96 at layer 8 and 0.76 at layer 16, but then stays at 0.76 while the floor grows from 61% to 82% of the response at 1,000 tokens, so part of the late fall in the pure fit is a growing floor. RL lowers the exponent in the middle and late layers and reduces the drop at the last layer (ρ_{28}: $-0.12 \to -0.03$), a step that includes the final normalisation. OLMo's local exponent rises from 0.25 at 128–512 tokens to 0.50 at 512–1,400 (Appendix F).

Table 1: Every post-training step. Δp is the change in the power-law decay exponent (negative = slower decay) for the state (S), random (R) and nuisance (N) perturbations; ΔSNPR is the relative change in state-to-nuisance response energy. Brackets: paired bootstrap 95% CI over prompts; *: CI excludes zero.

Step	Probe	Δp_S [95% CI]	late-source half	Δp_R	Δp_N	ΔSNPR
Qwen: R1-Distill→DeepScaleR (prompt set 1)	own	-0.021 $[-0.026, -0.016]^*$	-0.021^*	-0.020	-0.049	-2.7%
Qwen: R1-Distill→DeepScaleR (prompt set 2)	own	-0.014 $[-0.018, -0.010]^*$	-0.013^*	-0.016	-0.050	-0.5%
OLMo: DPO→RLVR	own	-0.018 $[-0.020, -0.015]^*$	-0.014^*	-0.019	-0.004	$+16.9\%$
Qwen: R1-Distill→DeepScaleR	raw	-0.001 $[-0.003, +0.002]$	-0.002	-0.000	$+0.001$	-8.2%
OLMo: DPO→RLVR	raw	-0.001 $[-0.002, +0.001]$	-0.000	$+0.001$	$+0.001$	-1.4%
Qwen: base→R1-Distill (distill.)	raw	-0.045 $[-0.073, -0.019]^*$	-0.012	-0.042	-0.045	-51.1%
OLMo: SFT→DPO	raw	-0.038 $[-0.042, -0.034]^*$	-0.041^*	-0.043	-0.012	-9.2%
OLMo: SFT→DPO	own	-0.037 $[-0.044, -0.030]^*$	-0.024^*	-0.053	$+0.032$	$+6.1\%$

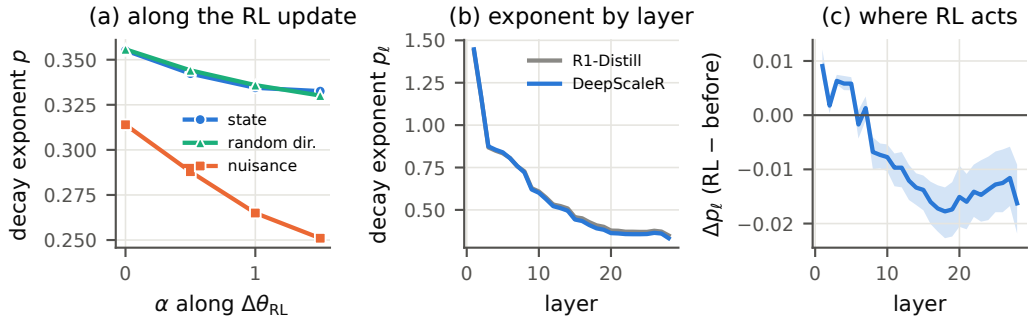

Figure 2: Qwen pair (R1-Distill→DeepScaleR, reasoning-style probe). (a) Decay exponents along the interpolated RL update $\theta(\alpha) = \theta_{\text{R1D}} + \alpha\Delta\theta_{\text{RL}}$. (b) Exponent at every layer. (c) Change in the exponent at every layer, with 95% bootstrap band (24 prompts).

5.2 Training as an intervention: RL slows decay in Qwen on the reasoning-style probe

Figure 1 and Table 1 summarise every step. On the reasoning-style probe, RL lowers p_S in the Qwen pair by -0.021 $[-0.026, -0.016]$, replicated on an independent prompt set (-0.014 $[-0.018, -0.010]$). Both changes survive the position control and a fit with a constant floor (-0.024 $[-0.048, -0.001]$ and -0.033 $[-0.049, -0.020]$; Table 7). With a floor, RL both raises the floor (by 19–23%) and slows the decaying part, so it raises long-range influence in two ways. The changes are also far larger than generic drift: random updates with the same per-tensor norms as the RL update change p_S by only -0.0002 (three seeds, all CIs within $[-0.0004, +0.0001]$), against -0.015 $[-0.018, -0.011]$ for RL on the same 60 prompts (Table 10). Fit-free, DeepScaleR carries 23% more state signal than R1-Distill at 1,000 tokens on this probe, against a flat 3% offset on raw text (Figure 3, Appendix D). On raw text, the same step leaves p_S essentially unchanged (-0.001, CI containing zero; a floor fit gives a small increase, $+0.007$ $[+0.000, +0.016]$), although R1-Distill recalls only a third of the raw-text values there. In OLMo, the pure power-law fit gives -0.018 $[-0.020, -0.015]$ for RLVR on the reasoning-style probe, but the floor fit reverses the sign ($+0.082$), the floor carries 96% of the signal at 1,000 tokens, and RLVR carries 3–5% less state signal than DPO over most of the window. We therefore do not claim that RL slows decay in OLMo. The probe dependence in Qwen is consistent with Theorem 1 if the reasoning-style probe is more kernel-aligned with the rollouts, which we do not measure; the probes differ in template, filler and number of competitors, and DeepScaleR's RL also lengthened its context from 8K to 24K tokens, so long-context exposure may contribute (Theorem 3). Resolved by distance (Table 5), Qwen's change concentrates at the longest range, while OLMo's local exponents move in opposite directions at short and long range.

5.3 Does the change reach the answer?

G is not direction-selective (Section 5.4), so it could change without affecting what the model reads out. We therefore measure, for each prompt, how much the S perturbation changes the answer's logit gap between the true and a counterfactual digit (Table 2). The change points toward the counterfac-

Table 2: Answer-level sensitivity to the state: ratio of the mean absolute change in the answer's logit gap under the S perturbation, after vs. before each step, with paired bootstrap 95% CIs over 120 prompts (*: excludes 1). Sign ok: share of prompts whose effect points toward the counterfactual (minimum of before and after). S/R: effect of the state direction relative to a random direction at the same token, before→after, with their ratio. Clean recall: teacher-forced exact match of the unperturbed answer, with an exact McNemar p when below 0.05.

Step	Probe	$\lvert\Delta\text{ans}_S\rvert$ after/before [95% CI]	late-source half	sign ok	S/R before→after	clean recall
Qwen: R1-Distill→DeepScaleR (set 1)	own	1.211 [1.192, 1.231]*	1.186*	100%	15.4 → 14.6 (×0.95*)	1.00→1.00
Qwen: R1-Distill→DeepScaleR (set 2)	own	1.191 [1.168, 1.212]*	1.175*	100%	13.2 → 12.7 (×0.96*)	1.00→1.00
Qwen: R1-Distill→DeepScaleR	raw	0.983 [0.944, 1.025]	0.980	91%	8.4 → 8.4 (×1.00)	0.33→0.32
OLMo: DPO→RLVR	own	0.995 [0.984, 1.006]	0.986	100%	7.7 → 7.7 (×1.00)	0.85→0.88
OLMo: DPO→RLVR	raw	0.982 [0.974, 0.989]*	0.970*	100%	8.6 → 8.5 (×0.99)	0.97→0.95
OLMo: SFT→DPO	own	1.242 [1.195, 1.289]*	1.213*	100%	7.8 → 7.7 (×0.98)	0.97→0.85 (p=6e − 05)
OLMo: SFT→DPO	raw	1.574 [1.537, 1.613]*	1.581*	100%	7.5 → 8.6 (×1.15*)	0.95→0.97

Table 3: Per-prompt discrimination of recall failures within each checkpoint that has failures: AUC for success (0.5: no signal) with stratified bootstrap 95% CIs. Slow decay: minus the per-prompt exponent, fitted as the population exponent. Last column: paired AUC difference between answer sensitivity and slow decay (*: CI excludes 0). All measures use the known answer.

Checkpoint	failures	slow decay ($-p$)	G_S, $r \in [512, 1024)$	G_S at answer	$\lvert\Delta\text{ans}_S\rvert$	answer log-prob	$\lvert\Delta\text{ans}_S\rvert$ vs. $-p$
OLMo DPO (own)	18/120	0.52 [0.38, 0.67]	0.57 [0.43, 0.70]	0.58 [0.43, 0.72]	0.65 [0.51, 0.79]	1.00 [0.99, 1.00]	+0.13 [−0.04, +0.29]
OLMo RLVR (own)	15/120	0.60 [0.45, 0.75]	0.64 [0.51, 0.77]	0.60 [0.45, 0.75]	0.70 [0.54, 0.83]	1.00 [0.99, 1.00]	+0.10 [−0.08, +0.27]
Qwen R1-Distill (raw)	80/120	0.50 [0.40, 0.62]	0.50 [0.39, 0.61]	0.59 [0.48, 0.69]	0.76 [0.67, 0.84]	0.88 [0.81, 0.94]	+0.26 [+0.11, +0.40]*
Qwen DeepScaleR (raw)	82/120	0.49 [0.38, 0.61]	0.47 [0.36, 0.58]	0.57 [0.46, 0.67]	0.73 [0.62, 0.82]	0.89 [0.83, 0.95]	+0.24 [+0.08, +0.38]*

