ROUTING THE PAST: HOW RETRIEVAL CIRCUITS CONTROL APPARENT MEMORY CAPACITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Ask a language model for the current value of a variable that has been reassigned 32 times and it answers at chance. This is usually read as working memory running out. We show that much of this apparent capacity is set by which query-time circuit retrieves the answer, not by what the model stores. The evidence is causal. An answer-agnostic scalar gain on a few copy heads, applied only at the query token so that it cannot alter the context, flips genuinely binding-limited base models from failing to passing (Llama-3.1-8B $0.03 \to 0.60$, Pythia-6.9B $0.12 \to 0.71$; positive in 20 of 25 models at one fixed setting; budget-matched random heads reach only 0.18 against 0.60), and ablating the same heads runs the switch in reverse. The lift survives free generation and interleaved distractors that remove the "last line" cue. The value was in the residual stream all along: under adversarial controls, a linear probe reads it above the model's own output in 10 of 12 models across seven families. The mechanism transfers to code, dialogue, and documents (code retrieval $0.08 \to 0.78$ with no code-specific labels), and it reframes evaluation: under a raw-completion harness the capacity ranking of 19 chat models dissolves and the scale trend vanishes. Across 47 models (70M–32B) and three benchmarks (PI-LLM, bAbI, MultiWOZ), we characterize the occurrence-order code that crowds with load, prove that position-only routes can reach only endpoints, and identify a query-conditioned pointer-head signature behind the rare interior access of gemma-4-31B. Apparent working-memory capacity is, in large part, retrieval routing.

1 INTRODUCTION

A coding agent reassigns a variable, then writes code using the value from three edits ago. A retrieval pipeline quotes a refund limit that a later amendment replaced. A support bot ships to an address the customer changed twice. In each case the model returns a stale value of an entity that was updated in plain view, fluently and without hedging, and the failure is filed under "the model forgot." We show instead that a different query-time circuit can recover the answer from the unchanged context at exactly the loads where the default circuit fails. Rebind one key to u successive values and ask a language model for the latest one. Accuracy collapses toward chance as u grows (Gong et al., 2024; Wang & Sun, 2025; xio), a limit that resembles the four-chunk bound of human working memory (Cowan, 2001; Luck & Vogel, 1997). The natural reading is that the model has run out of storage. We show that the failure is set instead by how the model retrieves. The context is not blank at failure: a linear probe reads the current value off the residual stream above the model's own output even under adversarial controls (§4.1), and every model whose geometry we measure (30 models, 70M–32B) carries a code that indexes the order of mentions. What varies is the circuit that reads this code. An order-based read can address an earlier binding but crowds as the spacing of its low-dimensional manifold shrinks with load. A copy of the most recent mention never crowds but can address nothing else. A benchmark score is therefore a diagnostic of which route a model uses, not a measure of what it stores. The limit is causally switchable in both directions: amplifying a binding-limited model's latent copy heads with an answer-agnostic gain at the query token lifts high-load retrieval to near ceiling under null controls, and ablating a recency-copy model's copy heads exposes the latent limit (Fig. 1). This reinterprets the working-memory benchmark: its interference limit reflects which retrieval circuit a model engages as much as how much it stores. The copy route escapes retrieve-latest but cannot random-access a non-final binding, so most capable models trade away non-final addressing (gemma-4-31B is the one exception).

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.