tual, as expected, in every reasoning-style prompt of both families. **Magnitude.** In the Qwen pair, RL raises the answer effect by 21% [19, 23] on the reasoning-style probe (19% on the second prompt set, and robust to the position control) and leaves it unchanged on raw text, in line with the change in p_S. In the OLMo pair, RLVR leaves the answer effect unchanged; OLMo's answer-level change comes from the DPO step. **Selectivity.** The state direction moves the answer 7–15 times more than a random direction at the same token, but RL does not increase this ratio. In the Qwen pair, the effects of the random direction and of the colour word rise by 28% and 37%, so S/R falls slightly ($\times 0.95$ $[0.91, 0.99]$). RL thus makes the answer more sensitive to distant tokens in general. Part of this may reflect sharper output logits (the answer's log-probability rises from -0.091 to -0.067); we cannot separate the two. **Recall.** On the reasoning-style probe, OLMo's DPO step lowers recall from 97.5% to 85% (15 prompts lost, none gained; McNemar $p = 6 \times 10^{-5}$), even though it raises the state response 1,000 tokens away by 73% and at the answer by 37% (Table 7), and the answer's sensitivity to the state by 24%. The prompts it broke gained as much far-range influence as those it kept (response at $r \in [512, 1024)$: $\times 1.72$ vs. $\times 1.76$), but their answer sensitivity barely rose ($\times 1.06$ vs. $\times 1.28$). **Per-prompt discrimination** (Table 3). In the four checkpoints with at least 15 failures (15–82 of 120 prompts), we detect no reliable link between the per-prompt exponent and failure: its AUC is 0.49–0.60, with all 95% intervals inside $[0.38, 0.75]$ ($[0.38, 0.62]$ for Qwen), which rules out strong discrimination. The fit-free far-range and answer-position responses do no better. Answer-level sensitivity reaches an AUC of 0.65–0.76 and beats the exponent in paired comparisons in both Qwen conditions ($+0.26$, $+0.24$), though not significantly in OLMo. The answer's own log-probability does better still (0.88–1.00), and on Qwen, answer sensitivity adds no significant information beyond it (likelihood-ratio $p = 0.21$ and 0.97; in OLMo the separation is too complete to test). All of these measures use the known answer, so they are retrospective analyses, not failure detectors. Long-range propagation, answer-level sensitivity and correct selection are therefore distinct quantities. The exponent measures influence across a population of prompts, not memory in an individual prompt, and slower decay is not evidence of better recall.

5.4 Where RL acts (Qwen pair)

By layer. RL makes layers 1–5 slightly *faster*-decaying ($+0.002$ to $+0.009$), leaves layers 6–7 unchanged, and makes layers 8–28 slower-decaying, reaching -0.018 at layer 18 (Figure 2c; significant at every layer from 8 to 28). The slower decay thus arises in the middle and late layers. **The retrieval route: heads.** This analysis measures the route by which distant positions read the state, not whether retrieval succeeds. Distant tokens rarely read the state token directly: only 13 of 336 heads attend to it from $r \geq 256$ more than a uniform spread would, against 285 heads from $r < 16$. Long-range signal

therefore travels through diffuse attention and relaying across layers. After RL, far-range attention to the state rises in 231 of 336 heads ($+4.1\%$ in total, against -0.1% at near range), a broad shift rather than a change in a few heads. **Ablation and behaviour** (Table 9). We zero the 12 heads with the most far-range attention to the state in R1-Distill, or one of five control sets of 12 heads (three random, two matched to the same layers), on 40 prompts per probe. On the reasoning-style probe, the targeted ablation steepens decay by $+0.077$ [$+0.063, +0.092$], more than any control (-0.022 to $+0.034$), so these heads carry a disproportionate share of the tail. Yet recall is unchanged (40/40; on raw text 14/40 before and 13/40 after, not significant), and the answer effect falls by 15%, within the range of the controls. Conversely, on raw text, where the value must be selected among many competing assignments, two control sets that share no head abolish recall (14/40 to 0/40; McNemar $p = 10^{-4}$) while leaving reasoning-style recall intact (at least 39/40). The effect on decay does not predict the effect on recall: two random sets change raw-text decay by $+0.028$ and $+0.029$, yet one leaves recall untouched and the other removes it entirely. This is not generic damage: the two recall-abolishing sets leave the perplexity of the raw-text prompts unchanged ($\times 0.97$), whereas random set 1 raises perplexity by 17–19% without affecting recall (Table 10). The heads that carry long-range influence are thus not the heads that recall depends on, which makes the dissociation between influence and recall causal rather than only correlational. **Not direction-selective.** In every RL step, the random direction at the state token changes as much as the state direction (Δp_R vs. Δp_S: $-0.020/-0.021$, $-0.016/-0.014$, $-0.019/-0.018$). In the residual stream, the change acts on the token as a whole rather than on a feature direction, even though the answer itself is direction-selective (Section 5.3). State-versus-nuisance comparisons (SNPR: -2.7% and -0.5% for DeepScaleR, $+16.9\%$ for OLMo RLVR) involve different tokens and positions and are confounded by token type, so we do not read them as task selectivity.

5.5 Supervised objectives versus RL

DPO raises long-range influence in OLMo on both probes (the state response 1,000 tokens away rises by 73% and 61%; Table 7). Its pure power-law exponent falls on both probes (-0.037 and -0.038, both surviving the position control), but on the reasoning-style probe a floor fit reverses the sign ($+0.36$). Tülu-3-style DPO also uses partly on-policy completions, so it does not cleanly separate offline from on-policy training. The only cross-entropy step among the public checkpoints, Qwen distillation (-0.045), is driven by prompts whose source lies near the start and loses significance under the position control (-0.012 [$-0.044, +0.017$]). We therefore make no claim about cross-entropy from the public checkpoints and test it directly instead. **Fine-tuning the same model.** We fully fine-tune R1-Distill with cross-entropy for 250 steps of 8 sequences on synthetic reasoning traces. The two arms are identical except that a stored value is restated every four lines (short gaps) or never (long gaps), and the traces use different templates from the probe (Appendix E). Both arms *steepen* decay on the reasoning-style probe: $\Delta p_S = +0.065, +0.076, +0.091$ (long gaps) and $+0.069, +0.078, +0.116$ (short gaps) at steps 25, 100 and 250, all with 95% CIs excluding zero, at $\|\Delta\theta\| = 3.1$–5.6 (Table 8). The change transfers to raw text ($+0.027$ and $+0.036$). RL from the same starting model instead *flattens* decay (-0.021 at $\|\Delta\theta\| = 1.83$) and does not transfer. The two procedures therefore move the response in opposite directions. They differ in data, difficulty, budget and update size as well as in objective, with a single run per arm, so this comparison does not isolate the objective. Our prediction that long-gap training would slow decay failed: long gaps only reduced the steepening ($+0.091$ vs. $+0.116$ at step 250). Two parts of our model are consistent with this, although we reconciled them with the result only after seeing it. First, the recall loss reached zero within 25 steps in both arms, so the long gaps were well within the model's horizon and, by the learnability condition of Theorem 3, created no pressure. Second, only 8 of about 1,500 supervised tokens per trace are long-range, far below the long-range fraction at which the model of Proposition 3 turns the net pressure negative (about 37% for its illustrative parameters). This agrees with reports that chain-of-thought fine-tuning can hurt long-range retrieval in hybrid linear-attention models (Zhou et al., 2026).

5.6 A controlled test of exposure

To separate exposure from reward, we train a 4-layer transformer with learnable attention decay (ALiBi slopes; Press et al., 2022) on a task in which a state is read at every step through a query that later positions cannot see, and the model may restate the state. We measure ℓ^*, the age at which read accuracy with restating disabled falls below 75%, over two seeds. All continuations start from the

same supervised model, trained on traces that restate about every four steps ($\ell^* = 10/9$ for the two seeds).

- **Exposure without reward.** Further training on the same short-gap data leaves ℓ^* at $10/10$. Supervised training on traces that never restate raises it to $46/37$. Since this arm trains directly on the unrestated condition that ℓ^* measures, we read it as evidence that exposure is enough, not as a comparison of objectives.
- **Passive versus active persistence.** RL with free restating keeps restating every 4.3 steps and leaves ℓ^* at $10/9$. With a cost per restatement, it restates about twice per 32-step episode, at ages near 12 (Figure 4c), *and* raises ℓ^* to 14–16, so it restates at an age close to its own horizon (Appendix G), as the restating law predicts.

6 Discussion and limitations

What did not hold. Our starting hypothesis was local advection–diffusion–reaction, with one decay pole per Fourier mode that RL moves for task-relevant modes. It failed on two assumptions: a short-range kernel renormalises into local transport with exponential decay, whereas the data are power-law-like; and identical layers would give poles in depth, whereas the layers differ and a finite stack has none. Showing that the averaged envelope depends on distance also does not justify Fourier-diagonalising the linearised operator. Task-selective stabilisation failed in the residual stream, as did the prediction that supervised training on long gaps would slow decay (Section 5.5).

What the probe does and does not show. The distance dependence of the response is a reproducible, controlled measurement, and post-training changes it in a probe-dependent way, robustly in Qwen. But it measures influence, not storage. In causal decoding, the keys and values at the source are not overwritten as tokens accumulate, so slower decay of G means that the source keeps more influence on later positions, not that a stored representation degrades less. We detect no reliable per-prompt link between the exponent and recall failures, and a DPO step that raises long-range influence lowers recall. Head ablation separates the heads that carry influence from the heads that recall needs, but we have not separated storage failures from retrieval failures within a prompt, and the tree task is at floor.

Limitations. The headline RL effect is small (about 6% of p), robust to the fitted form only in Qwen, and comes from one run per pair; our confidence intervals cover prompt sampling, not variation across RL runs, and the null baseline uses three random directions at one scale. DeepScaleR's context lengthening confounds RL with long-context exposure. The layer, head and ablation analyses cover only the Qwen pair (24 and 40 prompts), with ablations in R1-Distill only, and the models have 1–1.5B parameters. G, a ratio, measures propagated influence, not decodable information; we have not measured decodability with probes. Separating the three differences between our probes (template, filler and number of competitors) would need a factorial design. The per-prompt analyses cover four conditions in two families, rely on the known answer, and lack the unperturbed logit margin and output entropy as further baselines. The fit range spans less than two decades, and the fractional model is not identified.