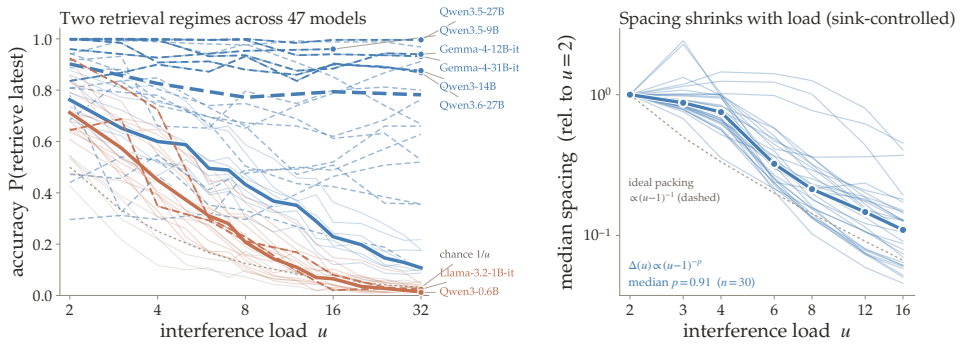


Figure 1: Apparent retrieve-latest capacity is largely controlled by retrieval routing. Left: retrieve-latest accuracy against interference load for 47 models (70M–32B), which split into two regimes that read one shared context in two ways: binding-limited (decays to chance) and recency-copy (escapes the limit). Dashed curves are chat models under fair chat elicitation. Right: in every one of the 30 sink-controlled models the order-code spacing shrinks with load (crowding exponent median 0.91; §4.2). The load-bearing causal result is the switch itself: an answer-agnostic gain on latent copy heads at the query token moves a binding-limited model into recency-copy (u=32: 0.06 → 0.85; §4.7).

Our contributions:

1. **Bidirectional causal control of the capacity regime (§4.7).** An answer-agnostic gain on latent copy heads, applied at the query token only, converts binding-limited base models to recency-copy under their fair interface (Llama-3.1-8B 0.03 → 0.60, Pythia-6.9B 0.12 → 0.71, up to 0.06 → 0.85), against random-head, budget-matched, wrong-position, and norm-matched nulls; ablating a recency-copy model's copy heads reverses it. The direction holds in 24 of 25 models.
2. **The information is present but under-routed (§4.1).** A deconfounded linear probe decodes the current value above both chance and the model's own output in 10 of 12 models across seven families, with the largest margin where the model routes it least.
3. **A conserved occurrence-order code and an endpoint bound (§3, §4.2).** Every sink-controlled model writes competing bindings onto a low-dimensional ordinal manifold whose spacing crowds with load within a fixed extent. We prove that a position-only route can strictly prefer only the first or the latest binding (Prop. 1); interior access must pay the crowding cost.
4. **Random access is rare, not bought by scale, and has an attention signature (§4.3, §4.6).** Under one ordinal template gemma-4-31B answers the middle slot at 0.86 where 13 copy-route models, including larger ones, collapse. It carries a query-conditioned pointer-head signature.
5. **Scale, interface, and real tasks (§4.5, §4.8).** Under fair elicitation, scale escapes the limit only in chat-era families and only through the chat interface; in raw completion the chat models' ranking dissolves. The limit reproduces on PI-LLM, bAbI, and real MultiWOZ dialogue, and copy heads frozen from the synthetic task repair code retrieval (0.08 → 0.78).

2 Related Work

Binding and order codes. Feng & Steinhardt (2024) extract additive binding-ID vectors that causally mediate retrieval; Feng et al. (2025) identify a binding subspace; Dai et al. (2024; 2026); Bassi & Tomar (2026) argue the code is largely ordinal on a near-1-D manifold, the substrate for successor heads (Gould et al., 2023). That axis is occurrence order, not meaning (Singh, 2026), which our controls confirm. Oh & Demberg (2026) isolate a rebinding circuit for a single overwrite, Campbell et al. (2025) tie capacity to interference probability, and Tang et al. (2026) show retrieval is a one-shot final-token aggregation, the operation our model formalizes. Entity tracking is the behavioral face of the same machinery (Kim & Schuster, 2023; Prakash et al., 2024; Wang & Sun, 2026). All of these probe two to three items; we add the load axis.

Retrieval routing. Retrieval is carried by induction heads (Olsson et al., 2022), copy-suppression heads (McDougall et al., 2023), and retrieval heads (Wu et al., 2025), and which route wins is

contested (Gur-Arieh et al., 2026; Ortu et al., 2024; Prakash et al., 2026), so far correlationally. Our interventions follow standard activation patching (Meng et al., 2022; Zhang & Nanda, 2024; Conmy et al., 2023); reweighting heads is itself established (Zhang et al., 2024). What is new is that one scalar gain moves a model between two capacity regimes in a controlled direction. The copy route is the induction circuit that forms in a pretraining phase change (Olsson et al., 2022; Cabannes et al., 2024); our era-by-interface result adds a post-pretraining degree of freedom, namely whether a model routes this task through it.

Working-memory limits and capacity theory. Gong et al. (2024) find a human-like n-back limit and Wang & Sun (2025) a log-linear proactive-interference decline; Liu et al. (2024); Hsieh et al. (2024) document position-dependent loss, which we control for. The closest prior work, xio, suppresses entangled representations, editing the store in one direction. We edit only the query token, leaving every stored representation untouched by construction, and move the regime in both directions. Capacity theory prices superposed storage (Frady et al., 2017; Adler & Shavit, 2026; Barnfield et al., 2026) and softmax dispersion (Veličković et al., 2025; Gong & Zhang, 2024), a rival we test and reject at the identified head (App. E.3); it assumes synthetic codes, whereas we measure the geometry. From cognitive science we import tensor-product binding (Smolensky, 1990), binding-specific limits (Oberauer, 2019), fixed-extent temporal crowding (Brown et al., 2007), and Shepard's law (Shepard, 1987) as measurable signatures.

3 Formal framework

Task. A context presents a stream of "key: value" updates. A tracked key is rebound u times to distinct single-token values $v_1, \ldots, v_u$; the model is queried and must emit the latest value v_u, the $u-1$ stale values being distractors. The interference load u is our capacity axis, and accuracy $A(u)$ is the constrained choice (argmax of the final-token logits over the u presented values; chance $1/u$). Slot $k \in \{0, \ldots, u-1\}$ denotes the k-th update. Default runs have no filler distractors (nd=0), so a copy of the last line can solve the task; every headline is therefore repeated with interleaved distractors (nd$\geq$1) and with cued recall of a non-final slot, which no copy shortcut can solve.

Storage. Let $\mathbf{h}_t^{(\ell)}(x) \in \mathbb{R}^d$ be the layer-ℓ residual at position t of prompt x, and $t_k(x)$ the position of the tracked key's k-th occurrence. Averaging over n prompts with resampled identities cancels content and leaves the order code for slot k (Feng & Steinhardt, 2024):

$$\boldsymbol{c}_k \;=\; \frac{1}{n}\sum_{x} \mathbf{h}^{(\ell)}_{t_k(x)}(x) \;\in\; \mathbb{R}^d, \qquad k = 0, \ldots, u-1. \tag{1}$$

The set $\{\boldsymbol{c}_0, \ldots, \boldsymbol{c}_{u-1}\}$ is the *order manifold*, with mean $\bar{\mathbf{c}}$, extent $E(u) = \max_{j,k} \|\boldsymbol{c}_j - \boldsymbol{c}_k\|_2$, and median nearest-slot spacing $\Delta(u) = \mathrm{median}_k \|\boldsymbol{c}_{k+1} - \boldsymbol{c}_k\|_2$. We use the median because for near-collinear codes with conserved extent the mean spacing is $E/(u-1)$ by construction. We summarize the shrinking spacing by the descriptive fit

$$\Delta(u) \;=\; a\,(u-1)^{-p}, \tag{2}$$

with measured median p=0.91 (range 0.26–1.31). We do not read p as a packing law; the load-bearing fact is that spacing shrinks while extent does not (§4.2).

Retrieval. Retrieval is a last-token associative lookup (Tang et al., 2026; Bricken & Pehlevan, 2021). With query code $\mathbf{q}$ we score each slot by the mean-removed cosine $s_u[k] = \cos(\mathbf{q}-\bar{\mathbf{c}},\ \boldsymbol{c}_k-\bar{\mathbf{c}})$ and model the report as a Luce choice (Cane, 1960) over these scores plus a recency prior,

$$P(\text{report} = k \mid u) \;=\; \mathrm{softmax}_k\big(\beta\, s_u[k] \;+\; \lambda\, \rho_k\big), \qquad \rho_k = k - (u-1), \tag{3}$$

with geometry gain β and recency prior λ, fit by maximum likelihood on training loads (App. C). Read as a retrieval head's attention over the u occurrences, the three observed regimes are three settings of this logit: binding-limited (order-dominated, finite capacity), recency-copy (recency-dominated, no limit in our range), and incapable.