7 Conclusion

We separated two reasons a Transformer can fail to revisit the right state, its influence not reaching the query and the query failing to select and read it out, and showed that they come apart. Post-training can strengthen long-range influence: in the Qwen pair, RL slows its decay on a reasoning-style probe, far beyond what random updates of the same size do. Yet influence does not buy recall. A DPO step that raises long-range influence lowers recall, the decay exponent shows no reliable link to which prompts fail, and removing the heads that carry the most long-range influence spares recall, whereas removing others, without general damage, abolishes it when the value must be selected among competitors. Influence measures, such as Green's-function exponents and attention distance, are therefore not proxies for memory: methods meant to help models revisit earlier states, whether training objectives or architectural changes such as momentum-like depth dynamics and structured addressing, should be evaluated by retrieval under competition and by readout.

AI USE STATEMENT

The research idea, the dynamical-systems framing and the choice of experiments are the authors'. The authors ran the main model experiments. Generative AI (Claude) was used to implement the experiment and analysis code, including the perturbation probe and the controlled-transformer experiment; to search the literature; and to draft the text and figures. Every reported number is produced by scripts included in the supplementary material, and the authors reviewed all AI-assisted work and take responsibility for the content.

REPRODUCIBILITY STATEMENT

Probe, analysis and training code, per-prompt response curves for every checkpoint, and scripts that produce every figure and table from those files are in the supplementary material. All checkpoints are public; Section 3 and Appendices D and G give every hyperparameter.

REFERENCES

N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular Variation*. Cambridge University Press, 1987.

Matthias Brändel, Stephan Köhler, and Oliver Rheinbach. How token influence decays with distance: A green-function view of trained language models. *arXiv preprint arXiv:2606.29139*, 2026.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

William Feller. *An Introduction to Probability Theory and Its Applications, Vol. II*. Wiley, 2nd edition, 1971.

Sergey Foss, Dmitry Korshunov, and Stan Zachary. *An Introduction to Heavy-Tailed and Subexponential Distributions*. Springer, 2 edition, 2013.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. *arXiv preprint arXiv:2312.10794*, 2023.

Peter Holderrieth et al. Generator matching: Generative modeling with arbitrary Markov processes. In *International Conference on Learning Representations*, 2025.

Hilbert J. Kappen. Linear theory for control of nonlinear stochastic systems. *Physical Review Letters*, 95:200201, 2005.

Tomasz Korbak, Ethan Perez, and Christopher L. Buckley. RL with KL penalties is better viewed as Bayesian inference. In *Findings of the Association for Computational Linguistics: EMNLP*, 2022.

Nathan Lambert et al. Tülu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Bin Lei, Caiwen Ding, Jiachen Yang, Ang Li, and Xin Eric Wang. Stateful reasoning via insight replay. *arXiv preprint arXiv:2605.14457*, 2026.

Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018.

Henry W. Lin and Max Tegmark. Critical behavior in physics and probabilistic formal languages. *Entropy*, 19(7):299, 2017.

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*, 2024.

Yiping Lu, Zhuohan Li, Di He, Zhiqing Sun, Bin Dong, Tao Qin, Liwei Wang, and Tie-Yan Liu. Understanding and improving transformer from a multi-particle dynamic system point of view. *arXiv preprint arXiv:1906.02762*, 2019.

Michael Luo et al. DeepScaleR: Surpassing O1-Preview with a 1.5B model by scaling RL. Agentica project blog post and model card (agentica-org/DeepScaleR-1.5B-Preview), 2025.

Aman Mehta and Anupam Datta. Plans don't persist: Why context management is load bearing for LLM agents. *arXiv preprint arXiv:2606.22953*, 2026.

Sagnik Mukherjee, Lifan Yuan, Dilek Hakkani-Tür, and Hao Peng. Reinforcement learning finetunes small subnetworks in large language models. *arXiv preprint*, 2025.

Dong Nie. Post-training is about states, not tokens: A state distribution view of SFT, RL, and on-policy distillation. *arXiv preprint arXiv:2605.22731*, 2026.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*, 2022.

Cheng Kevin Qu, Andrew Ly, and Pulin Gong. Fractional neural attention for efficient multiscale sequence processing. *arXiv preprint arXiv:2511.10208*, 2025.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023.

Rafael Rafailov, Joey Hejna, Ryan Park, and Chelsea Finn. From r to q^*: Your language model is secretly a Q-function. *arXiv preprint arXiv:2404.12358*, 2024.

Stefan G. Samko, Anatoly A. Kilbas, and Oleg I. Marichev. *Fractional Integrals and Derivatives: Theory and Applications*. Gordon and Breach, 1993.

Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. RL's razor: Why online reinforcement learning forgets less. *arXiv preprint*, 2025.

Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A general framework for inference-time scaling and steering of diffusion models. *arXiv preprint arXiv:2501.06848*, 2025.

Team OLMo. 2 OLMo 2 furious. *arXiv preprint arXiv:2501.00656*, 2025.

Emanuel Todorov. Linearly-solvable Markov decision problems. In *Advances in Neural Information Processing Systems*, 2006.

Jesse Vig and Yonatan Belinkov. Analyzing the structure of attention in a transformer language model. In *Proceedings of the ACL Workshop BlackboxNLP*, 2019.

Peihao Wang, Wenqing Zheng, Tianlong Chen, and Zhangyang Wang. Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice. In *International Conference on Learning Representations*, 2022.

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

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

An Yang et al. Qwen2.5-Math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

Yue Yu et al. Nonlocal attention operator: Materializing hidden knowledge towards interpretable physics discovery. In *Advances in Neural Information Processing Systems*, 2024.

Honglin Zhang, Qianyue Hao, Fengli Xu, and Yong Li. Reinforcement learning fine-tuning enhances activation intensity and diversity in the internal circuitry of LLMs. *arXiv preprint arXiv:2509.21044*, 2025.

Yukun Zhang and Xueqing Zhou. Understanding transformer architecture through continuous dynamics: A partial differential equation perspective. *arXiv preprint arXiv:2408.09523*, 2024.

Stephen Zhao, Rob Brekelmans, Alireza Makhzani, and Roger Grosse. Probabilistic inference in language models via twisted sequential Monte Carlo. In *International Conference on Machine Learning*, 2024a.

Xinyu Zhao, Fangcong Yin, and Greg Durrett. Understanding synthetic context extension via retrieval heads. *arXiv preprint arXiv:2410.22316*, 2024b.

Xinyu Zhou, Boyu Zhu, Yi Xu, Zhiwei Li, Yingfa Chen, Huiming Wang, and Zhijiang Guo. Attention amnesia in hybrid LLMs: When CoT fine-tuning breaks long-range recall, and how to fix it. *arXiv preprint arXiv:2606.11052*, 2026.

Hanqing Zhu et al. The path not taken: RLVR provably learns off the principals. *arXiv preprint arXiv:2511.08567*, 2025.

A Proofs

Proposition 1 (Spectral representation)**.** *If $r \mapsto G(r)$ is completely monotone (we do not verify this for the data) then $G(r) = \int_0^\infty e^{-\gamma r}\mu(d\gamma)$ for a unique $\mu \geq 0$, and $\hat{G}(k) = \int \mu(d\gamma)/(\gamma + ik)$: each decay rate is a pole at $k = i\gamma$. A single pole gives $e^{-\gamma_0 r}$; if μ has density $\sim C\gamma^{p-1}$ near 0 ($0 < p < 1$), the poles form a branch cut reaching $k = 0$ and $G(r) \sim C\Gamma(p)r^{-p}$, $|\hat{G}(k)| \sim C'k^{p-1}$.*

Proof. Bernstein's theorem (Feller, 1971, XIII.4) gives $G(r) = \int_0^\infty e^{-\gamma r}\mu(d\gamma)$ for a unique $\mu \geq 0$ whenever G is completely monotone. By Fubini, for $k \neq 0$, $\int_0^\infty e^{-ikr}G(r)\,dr = \int \mu(d\gamma)\int_0^\infty e^{-(\gamma+ik)r}dr = \int \mu(d\gamma)/(\gamma + ik)$, finite whenever $\int \mu(d\gamma)/(1+\gamma) < \infty$. If $\mu(d\gamma) = C\gamma^{p-1}(1 + o(1))\,d\gamma$ as $\gamma \to 0$, Karamata's Tauberian theorem (Feller, 1971, XIII.5); (Bingham et al., 1987) gives $G(r) \sim C\,\Gamma(p)\,r^{-p}$, and substituting $\gamma = kt$ gives $\hat{G}(k) = k^{p-1}C\int_0^\infty t^{p-1}(t+i)^{-1}dt\,(1 + o(1))$ as $k \to 0$ for $0 < p < 1$. We do not verify complete monotonicity, which constrains derivatives of every order. Measured envelopes decrease in 69–93% of log-spaced distance bins and are convex in 50–64%, which tests only the first two derivatives; the violations come from the line-period oscillations and from bumps near the source. The proposition is therefore a conditional representation, not a verified property of the data. *Numerical check* (`toy_theorems_check.py`): with $\mu(d\gamma) = \gamma^{p-1}d\gamma$ on $(0, 1]$, tail slopes are -0.500 and -0.800 for $p = 0.5, 0.8$, and low-frequency slopes are -0.497 and -0.205 (expected -0.5, -0.2) for $k \leq 10^{-7}$. At realistic ranges ($k \in [10^{-4}, 10^{-2}]$) the $p = 0.8$ slope is biased (-0.258), which is why we estimate p in the distance domain.

Let $F(\theta; x)$ be any differentiable property of the internal transport at a context x, for example $-p$ or $G_{L,S}(r)$. An optimiser step is $\Delta\theta = \eta Pg$ with a positive-definite preconditioner P.