Why random access is the expensive case. Write the logit for slot k in its most general position-plus-content form,

$$z_k \;=\; \underbrace{\beta\, s_u[k]}_{\text{content}} \;+\; \underbrace{\lambda\, r(u{-}1{-}k) \;+\; \mu\, \pi(k)}_{\text{position}}, \qquad \beta, \lambda, \mu \geq 0, \tag{4}$$

where r and π are a recency kernel and a primacy kernel, assumed discretely convex (non-decreasing first differences) but not necessarily monotone. This nests Eq. equation 3.

Proposition 1 (Endpoint access) *If the head is position-only (β=0) and r, π are discretely convex, then $\max_k z_k = \max\{z_0, z_{u-1}\}$: a position-only head can strictly prefer only the first or the latest binding. Strictly selecting an interior slot $0 < k < u-1$ therefore requires $\beta > 0$, and its winning margin is bounded by the adjacent-slot order contrast,*

$$z_k - \max_{j \neq k} z_j \;\leq\; \beta \max_{j \in \{k-1,\, k+1\}} \big(s_u[k] - s_u[j]\big). \tag{5}$$

The proof is in App. C. Under two regularity assumptions on the sink-controlled geometry (quasi-uniform packing and non-degenerate codes, whose measured ranges we report rather than assume), the right-hand side is $O(\Delta(u))$ and vanishes as the spacing crowds. So within a recency/primacy route, endpoint access is free and interior access pays a geometry cost that grows with load. The bound leaves one escape: a query-conditioned read with $\beta > 0$ whose match depends on the cue. gemma-4-31B, which reads the interior slot at 0.86, is exactly such a model (§4.6), and it still pays the predicted rate: its interior accuracy crowds $0.99 \to 0.67$ over u=4–32 while its endpoint stays at 1.00, and its pointer-free sibling gemma-4-12B collapses at the interior to 0.07–0.25 (App. Fig. 9).

Regimes. A curve is *incapable* if within 0.06 of chance at $u \leq 4$, *binding-limited* if it reaches chance by u=32, and *recency-copy* otherwise. Parameters are fit on train loads, the layer is selected on validation loads, and every number is reported on held-out test loads (App. A).

4 Experiments

Qwen3-1.7B is the primary model; §4.4 analyzes 47 models. Base models are measured by raw completion; chat, instruct, and reasoning models with their chat template, since their raw-completion scores are a harness artifact (§4.5; cf. Sclar et al., 2024). Single-model mechanistic sections use completion. The order manifold uses 300 content-randomized prompts per load, behavioral curves 500 (App. A).

4.1 The latest binding is decodable but under-routed

If the benchmark reads chance while the value is still present, a probe should find it. A linear probe on the final query-token residual does, but probes are easy to fool, so we run four adversarial controls (App. Fig. 7). The naive probe reads the value at 0.83 (model: 0.05 at u=32); disjoint-value training drops it to 0.46 (value memorization), removing the positional shortcut (train nd=0, test nd=2) to 0.165 against the model's 0.041 (chance 0.0625); a shuffled-label probe sits at chance. Fully deconfounded, the probe still reads the value above chance and the model's output in 10 of 12 models across seven families (Fig. 2), with the largest margin where the model routes the value least. The one model whose probe falls below its output, OLMo-2-7B, is also the one clean non-converter in §4.7, as expected if it has no latent copy route to amplify. Decodability establishes presence, not use; the causal evidence is the switch of §4.7.

4.2 The order-code geometry crowds with load

Retrieve-latest accuracy declines from 0.57 at u=2 toward chance at u=32; a recency-only account predicts a plateau and fails ($R^2 = -17.3$). We measure the order manifold with the position-0 activation sink controlled by a neutral header line, without which one massive-activation token dominates the geometry (App. D). Sink-free, in all 30 models the u order codes lie on a low-dimensional manifold (participation ratio 1.1–2.1) whose extent is nearly load-independent (median CV 0.11) while its median spacing shrinks with load in every model (exponent median 0.91). Conserved extent with more items packed in is the signature of a resource account, not a slot model (Ma et al., 2014). On a shared plane the codes form a fixed ruler whose ticks crowd into the retrieval-noise disk; accuracy collapses where the spacing-to-noise ratio crosses 1, at $u \approx 8$ (App. Fig. 8).