Theorem 1 (Distribution-bound transport)**.** *Define the tangent kernel $K_P(x; s, a) = \langle \nabla_\theta F(\theta; x), P\nabla_\theta \log \pi_\theta(a \mid s)\rangle$. For a policy-gradient step, $g = \mathbb{E}_{\tau\sim\pi_\theta}[\sum_t A_t \nabla_\theta \log \pi_\theta(a_t \mid s_t)]$, and*

$$\Delta F(x) = \eta\,\mathbb{E}_{\tau\sim\pi_\theta}\Big[\sum_t A_t\,K_P(x; s_t, a_t)\Big] + O(\eta^2).$$

For a supervised step on data $\mathcal{D}$, $\Delta F(x) = \eta\,\mathbb{E}_{(s,a)\sim\mathcal{D}}[K_P(x; s, a)] + O(\eta^2)$. Consequently, if $|K_P(x; s, a)| \leq \kappa(x)$ on the support of the training states, then $|\Delta F(x)| \leq \eta\,\kappa(x)\,\mathbb{E}\sum_t |A_t| + O(\eta^2)$: the transport at x can move only as far as x is kernel-aligned with the states the objective trains on.

Proof. Let $F(\cdot; x)$ be twice differentiable with $\|\nabla^2_\theta F\| \leq M$ along the update. Taylor's theorem gives $F(\theta + \Delta\theta; x) - F(\theta; x) = \langle \nabla_\theta F(\theta; x), \Delta\theta\rangle + R$ with $|R| \leq \frac{M}{2}\|\Delta\theta\|^2 = O(\eta^2)$. For $\Delta\theta = \eta Pg$ and the policy gradient $g = \mathbb{E}_{\tau\sim\pi_\theta}[\sum_t A_t\nabla_\theta \log \pi_\theta(a_t \mid s_t)]$, linearity of the inner product and of the expectation gives $\langle \nabla_\theta F, \eta Pg\rangle = \eta\,\mathbb{E}_\tau[\sum_t A_t\langle \nabla_\theta F, P\nabla_\theta \log \pi_\theta(a_t \mid s_t)\rangle] = \eta\,\mathbb{E}_\tau[\sum_t A_t K_P(x; s_t, a_t)]$. For supervised learning, $g = \mathbb{E}_{(s,a)\sim\mathcal{D}}[\nabla_\theta \log \pi_\theta(a \mid s)]$ (ascent on the log-likelihood), which gives the second formula. The bound follows from $|\mathbb{E}\sum_t A_t K_P| \leq$

$\mathbb{E}\sum_t |A_t||K_P| \le \kappa(x)\,\mathbb{E}\sum_t |A_t|$. *Many steps.* Summing over steps k, $F(\theta_n;x) - F(\theta_0;x) = \sum_k \eta_k \langle \nabla_\theta F(\theta_k;x), P_k g_k\rangle + \sum_k O(\eta_k^2)$. Each first-order term is again a kernel average over the *then-current* policy's states, so the conclusion holds along the whole run as long as the probe stays kernel-misaligned with the rollouts. The interpolation experiment (Figure 2a) measures the aggregate change along the realised update directly. The theorem predicts that the size of ΔF on a probe distribution scales with its kernel alignment to the training states; computing K_P between our probes and actual RL rollouts is a direct test we leave to future work.

Proposition 2 (Horizon amplification)**.** *The usable horizon is $a^* = (C/\sigma)^{1/p}$, so $\Delta \ln a^* = -\frac{\Delta p}{p^2}\ln\frac{C}{\sigma} + \frac{1}{p}\frac{\Delta C}{C}$ to first order. For exponential memory $G = Ce^{-\gamma a}$, $\Delta \ln a^* = -\Delta\gamma/\gamma + O(\Delta C)$: a power law multiplies a relative exponent change by $\ln(C/\sigma)/p$.*

With $p \approx 0.355$ and $\Delta p = -0.021$ (Section 5.2), the horizon grows by a factor $e^{0.167\ln(C/\sigma)}$, e.g. 1.6–2.3× for $\ln(C/\sigma) \in [3,5]$. The readout threshold C/σ is not measured here, and Section 5.3 finds no per-prompt link between G and recall, so this horizon gain is conditional on a link our data do not support. *Proof.* With a threshold readout, or a Gaussian readout with the horizon defined where $G(a^*) = \sigma$ (i.e. $\Phi(1)$ success), $Ca^{*-p} = \sigma$ gives $\ln a^* = \frac{1}{p}\ln\frac{C}{\sigma}$. Differentiating, $\partial \ln a^*/\partial p = -\frac{1}{p^2}\ln\frac{C}{\sigma}$ and $\partial \ln a^*/\partial \ln C = \frac{1}{p}$. For $G = Ce^{-\gamma a}$, $a^* = \gamma^{-1}\ln\frac{C}{\sigma}$, so $\partial \ln a^*/\partial \ln\gamma = -1$ and $\partial \ln a^*/\partial \ln C = 1/\ln\frac{C}{\sigma}$. The ratio of the relative sensitivities to the decay parameter is $\ln(C/\sigma)/p$.

Corollary 1 (Restating law under power-law memory)**.** *If the state is read at every step and each restatement costs c, the cost-optimal restating interval is $\Delta^* = a^* u^*(c/a^*; p)$, where u^* depends only on p and c/a^*. The number of restatements over a horizon T is $T(\sigma/C)^{1/p}/u^*(c/a^*;p)$; because u^* also depends on p, the explicit factor alone does not give the dependence on p.*

Proof. With $G(a) = Ca^{-p}$ and success probability $\Phi(G(a)/\sigma)$, $G(a)/\sigma = (a/a^*)^{-p}$, so the per-read hazard is $H(a) = h_p(a/a^*)$ with $h_p(x) = -\ln\Phi(x^{-p})$, which is increasing. Restating every Δ steps costs $C(\Delta) = [\int_0^\Delta H(a)\,da + c]/\Delta$ per step (continuous relaxation). Substituting $u = \Delta/a^*$, $C = [\int_0^u h_p + c/a^*]/u$, which is Theorem 4 with $\gamma = 1/a^*$ and $H = h_p$. The optimum u^* solves $m_p(u) = uh_p(u) - \int_0^u h_p = c/a^*$, unique when $c/a^* < m_p(\infty)$ since $m_p'(u) = uh_p'(u) > 0$. Hence $\Delta^* = a^* u^*(c/a^*;p)$, and the number of restatements over T steps is $T/\Delta^* = T(\sigma/C)^{1/p}/u^*$.

Theorem 2 (Fractional Green's functions and residual stacks)**.** *Let $0 < \alpha < 1$. The one-sided fractional derivative has Fourier symbol $(ik)^\alpha$. (i) One application has a kernel with tail $\propto r^{-1-\alpha}$. (ii) The steady state of $c\,\mathcal{D}_+^\alpha u = \delta_{x_0}$ is $u(x_0+r) = r^{\alpha-1}/(c\,\Gamma(\alpha))$, a power law with $p = 1-\alpha$. (iii) A discrete residual stack does not produce (ii). If layer ℓ acts as $u \mapsto u + \epsilon_\ell K_\ell * u$, where each K_ℓ is a one-sided probability kernel with $K_\ell(r) \sim \kappa_\ell r^{-1-\alpha}$, then $G_L(r) \sim \big[\sum_\ell \epsilon_\ell\kappa_\ell \prod_{j\neq\ell}(1+\epsilon_j)\big] r^{-1-\alpha}$ as $r\to\infty$: depth rescales the single-layer tail but keeps its exponent. (iv) A finite stack has no poles in depth: for layers with transfer $m_\ell(k)$, the depth transform $\sum_{\ell\le L} z^{-\ell}\prod_{j\le\ell} m_j(k)$ is a polynomial in z^{-1}. If each layer acts on the envelope as a scalar multiplier, the local rates $\lambda_\ell(k) = \ln m_\ell(k)$ sum to $\ln\hat{G}_L(k)$.*

Proof. The symbol of the one-sided (Marchaud or Riemann–Liouville) derivative of order $\alpha \in (0,1)$ is $(ik)^\alpha$ (Samko et al., 1993). (i) Its Marchaud form $\int_0^\infty (u(x) - u(x-r))r^{-1-\alpha}dr$ is a jump operator with one-sided jump density $\propto r^{-1-\alpha}$. (ii) The Riemann–Liouville integral $I^\alpha f(x) = \Gamma(\alpha)^{-1}\int_{-\infty}^x (x-y)^{\alpha-1} f(y)\,dy$ inverts $\mathcal{D}_+^\alpha$ on suitable function spaces (Samko et al., 1993), so $u = c^{-1}I^\alpha\delta_{x_0}$. (iii) The linearised stack maps δ_{x_0} to $\prod_{\ell=1}^L (I + \epsilon_\ell K_\ell)\delta_{x_0}$. The factors are convolutions and commute, so expanding the product gives $\sum_{S\subseteq\{1..L\}} \big(\prod_{\ell\in S}\epsilon_\ell\big)\big(*_{\ell\in S} K_\ell\big)$, with $S=\emptyset$ contributing only at $r=0$. For identical layers this is $\sum_j \binom{L}{j}\epsilon^j K^{*j}$, not the unweighted sum $\sum_j K^{*j}$. Densities with regularly varying, eventually decreasing tails of a common index are locally subexponential, so a convolution of finitely many of them satisfies $(*_{\ell\in S}K_\ell)(r) \sim \sum_{\ell\in S}\kappa_\ell r^{-1-\alpha}$: one large jump dominates (Foss et al., 2013). Summing over S, $\sum_S \prod_{j\in S}\epsilon_j \sum_{\ell\in S}\kappa_\ell = \sum_\ell \epsilon_\ell\kappa_\ell\prod_{j\neq\ell}(1+\epsilon_j)$. The Riesz law $r^{\alpha-1}$ of (ii) would need the unweighted sum over hop counts, which a residual stack does not produce. *Numerical check* (`paper/check_depth_tree.py`, 2^{16}-point FFT, identical layers): for $\alpha \in \{0.45, 0.6\}$, $L \in \{16, 28\}$ and $\epsilon \in \{0.05, 0.3, 1\}$, the slope on $r \in [4096, 16384)$ is 1.41–1.45 ($\alpha = 0.45$) and 1.60–1.62 ($\alpha = 0.6$), and $G(8192)$ matches the predicted coefficient to within 5%. With strong coupling, multi-hop paths flatten the response at short

range (slope on $[16, 256)$ down to 0.02 for $\alpha = 0.45$, $L = 28$, $\epsilon = 1$). (iv) $\prod_{j \le \ell} m_j$ is finite for every $\ell \le L$, so the sum is a polynomial of degree L in z^{-1}. For a scalar envelope, $\ln \prod_\ell m_\ell = \sum_\ell \ln m_\ell$ exactly. This holds for tied and untied layers alike. A pole needs infinite depth: for identical layers, $\sum_{\ell \ge 0} z^{-\ell} m^\ell = 1/(1 - m/z)$ has its pole at $z = m$; an infinite sequence of distinct layers can also have one, for example when $m_\ell \to m$.

Let a target at distance a be read through $G(r) = Cr^{-p}$ against interference from all more distant tokens, with signal-to-noise ratio $\mathrm{SNR}(a) = Ca^{-p}/\sqrt{\sigma^2 + \kappa^2 \sum_{r>a} C^2 r^{-2p}}$.

Proposition 3 (Interference pressure). *$\partial_p \log \mathrm{SNR}(a) > 0$ for $a = 1$ and < 0 for a near the context length: short-range targets favour faster decay and long-range targets slower decay. A loss whose targets are mostly short-range therefore pushes p up.*

Proof. Write $S(a, p) = \sum_{r>a} r^{-2p}$. Then $\partial_p \log \mathrm{SNR}(a) = -\ln a + \phi(a)\, \overline{\ln r}(a)$, where $\overline{\ln r}(a) = \sum_{r>a} r^{-2p} \ln r / S(a, p) \ge \ln a$ and $\phi(a) = \kappa^2 C^2 S/(\sigma^2 + \kappa^2 C^2 S) \in (0, 1)$. At $a = 1$ the first term vanishes and the second is positive. As a approaches the context length, $S \to 0$, so $\phi \to 0$ and the derivative tends to $-\ln a < 0$. Numerically ($p = 0.35$, 1,500 tokens, $\sigma = 0.05$, $\kappa = 1$): $+4.35$, $+3.74$, $+2.55$, $+0.81$, -1.56 and -6.22 at $a = 1, 2, 8, 64, 512, 1400$. A mixture of targets at $a = 2$ and $a = 1400$ changes sign at a long-range fraction $f^* = 0.375$.

In a post-order traversal of a dense binary tree with N nodes, returning to node v requires a state written $|T(v)| - 1$ steps earlier. In inorder and pre-order traversal, the corresponding age is the size of v's left subtree. Suppose, as an idealisation, that reads fail independently, that a read of age a succeeds from memory with probability $q_{\text{mem}}(a) = \Phi((a/a^*)^{-p})$, which falls to chance ($1/2$) as $a \to \infty$, and that if memory fails the pointer is recovered by retrieval, e.g. by looking up the parent in the edge list, with probability q_{ret}.

Proposition 4 (Tree depth law). *For $a^* \gg 1$ the traversal succeeds with probability $P = \exp\big(-\tilde{c}\,(N/a^*)(1 - q_{\text{ret}})\big)$, where $\tilde{c}$ lies between positive constants that depend only on p. The depth at which success falls to $1/2$ is $d_{1/2} = \frac{1}{p} \log_2 \frac{C}{\sigma} + \log_2 \frac{1}{1 - q_{\text{ret}}} + O(1)$ for post-order, and one level more for inorder and pre-order.*

Proof. In a dense tree of depth d ($N = 2^d - 1$), the node of height h has $|T(v)| = 2^h - 1$, and there are 2^{d-h} such nodes. Post-order needs age $2^h - 2$ at that node; inorder and pre-order need $2^{h-1} - 1$, one level less. Reads fail independently with probability $f(a) = (1 - q_{\text{mem}}(a))(1 - q_{\text{ret}})$, which rises from ≈ 0 for $a \ll a^*$ to $(1 - q_{\text{ret}})/2$ as $a \to \infty$, because $\Phi(0) = 1/2$. Then $-\ln P = \sum_h 2^{d-h}\big(-\ln(1 - f(2^h - 2))\big)$. Because $f \le (1 - q_{\text{ret}})/2 \le 1/2$, $f \le -\ln(1 - f) \le 2 \ln 2\, f$, so $-\ln P = \tilde{c}\, \frac{N+1}{a^*}(1 - q_{\text{ret}})$ with $c \le \tilde{c} \le 2 \ln 2\, c$, where $c = \frac{a^*}{N+1} \sum_h 2^{d-h}(1 - q_{\text{mem}}(2^h - 2)) \approx \sum_h g(x_h)$ with $x_h = 2^h/a^*$ and $g(x) = (1 - \Phi(x^{-p}))/x$ (ages $\gg 1$, i.e. $a^* \gg 1$). On the geometric grid x_h the sum converges at both ends: $g(x) \approx 1/(2x)$ as $x \to \infty$, and g is Gaussian-small as $x \to 0$. So c, and hence $\tilde{c}$, lies between positive constants that depend only on p (c is log-periodic in a^*). Setting $P = 1/2$ gives $N_{1/2} = a^* \ln 2/(\tilde{c}(1 - q_{\text{ret}}))$ and $d_{1/2} = \log_2 N_{1/2} + O(1)$, with $\log_2 a^* = \frac{1}{p} \log_2 \frac{C}{\sigma}$ by Proposition 2. *Restating.* Capping ages at Δ bounds each read's failure by $f(\Delta) > 0$, but every restatement is itself a read. Under independent failures, exact success still decays exponentially in the number of reads, so restating shifts the limit but does not remove it. *Numerical check* (`paper/check_depth_tree.py`): for $(a^*, p, q_{\text{ret}}) \in \{(200, 0.35, 0), (200, 0.35, 0.9), (400, 0.33, 0), (400, 0.33, 0.9)\}$, the small-failure constant is $c = 0.99, 0.99, 1.12, 1.12$, and the first depth with $P < 1/2$ is $8, 11, 9, 12$ for post-order and $9, 12, 10, 13$ for inorder, against continuous formula values $7.13, 10.45, 7.95, 11.27$ (post-order).

Proposition 5 (Transport, readout and selection). *Let $z(e) = \langle w, h_L(q; e) \rangle + b$ be the answer's logit gap as a function of the input embeddings, with h_L the final hidden state after the final normalisation and w the difference of the two unembedding rows, and $J = \partial h_L(q)/\partial e_{x_0}$. (i) $\delta z = \epsilon \, \|Jv\| \, \|w\| \cos\theta_v + O(\epsilon^2)$, where θ_v is the angle between Jv and w. Hence $\delta z_S/\delta z_R = (\|Jv_S\|/\|Jv_R\|)(\cos\theta_S/\cos\theta_R)$ for two directions at the same token, and under training $\Delta \log|\delta z| = \Delta \log \|Jv\| + \Delta \log(\|w\| \, |\cos\theta_v|)$ to first order in ϵ. (ii) The read prefers the true digit to the counterfactual iff the unperturbed margin is positive; producing the true digit requires the stronger condition $z_{\text{true}} > \max_{j \ne \text{true}} z_j$ over the vocabulary. Changing b changes this preference without changing any first-order quantity, so it is not a function of transport and coupling. (iii) If the value read at q is $\sum_{j=0}^{m} \pi_j V_j$ with attention weights $\pi_j \propto e^{s_j}$ over the correct state ($j = 0$)*

and m competitors, then $\pi_0 = (1 + \sum_{j\geq 1} e^{-\Delta_j})^{-1}$ with margins $\Delta_j = s_0 - s_j$, independent of the transported values V_j. So $\pi_0 \geq 1/2$ iff $\sum_j e^{-\Delta_j} \leq 1$, i.e. $\Delta \geq \ln m$ at equal margins, and the correct content dominates this read's contribution to z if $\pi_0\langle w, V_0\rangle > \sum_{j\geq 1} \pi_j |\langle w, V_j\rangle|$.

Proof. (i) Taylor expansion of z in e_{x_0} gives $\delta z = \epsilon\langle w, Jv\rangle + O(\epsilon^2)$, and $\langle w, Jv\rangle = \|w\|\|Jv\| \cos\theta_v$; taking logarithms of the ratio before and after training gives the split. (ii) b enters z additively and no derivative. (iii) Softmax algebra: $\pi_0 = e^{s_0}/\sum_j e^{s_j} = (1 + \sum_{j\geq 1} e^{s_j - s_0})^{-1}$; the dominance condition bounds the competitors' contribution in absolute value.

B Persistence and restating in a controlled model

Persistence–refresh model. A state from a large alphabet is written at step 0 and read at every step $t = 1, \ldots, T$. A read at age a (steps since the last write) succeeds with probability $q(a) = e^{-H(\gamma a)}$, where H is non-decreasing and C^1. A restatement resets the age to 0, and failed reads are unrecoverable, so a trajectory with read ages (a_t) succeeds with probability $S = \exp(-\sum_t H(\gamma a_t))$. For the Gaussian readout $H(u) = -\ln \Phi(e^{-u}/\sigma)$ and $\ell^* = \ln(1/\sigma)/\gamma$. When differentiating with respect to the persistence parameter γ, we hold the restating policy fixed. In the controlled Transformer, behaviour and representation share parameters, so the results below describe only the direct effect of γ.