4.3 Random access: reaching a non-final binding

The copy route returns only the most recent occurrence, a position-blind $O(1)$ read; the order route can address any past binding at a crowding-limited cost. Two facts tie random access to the order code. First, the limit is specific to a repeated binding, not item count: content-addressable retrieval of unique keys stays flat to N=64 (App. Fig. 15). Second, under a single matched ordinal cue only gemma-4-31B recovers a middle slot (0.86), while all 13 copy-route models collapse there with endpoints surviving, as Prop. 1 predicts (Fig. 3). We use the middle rather than the first slot because the first is an endpoint reachable by a primacy read. For copy-route models the behavioral limit is the price of position-general access, a price the one genuine random-access model does not pay.

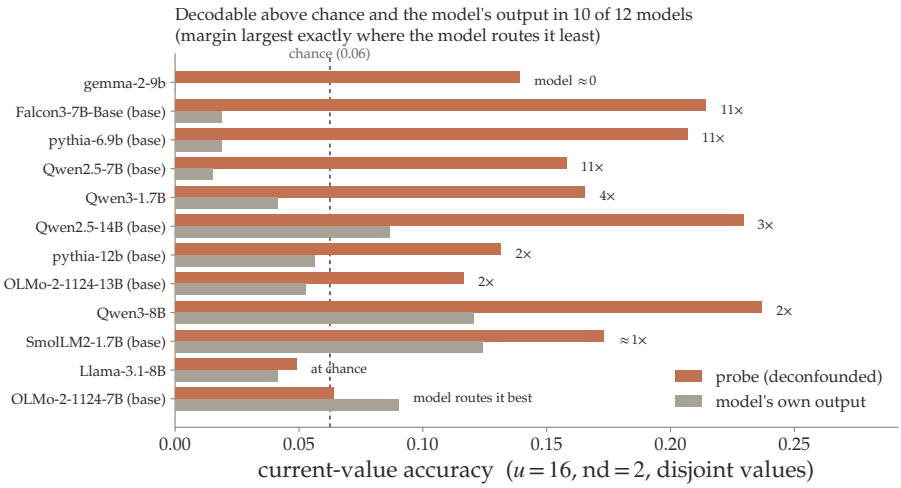


Figure 2: The current value is decodable above the model's own output, deconfounded and across models. The fully controlled probe (disjoint value vocabularies, positional shortcut removed; u=16, nd=2) reads the value above chance and the model's output in 10 of 12 models across seven families. The margin is largest where the model routes the value least, from gemma-2-9B (outputs it 0% of the time) to OLMo-2-7B, which already routes it best. Llama-3.1-8B is at chance for both.

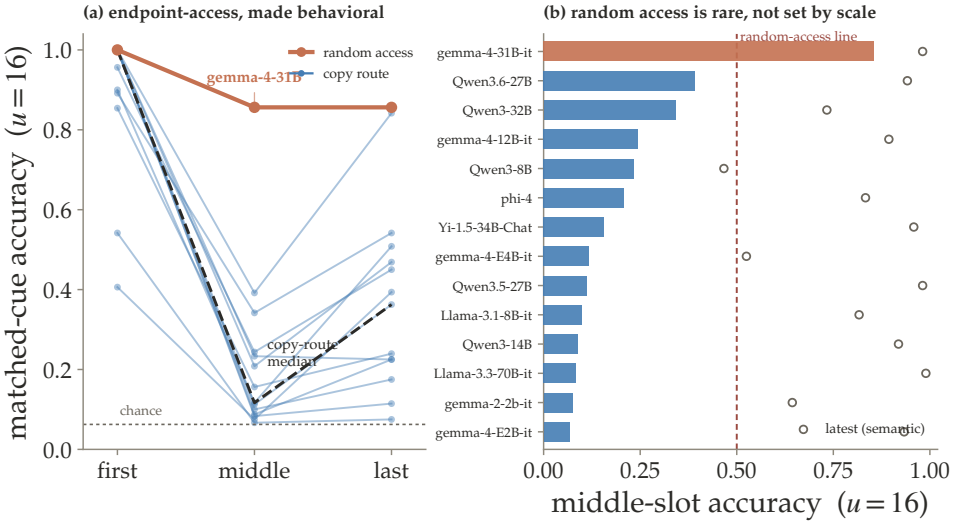


Figure 3: The endpoint bound made behavioral; random access is rare. (a) Matched-cue accuracy at the first, middle, and last slot under one ordinal template (u=16, chance ≈0.06). The 13 copy-route models (blue) keep both endpoints and collapse in the interior (dashed median); gemma-4-31B (clay) alone stays high at every slot. (b) The same middle-slot score across every model that clears retrieve-latest: only gemma-4-31B crosses the random-access line, and it is not the largest model tested.

4.4 A CROSS-MODEL REGIME MAP

Across 47 models (70M–32B; 11 families) crowding is ubiquitous: every sink-controlled model writes the low-dimensional order manifold whose spacing shrinks with load (Fig. 1, right). Whether it limits behavior depends on the retrieval, which sorts models into three regimes under fair elicitation (Fig. 1, left). *Binding-limited* (17): accuracy decays to chance across eight families, including two chat-elicited models, so the limit is not a raw-completion artifact. *Recency-copy* (26): induction heads emit the most recent occurrence, so accuracy stays above chance and exceeds what the geometry allows (18 of 20 chat-elicited models, plus GPT-2 and mid-sized Pythia). *Incapable* (4,

model (u=16)	ordinal "value k" 1st	mid	last	sem. recent
copy-route				
gemma-4-12B-it	0.96	0.24	0.45	0.89
Qwen3-14B	1.00	0.09	0.23	0.92
gemma-2-2B-it	0.41	0.08	0.39	0.64
random-access				
gemma-4-31B-it	1.00	**0.86**	0.86	0.98

model (completion)	u	base	**interv.**
shortcut present (nd0)			
Llama-3.1-8B	16	0.03	0.60
Pythia-6.9B	16	0.12	0.71
Pythia-12B	32	0.06	0.70
Qwen3-1.7B	32	0.06	**0.85**
shortcut removed (nd2; chance 0.03)			
Pythia-6.9B	32	0.00	0.07
Pythia-12B	32	0.02	0.07
Qwen2.5-7B	32	0.03	0.09
Qwen3-1.7B	32	0.02	0.48

Table 1: **Left: matched-cue random access** (u=16, chance 0.06, one ordinal template). Copy-route models collapse on the interior slot while endpoints survive (Prop. 1); gemma-4-31B answers it at 0.86. **Right: the regime is a causally switchable route.** An answer-agnostic gain on a few copy heads at the query token lifts high-load accuracy; ablating the same heads in GPT-2-large reverses it (recency-copy → binding-limited). All interventions pass random-head, wrong-position, and norm-matched nulls, and conversion survives removal of the positional shortcut (nd2), including on two base models where no last-line cue can help.

all base): at chance throughout. Qwen3-1.7B is binding-limited only under completion (its chat template reads recency-copy), so its conversion below is scoped to completion; genuinely binding-limited base models convert regardless (App. I). The map is not a last-line shortcut: under interleaved distractors (nd=2), where blind last-line copying scores at chance, the taxonomy survives (App. Fig. 13). Scale barely tracks capacity ($R^2_{\text{param}} = 0.02$), and all $8/8$ base recency-copy models sit above the capacity their geometry implies (sign test p=0.0039), reading through a route the geometry does not describe; we call this route-gated capacity (App. F, Fig. 27).