Theorem 3 (Exposure)**.** *Let D_E be the distribution of read ages in expert traces and $\tilde{D}_\pi$ the success-weighted distribution of read ages in the model's own rollouts. Then $\partial_\gamma L_{\mathrm{CE}} = T\,\mathbb{E}_{D_E}[aH'(\gamma a)]$ and $-\partial_\gamma \log J = T\,\mathbb{E}_{\tilde{D}_\pi}[aH'(\gamma a)]$. (a) If expert traces never leave the state unrestated for more than $w \leq \ell^*$ steps, $\partial_\gamma L_{\mathrm{CE}} \leq TwH'(\gamma w) = \exp(-\Theta(e^{-2\gamma w}/\sigma^2))$ for the Gaussian readout. (b) $-\partial_\gamma \log J \geq 0.0135\, T\ell^*\, \tilde{D}_\pi([\ell^* - 1/\gamma, \ell^*])$ when $\ell^* \geq 2/\gamma$.*

Proof. $L_{\mathrm{CE}} = \mathbb{E}_E[\sum_t H(\gamma a_t)]$ plus terms free of γ, which gives the first formula. For RL the restating decisions do not depend on γ, so $\partial_\gamma J = -\mathbb{E}_\pi[S \sum_t a_t H'(\gamma a_t)]$, which gives the second. (a) Every expert read has $a \leq w$. For the Gaussian readout $H'(u) = x\varphi(x)/\Phi(x)$ with $x = e^{-u}/\sigma$, decreasing in x on $x \geq 1$ (i.e. increasing in u for $u \leq \ln(1/\sigma)$), and $H'(\gamma w) \leq 2x_w\varphi(x_w)$ with $x_w = e^{-\gamma w}/\sigma$. (b) For $a \in [\ell^* - 1/\gamma, \ell^*]$, $x \in [1, e]$ and $H'(\gamma a) \geq e\varphi(e)/\Phi(e) \approx 0.0271 > 0.027$, since $x\varphi(x)/\Phi(x)$ decreases on $x \geq 1$, while $a \geq \ell^*/2$. □

Corollary 2 (Exposure without reward)**.** *Supervised training on traces whose unrestated gaps reach ℓ^* receives pressure on γ of the same order as RL (apply (b) to D_E).*

Corollary 3 (Learnability window)**.** *With a single read at horizon T and no restating, $|\partial_\gamma J| = Tx\varphi(x)$ with $x = e^{-\gamma T}/\sigma$. Its log-derivative in T is $1/T - \gamma(1 - x^2)$, which is strictly decreasing, so $|\partial_\gamma J|$ is unimodal with maximiser $T^* = \ell^* + O(1/(\gamma \ln(1/\sigma)))$.*

Theorem 4 (When to restate)**.** *With reads at every step and cost c per restatement, restating every Δ steps costs $C(\Delta) = [\int_0^\Delta H(\gamma a)\, da + c]/\Delta$ per step. If $c\gamma < m(\infty)$, where $m(u) = uH(u) - \int_0^u H = \int_0^u vH'(v)\, dv$, the minimiser is unique and satisfies $m(\gamma\Delta^*) = c\gamma$. The optimal rate γ/u^* increases with γ when H' is non-decreasing on $[0, u^*]$. At equal mean rate, periodic restating beats memoryless restating. If $c\gamma \geq m(\infty)$, never restating is optimal.*

Proof. With $u = \gamma\Delta$, $dC/d\Delta = \Delta^{-2}[\gamma^{-1}m(u) - c]$ and $m'(u) = uH'(u) \geq 0$, which gives existence and uniqueness. Using $u^{*\prime}(c\gamma) = 1/(u^*H'(u^*))$, $\frac{d}{d\gamma}\frac{\gamma}{u^*} = \frac{1}{u^*}[1 - \frac{m(u^*)}{u^{*2}H'(u^*)}] \geq \frac{1}{2u^*}$ when $m(u^*) \leq u^{*2}H'(u^*)/2$. For random cycle lengths, the renewal-reward cost is $\mathbb{E}F(\Delta)/\mathbb{E}\Delta \geq F(\mathbb{E}\Delta)/\mathbb{E}\Delta$ by Jensen, since $F(\Delta) = \int_0^\Delta H(\gamma a)\, da + c$ is convex. □

Numerical checks ($\gamma = 0.1$, $w = 5$, $T = 60$). With an RL policy that restates at the same mean rate as the expert, the ratio of RL to supervised pressure on γ is 1.2, 2.3, 14, 277 and 2.7×10^6 for $\sigma = 0.5, 0.3, 0.2, 0.15, 0.1$. At $\sigma = 0.15$ supervised pressure grows from 0.03 to 5.4 to 105 as the expert's gap grows from 5 to 10 to 20 (RL: 8.7). $|\partial_\gamma J|$ peaks at $T = 19$ against $\ell^* = 19.0$. $\gamma\Delta^* \in [0.66, 1.00]$ while Δ^* ranges over 4–43. A version of the model that reads only at the end is ill-posed (never restating always wins), which is why reads occur at every step.

C Selectivity under outcome reward

If a read at age a succeeds with probability $\Phi\big(e^{-\gamma_S a}/\sqrt{\sigma^2+\kappa^2 e^{-2\gamma_N a}}\big)$, where a nuisance mode with decay rate γ_N interferes with the state, then $\partial_{\gamma_S} J < 0$ and $\partial_{\gamma_N} J > 0$ whenever $\kappa > 0$. A supervised objective that also predicts nuisance tokens adds a term favouring nuisance retention, and its gradient on γ_N changes sign at some weight f^* (numerically $+1.09$ at $f = 0$, -0.01 at 0.25, -1.10 at 0.5). This prediction assumes nuisance content only interferes. The measured change is direction-agnostic at a token (Section 5.4), and state-versus-nuisance comparisons are confounded by token type, so the prediction is not tested cleanly.

D Probe details and validation

Raw-text probe: "`Keep track of every variable.`", 2–40 random assignments `name = dddd.`, the target assignment, "`Note: the sky today is <colour>.`", distractor assignments totalling $\approx 1{,}536$ tokens, then "`Question: what is the value of <name>? Answer: <name> = dddd.`". **Reasoning-style (own-format) probe:** the model's chat template with a fixed user message, the assistant turn opened (with `<think>` when the vocabulary has it), 10–40 arithmetic lines, "`Step 1 gives the key value: dddd.`", "`I'll mark that line in <colour>.`", arithmetic lines ($\approx 1{,}536$ tokens), then "`Now, going back to step 1, the key value was dddd.`". Counterfactual directions are normalised embedding differences to another single-token digit (S) or colour (N). G uses the final-layer hidden state after the final normalisation. Qwen tokenizes numbers digit by digit; OLMo groups digits into chunks of up to three, so its state token is the first three-digit chunk. Exponent: unweighted least squares of $\log G$ on $\log r$ over integer $r \in [16, R-8)$, where R is the curve length (1,506–1,535). Floor fit: $G = g + b(r+1)^{-p}$ on the same range (Table 7). SNPR $= \sum_{r\geq 1} G_S^2 / \sum_{r\geq 1} G_N^2$. Stationarity: median over 12 log-spaced bins of $|\log(G_{\text{late}}/G_{\text{early}})|$, splitting prompts at the median source position. Position control: the late-source half of the prompts. Bootstrap: 2,000 paired resamples of prompts.

Table 4: Framework checks. Linearity: response ratio for 2ϵ vs. ϵ. Power law wins: prompts for which a power law fits the log-binned single-prompt curve better than one exponential. Oscillation period: dominant period of $G/\text{trend} - 1$ in $[4, 40]$ tokens.

Model	Probe	linearity	stationarity	power law wins (prompts)	pow./exp. error	osc. period	tokens/line
Qwen R1-Distill	own	2.06	0.05	120/120	0.12 / 0.34	17.2	17.3
Qwen DeepScaleR	own	2.06	0.05	120/120	0.13 / 0.33	17.2	17.3
OLMo DPO	own	2.01	0.11	120/120	0.22 / 0.43	11.0	10.9
OLMo RLVR	own	2.01	0.14	120/120	0.23 / 0.44	11.0	10.9
Qwen R1-Distill	raw	2.05	0.39	119/120	0.05 / 0.26	8.1	8.2
Qwen DeepScaleR	raw	2.05	0.38	120/120	0.04 / 0.27	8.1	8.2
OLMo DPO	raw	2.00	0.13	93/120	0.05 / 0.15	6.1	6.2
OLMo RLVR	raw	2.00	0.12	95/120	0.04 / 0.16	23.1	6.2

Table 5: Band-resolved change in the decay exponent of the state for every RL step (*: 95% CI excludes zero). A distance band $r \approx 1/k$ is a frequency band; negative values mean the effective decay slowed.

RL step (probe)	$r \in [16, 64)$	$r \in [64, 256)$	$r \in [256, 1400)$
Qwen: R1-Distill→DeepScaleR (prompt set 1) (own)	-0.037 $[-0.056, -0.021]^*$	-0.035 $[-0.061, -0.013]^*$	-0.012 $[-0.020, -0.005]^*$
Qwen: R1-Distill→DeepScaleR (prompt set 2) (own)	-0.040 $[-0.058, -0.022]^*$	-0.005 $[-0.022, +0.011]$	-0.009 $[-0.016, -0.002]^*$
OLMo: DPO→RLVR (own)	$+0.023$ $[+0.013, +0.033]^*$	$+0.017$ $[+0.008, +0.025]^*$	-0.043 $[-0.048, -0.037]^*$
Qwen: R1-Distill→DeepScaleR (raw)	$+0.008$ $[+0.000, +0.016]^*$	-0.002 $[-0.009, +0.005]$	$+0.002$ $[-0.003, +0.006]$
OLMo: DPO→RLVR (raw)	$+0.006$ $[+0.001, +0.012]^*$	$+0.004$ $[+0.001, +0.008]^*$	-0.003 $[-0.006, +0.000]$

E Cross-entropy exposure experiment

Start model: R1-Distill-Qwen-1.5B, full fine-tuning in fp32 with bf16 autocast, AdamW (learning rate 10^{-5}, 10 warm-up steps), gradient clipping at 1, 250 steps of 8 sequences, gradient checkpointing. Traces use the model's chat template. The user asks for a multi-step calculation, and the

Table 6: Per-layer coefficients of the response, Qwen pair (reasoning-style probe, 24 prompts): effective decay exponent p_ℓ, reaction ρ_ℓ (log ratio of the response averaged over all $r \geq 1$ at layer ℓ to that at layer $\ell - 1$; only selected layers are shown, so the listed values do not sum to the total change; layer 28 is taken after the final normalisation), and amplitude c_ℓ (power-law prefactor at $r = 1$).