4.5 Scale escapes the limit only through the copy route

If capacity were stored competence it should grow with scale wherever models grow. Instead, under fair elicitation, high-interference accuracy (u=16) climbs with parameters inside reasoning-era families (Qwen3 ρ=+0.94, Qwen3.5 ρ=+0.90; pooled over 19 chat models ρ=+0.71, p=7×10^{-4}) and never in base-era families (all 27 base models below 0.37; pooled ρ=−0.37; App. Fig. 29). Every model above 0.40 is chat-elicited ($17/17$). Because every chat model runs under a template and every base model under completion, we measure three backbones four ways (base/instruct weights by completion/chat; App. Fig. 30): instruct weights escape only through their chat template (Llama-3.2-3B 0.42), the same template on base weights stays at chance (0.04), and instruction tuning lowers completion capacity ($0.09 \to 0.02$). The escape needs post-training and the interface together; it is a resolved routing decision, not added storage. An evals corollary follows: for the 19 chat models measured both ways, raw completion is not merely pessimistic (median 0.09 versus 0.79); its ranking no longer tracks the fair one (ρ=0.28, p=0.24) and it erases the scale trend (ρ=+0.06 versus +0.71; Steiger p=0.009; App. Fig. 31). Finally, copy-route models read the latest far better by content (semantic "most recent" 0.64–0.92) than by ordinal index ("value number u" 0.23–0.45; Table 1), so their interior collapse is a missing dereferencing route, not a counting failure: gemma-4-31B answers the identical middle cue at 0.86.

4.6 An attention signature of random access

Random access is rare and not set by scale (Fig. 3b): among models clearing retrieve-latest across eight families and up to 70B, only gemma-4-31B recovers the interior slot under the matched cue; every other model sits at 0.07–0.39. The mechanistic difference is a query-conditioned pointer head. Cueing gemma-4-31B for the middle slot moves one head's attention onto that occurrence (pointer selectivity π=0.81, layer 47; definition in App. D), supplying the content term $\beta\, s_u[k]$ that a position-only route cannot; gemma-4-12B has no such head (π<0.09) and answers "most recent" at ceiling yet cannot fetch an interior slot. Retrieve-latest and random access are two faces of one head-level dichotomy: the copy head returns the endpoint for free, the pointer dereferences an interior address and pays the crowding cost. The pointer is a predictive attention signature, not a localized cause: head ablations self-repair, and gemma-4's QK-normalized, soft-capped attention confounds edge patching (App. D).

4.7 Causal regime conversion, both directions

The regime map is correlational. Here we make it causal, in both directions and outside Qwen.

Ablation. In GPT-2-large (recency-copy) we identify the copy/induction heads as those whose ablation most reduces low-load retrieval and ablate them cumulatively (App. Fig. 38). Ablation steepens the whole curve monotonically (u=8: $0.41 \to 0.32 \to 0.22$ at $0/3/6$ heads) and by six

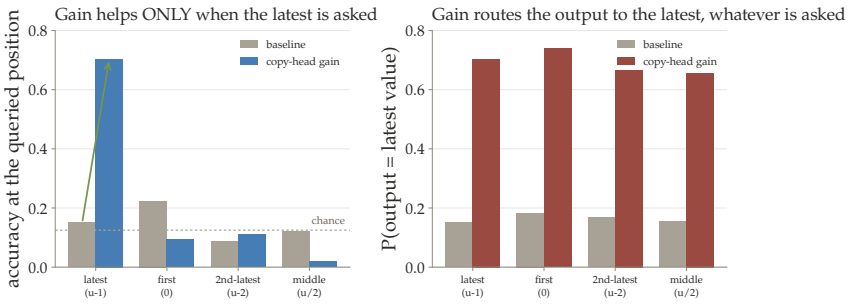


Figure 4: The query-token gain is a recency route, not a generic accuracy gain (Qwen3-1.7B, u=8, chance 0.125). Same stream; only the instruction changes. (a) The frozen copy-head gain lifts accuracy only for the latest query and degrades the others. (b) Across every query it drives the output to the latest value (P(output=latest) $\approx$ 0.66–0.74). Random-head and wrong-position controls (not shown) leave every bar at baseline.

heads pushes the high-load tail into the binding-limited regime. A matched-count random-head ablation leaves the curve recency-copy in all five draws (App. M), so the effect is specific to the copy heads; it reproduces directionally in Pythia-1.4B.