Layer	1	4	8	12	16	20	24	28
p_ℓ (R1-Distill)	1.44	0.85	0.73	0.53	0.45	0.38	0.37	0.35
p_ℓ (DeepScaleR)	1.45	0.85	0.72	0.52	0.43	0.36	0.36	0.33
ρ_ℓ (R1-Distill)	–	+0.05	+0.05	-0.01	+0.04	-0.00	-0.02	-0.12
ρ_ℓ (DeepScaleR)	–	+0.06	+0.07	-0.01	+0.04	-0.01	-0.00	-0.03
c_ℓ (R1-Distill)	7.1e-03	5.1e-03	4.6e-03	2.5e-03	2.1e-03	1.3e-03	1.0e-03	7.2e-04
c_ℓ (DeepScaleR)	7.5e-03	5.5e-03	4.7e-03	2.5e-03	2.0e-03	1.3e-03	1.1e-03	7.9e-04

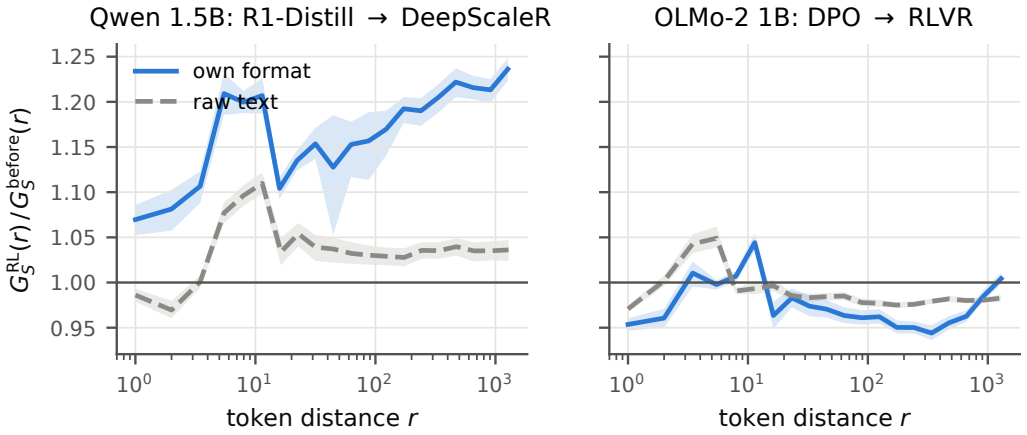

Figure 3: State response after vs. before RL, $G_S^{\mathrm{RL}}(r)/G_S^{\mathrm{before}}(r)$, log-binned, with 95% bootstrap bands. Solid: reasoning-style (own-format) probe; dashed: raw text.

assistant's reasoning writes "Let me record the quantity Q = dddd.", then 60–100 arithmetic lines (`Adding a and b gives s.` and similar). After every fourth line it adds either a restatement ("(Q is still dddd.)", short gaps) or a neutral line of similar length ("(Continuing the arithmetic.)", long gaps). It ends with "So the stored quantity Q was dddd." and a final answer. Loss is on all assistant tokens. Both arms use the same trace skeletons, and none of the templates occur in the probe. Probes at steps 25, 100 and 250 use the reasoning-style prompts of Section 3, and a raw-text probe runs at step 250. Recall loss (on the recalled digits) fell below 10^{-3} by step 25 in both arms. Clean recall on the probe stayed at 100% up to 2,048 tokens.

F Layer, head and ablation analyses (Qwen pair)

We re-ran the reasoning-style probe on the first 24 prompts of prompt set 1 for R1-Distill and DeepScaleR with eager attention in fp32, recording the response at every layer and the attention from every position back to the state token. The final-layer exponent reproduces the main run exactly on these prompts ($p_S = 0.3521$ and 0.3358; RL change -0.0163), with curves within 1%. **Layers:** a power law beats one exponential at every layer (log-error 0.11–0.21 vs. 0.35–0.93), and a tempered power law selects a zero cutoff at every layer after the first ($\lambda = 1.6 \times 10^{-4}$ at layer 1). **Heads:** a head reads the source "above uniform" if its mean attention to the source exceeds that of a uniform spread over the context. Far range ($r \geq 256$): 13/336 heads above uniform, none above $2.6\times$. Near range ($r < 16$): 285/336. The long-range share of attention to the source changes with $|z| > 3$ in 264 heads (209 up); far-range attention itself rises in 231 heads. **Ablation (first run):** we zero the outputs of the 12 heads with the most far-range attention to the source in R1-Distill, namely (layer, head) (16,1), (17,0), (6,3), (15,2), (17,3), (16,6), (15,3), (8,0), (7,9), (14,9), (26,6), (26,11), or 12 random heads, using the same 10 prompts for all conditions. The sig-

Table 7: Exponent change with a constant floor, $G = g + b(r+1)^{-p}$ (Brändel et al., 2026), fitted to the prompt-averaged curves on the same range as the pure power law, with paired bootstrap 95% CIs (*: excludes 0), and fit-free response ratios. Floor share: g/G at $r = 1000$ before the step; floor ratio omitted when the floor is below 1%. Last two columns: mean response after/before the step at $r \in [960, 1040)$ and at the answer tokens.

Step	Probe	Δp (pure)	Δp with floor [95% CI]	p with floor (before)	floor share at r=1000	floor g after/before	G after/before, $r \in [960, 1040)$	G after/before, answer
Qwen: R1-Distill→DeepScaleR (set 1)	own	−0.021	−0.024 [−0.048, −0.001]*	0.73	73%	×1.23	×1.21	×1.24
Qwen: R1-Distill→DeepScaleR (set 2)	own	−0.014	−0.033 [−0.049, −0.020]*	0.73	75%	×1.19	×1.20	×1.23
OLMo: DPO→RLVR	own	−0.018	+0.082 [+0.060, +0.104]*	1.28	96%	×0.99	×0.99	×0.98
Qwen: R1-Distill→DeepScaleR	raw	−0.001	+0.007 [+0.000, +0.016]*	0.47	35%	×1.09	×1.04	×1.01
OLMo: DPO→RLVR	raw	−0.001	+0.001 [−0.001, +0.002]	0.29	0%	–	×0.98	×0.99
OLMo: SFT→DPO	own	−0.037	+0.364 [+0.297, +0.462]*	0.92	84%	×2.06	×1.73	×1.37
OLMo: SFT→DPO	raw	−0.038	−0.042 [−0.046, −0.039]*	0.33	0%	–	×1.61	×1.31
Qwen: base→R1-Distill (distill.)	raw	−0.045	+0.088 [+0.043, +0.124]*	0.39	0%	–	×1.20	×0.92

Table 8: Cross-entropy fine-tuning versus RL from the same model (R1-Distill), reasoning-style probe, relative to R1-Distill. *: 95% CI excludes zero.

Model	$\|\Delta\theta\|$	Δp_S [95% CI]	Δp_N	Δp_R
RL (DeepScaleR)	1.83	-0.021 $[-0.026, -0.016]^*$	-0.049	-0.020
SFT long-gap, step 25	3.11	$+0.065$ $[+0.056, +0.076]^*$	$+0.019$	$+0.043$
SFT long-gap, step 100	4.41	$+0.076$ $[+0.063, +0.089]^*$	$+0.033$	$+0.043$
SFT long-gap, step 250	5.56	$+0.091$ $[+0.078, +0.104]^*$	$+0.018$	$+0.047$
SFT short-gap, step 25	3.13	$+0.069$ $[+0.058, +0.080]^*$	$+0.001$	$+0.049$
SFT short-gap, step 100	4.46	$+0.078$ $[+0.065, +0.092]^*$	$+0.010$	$+0.046$
SFT short-gap, step 250	5.63	$+0.116$ $[+0.104, +0.129]^*$	$+0.032$	$+0.067$

nal retained at $r \in [512, 1024)$ after the targeted ablation is $0.85\times$. **Ablation with behavioural readouts (second run):** same 12 targeted heads; control sets drawn with seeds 0–2 over all 336 heads (random) or as one random other head in the layer of each targeted head (layer-matched, seeds 100–101); 40 prompts per probe with the full probe (`ablate_behaviour.py`, sdpa attention, fp32). The two control sets that abolish raw-text recall are random set 3 and layer-matched set 2 (Table 9); they share no head. **Along the update (Figure 2a).** When we interpolate the weights, $\theta(\alpha) = \theta_{\text{RID}} + \alpha(\theta_{\text{DSR}} - \theta_{\text{RID}})$ with $\|\Delta\theta\| = 1.83$, every exponent falls monotonically with α (Figure 2a; $p_S = 0.355, 0.342, 0.334, 0.333$ at $\alpha = 0, 0.5, 1, 1.5$). The initial slope, $dp_S/d\alpha \approx -0.026$, is the directional derivative along the total RL displacement. It is not the sum of the per-step first-order terms of Theorem 1, but it shows that the change builds up smoothly along the update. **Controls** (`colab_controls.py`, Table 10). Null baseline: R1-Distill with each weight tensor perturbed along a random Gaussian direction whose norm equals that tensor's RL update (DeepScaleR minus R1-Distill, computed from the stored bf16 weights, total norm 1.64; 1.83 at full precision), three seeds, 60 reasoning-style prompts; DeepScaleR is measured on the same prompts from the same bf16 weights. Perplexity: mean next-token loss over all prompt tokens under each ablation condition, on the 40 prompts per probe of the ablation run. **Band exponents:** for OLMo the final-layer exponent depends on how distance bins are weighted, because the exponent varies with distance. Band exponents (Table 5) are the reliable description there. The local exponent in finer bands ($r \in [8, 32), [32, 128), [128, 512), [512, 1400)$) is $0.609, 0.426, 0.246, 0.504$ for OLMo DPO and $0.649, 0.433, 0.247, 0.451$ for OLMo RLVR (reasoning-style probe; on $[64, 256)$, $0.386 \to 0.402$, the $+0.017$ of Table 5), against $-0.06, 0.51, 0.30, 0.32$ for R1-Distill and $-0.02, 0.48, 0.27, 0.31$ for DeepScaleR. OLMo steepens at long range, while Qwen does not.

G Controlled transformer

Vocabulary: 32 state values, 32 restatement tokens (one per value), and markers. Each step is six tokens `NEXT D FIL u Q c`: D is the model's decision (`GO` or restate-with-value v), u a random filler value, and `Q c` a membership read ("is c the state?", c equal to the state with probability $1/2$), answered at c and masked from all later positions. Model: 4 layers, width 128, 4 heads, pre-LayerNorm, a learnable ALiBi slope per head (softplus, initialised at 0.3 per token), no other positional encoding. Start model: 2,000 supervised steps (batch 64, AdamW 10^{-3}) on expert traces that restate with probability 0.9 once the age reaches 4 and 0.1 otherwise. Continuations: 300 supervised steps at batch 128 (lr 3×10^{-4}), or 300 GRPO updates with 16 environments $\times$ 8 samples (lr 10^{-4}), reward $= \mathbb{1}$[all reads correct] $- c \cdot$ #restatements, no KL. Learning rates differ between the

Table 9: Head ablation in R1-Distill with behavioural readouts (40 prompts per probe). First row per probe: no ablation (p_S, S/R, recall, answer log-probability). Other rows: change in p_S and in the answer log-probability, ratio of $|\Delta\text{ans}_S|$, and recall lost/gained with an exact McNemar p when below 0.05 (paired bootstrap 95% CIs; *: CI excludes 0, or 1 for the ratio).

Probe	Ablated heads	p_S / Δp_S [95% CI]	$\lvert\Delta\text{ans}_S\rvert$ ratio	recall lost/gained	S/R	recall	Δ answer log-prob
own	none	0.352	–	–	15.1	1.00	−0.09
own	top-12 far-reading	+0.077 [+0.063, +0.092]*	0.85 [0.78, 0.93]*	−0/+0	14.6	1.00	−0.07 [−0.09, −0.04]*
own	random 1	+0.018 [−0.007, +0.043]	0.83 [0.71, 0.94]*	−0/+0	12.9	1.00	−0.04 [−0.06, −0.02]*
own	random 2	−0.022 [−0.034, −0.009]*	0.87 [0.81, 0.92]*	−0/+0	13.8	1.00	−0.00 [−0.02, +0.01]
own	random 3	−0.019 [−0.045, +0.006]	1.19 [1.12, 1.27]*	−0/+0	15.0	1.00	+0.01 [+0.00, +0.02]*
own	layer-matched 1	+0.034 [+0.024, +0.043]*	0.73 [0.68, 0.79]*	−0/+0	15.1	1.00	−0.17 [−0.22, −0.13]*
own	layer-matched 2	+0.010 [+0.001, +0.020]*	0.60 [0.54, 0.66]*	−1/+0	14.5	0.97	−0.59 [−0.67, −0.53]*
raw	none	0.321	–	–	6.5	0.35	−4.13
raw	top-12 far-reading	+0.026 [+0.005, +0.050]*	1.10 [0.86, 1.46]	−4/+3	6.4	0.33	−0.69 [−1.00, −0.35]*
raw	random 1	+0.002 [−0.024, +0.029]	1.31 [0.98, 1.89]	−8/+8	5.2	0.35	+0.27 [−0.37, +0.91]
raw	random 2	+0.028 [+0.009, +0.049]*	1.06 [0.89, 1.28]	−3/+3	7.6	0.35	+0.24 [−0.02, +0.50]
raw	random 3	+0.029 [+0.006, +0.055]*	0.27 [0.21, 0.35]*	−14/+0 ($p = 1 \times 10^{-4}$)	9.8	0.00	−4.03 [−4.50, −3.54]*
raw	layer-matched 1	+0.014 [−0.004, +0.032]	0.70 [0.47, 0.98]*	−5/+0	5.7	0.23	−0.41 [−0.75, −0.06]*
raw	layer-matched 2	+0.047 [+0.031, +0.063]*	0.18 [0.12, 0.25]*	−14/+0 ($p = 1 \times 10^{-4}$)	6.1	0.00	−3.59 [−4.06, −3.10]*

Table 10: Controls. (a) Change in p_S from R1-Distill for random updates with the RL update's per-tensor norms and for the RL update itself, on the same prompts (paired bootstrap 95% CIs). (b) Prompt perplexity under each head ablation, with its ratio to no ablation.

(a) Null baseline, R1-Distill, reasoning-style probe (60 prompts)

Update	Δp_S [95% CI]
random, per-tensor norms of the RL update (seed 1)	-0.0002 [-0.0004, $+0.0001$]
random, per-tensor norms of the RL update (seed 2)	-0.0002 [-0.0004, -0.0001]
random, per-tensor norms of the RL update (seed 3)	-0.0002 [-0.0004, -0.0000]
RL (DeepScaleR)	-0.0149 [-0.0184, -0.0111]

(b) Prompt perplexity under head ablation (40 prompts per probe)

Ablated heads	perplexity, own / raw (ratio to none)
none	3.13 (×1.00) / 5.91 (×1.00)
top-12 far-reading	3.17 (×1.01) / 5.92 (×1.00)
random 1	3.73 (×1.19) / 6.94 (×1.17)
random 2	3.21 (×1.03) / 5.92 (×1.00)
random 3	3.24 (×1.04) / 5.73 (×0.97)
layer-matched 1	3.20 (×1.02) / 5.86 (×0.99)
layer-matched 2	3.23 (×1.03) / 5.71 (×0.97)

supervised and RL continuations, so we do not compare their magnitudes. ℓ^*: first age at which the 3-step-smoothed read accuracy with restating disabled (64 steps, 512 sequences) falls below 0.75.

Table 11: Controlled transformer, both seeds. Restate interval: episode length (31 steps after the write) divided by the mean number of restatements per episode; it overstates the cycle length when the last cycle is unfinished.

Run	ℓ^* (seed 0 / 1)	restate interval
SFT, short gaps (start)	10 / 9	3.6
+ SFT, short gaps (control)	10 / 10	3.6
+ SFT, long gaps (no reward)	46 / 37	never
+ RL, restating free ($c = 0$)	10 / 9	4.3
+ RL, $c = 0.01$	16 / 15	15.5
+ RL, $c = 0.03$	16 / 15	15.6
+ RL, $c = 0.1$	15 / 14	15.5
+ RL, no restating, $T = 0.5\ell_0^*$	9 / 8	–
+ RL, no restating, $T = \ell_0^*$	12 / 10	–
+ RL, no restating, $T = 2\ell_0^*$	11 / 11	–
+ RL, no restating, $T = 4\ell_0^*$	11 / 9	–

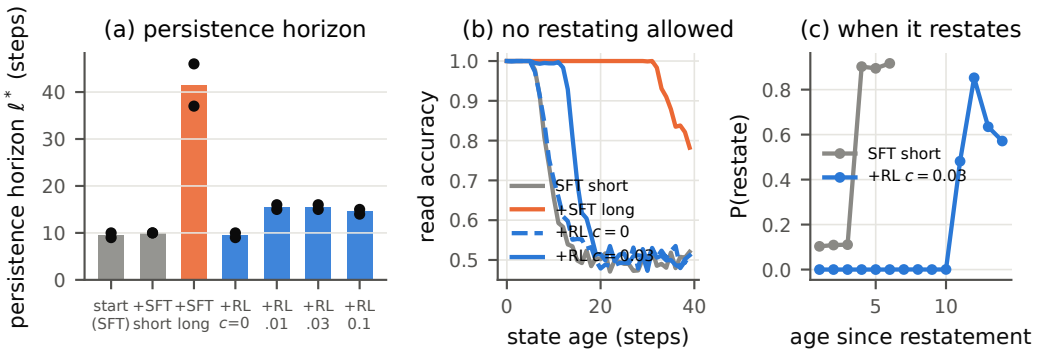

Figure 4: Controlled transformer (two seeds; dots are seeds). (a) Persistence horizon after each continuation. (b) Read accuracy vs. age with restating disabled. (c) Per-step probability of restating at each age since the last restatement (a hazard, averaged over seeds). With cost, the hazard implies cycles of about 12 steps; the interval in Table 11 is larger because it divides the episode length by the number of restatements, which counts the unfinished last cycle.

Learnability window (weak). RL without restating at training horizons $0.5/1/2/4 \times \ell^*$ changes ℓ^* by $-1/+1.5/+1.5/+0.5$ on average (Figure 5): short horizons do not help, but the peak is shallow.

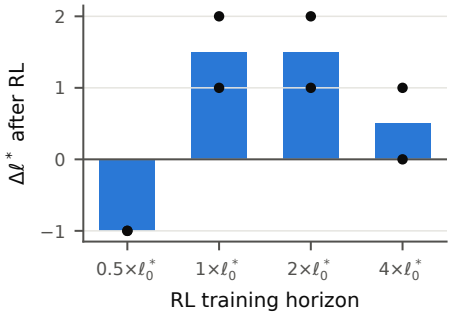

Figure 5: Change in ℓ^* after RL without restating, by training horizon (dots: seeds).

H Tree traversal at 1.5B scale

We ran inorder traversal of dense binary trees (depth 3–7, 40 trees per depth) on the Qwen chain. With greedy decoding, the R1-style models hit the 12,000-token limit on 50–100% (R1-Distill) and 98–100% (DeepScaleR) of chain-of-thought runs. With the chat template and sampling ($T = 0.6$, top-p 0.95) this drops to 20–50% and 5–32%, but exact-match accuracy stays at 0–5%, and 75–84% of first errors occur on the first node. The task is beyond these models and cannot test persistence at this scale. The rate of runaway generations depends on decoding: greedy decoding and sampling reverse the ordering of the two models.