Amplification. We raise a binding-limited model's high-load retrieval by strengthening its copy route: a scalar gain g on a few latent copy heads' output, applied at the final query token only. The cleanest cases are base models binding-limited under their own fair interface. The gain converts Llama-3.1-8B (0.03 $\rightarrow$ 0.60) and Pythia-6.9B (0.12 $\rightarrow$ 0.71) at u=16 and a $>$ 10B base (Pythia-12B 0.06 $\rightarrow$ 0.70 at u=32). The most heavily controlled case is Qwen3-1.7B, where the same edit on four heads (L19H3, L18H15, L7H10, L15H9; disjoint from the order-reading head L17H6) lifts u=32 retrieval from 0.06 (chance 0.03) to 0.85, dose-dependently (Fig. 5); since Qwen3-1.7B is binding-limited only under completion, the base-model conversions carry the interface-independent claim. The edit is not answer injection: it touches only the last token, leaving the stored order code untouched by construction, and it scales what the heads already write rather than changing their attention, adding no bit about which value is correct. The route is latent, not installed: these models are not induction-poor (task-external induction 0.92), so the regimes differ in whether retrieval is routed through the copy heads, not in whether they exist.

The conversion survives free generation. Amplification does not merely reshuffle logits among the presented values. On base models the gain converts free greedy generation, the full-vocabulary argmax of the emitted token, at u=16: Pythia-6.9B 0.05 $\rightarrow$ 0.29, Pythia-12B 0.03 $\rightarrow$ 0.42, Llama-3.1-8B 0.00 $\rightarrow$ 0.35, with random-head and wrong-position controls null on every model (App. Tab. 2). Two-token values convert as well (App. D).

The gain is a recency route, not a generic accuracy gain. Holding the stream fixed and varying only the instruction to report the value at a given position (latest, first, second-latest, middle; Qwen3-1.7B, u=8, chance 0.125; Fig. 4), the gain lifts accuracy only for the latest query (0.15 $\rightarrow$ 0.70) and degrades the rest (first 0.22 $\rightarrow$ 0.09, middle 0.12 $\rightarrow$ 0.02). Across all four queries it drives the output to the latest value (P(output=latest) 0.66–0.74 whatever is asked). A generic accuracy intervention would lift all four positions; a recency route lifts exactly one, as the account predicts. The dissociation reproduces in Qwen3-4B and Qwen2.5-1.5B, and controls leave every query at baseline. A counterfactual value swap confirms the edit carries no fixed answer: replacing only the latest value with a fresh token w makes the gain return w (0.21 $\rightarrow$ 0.59) and the removed original on 0.000 of prompts (baseline 0.037). Further controls, including interface patching and multi-token values, are in App. D.

Controls (Table 1). On Qwen3-1.7B (n=250), random heads, the same heads at a non-final position, and a norm-matched random vector all stay at baseline. The control is matched in budget, not just setting: on the same $k \leq 8$, $g \leq 32$ grid scored at its maximum (u=32, n=150), the random-head arm reaches 0.18 against the copy heads' 0.60, and the same matched-budget null holds on base models (Pythia-12B 0.72 versus 0.17; Pythia-6.9B 0.32 versus 0.07). Under interleaved distractors (nd=2) the experiment still converts 0.02 $\rightarrow$ 0.48 on Qwen3-1.7B with the copy heads transferred unchanged from the nd=0 screen, so they are not last-line heads. Decisively, the no-shortcut con-

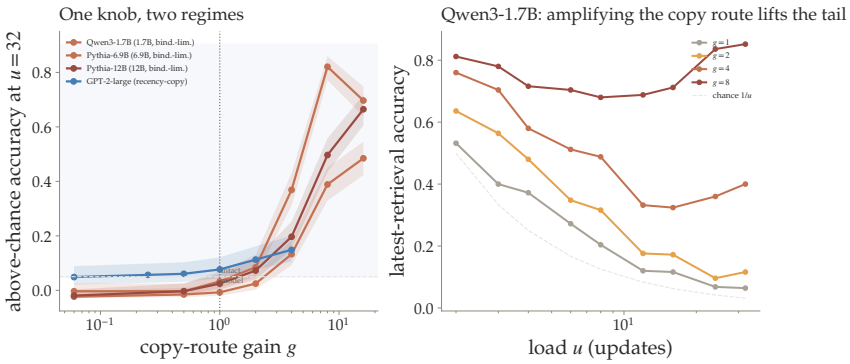


Figure 5: A copy-route gain at the query token moves capacity in both directions. (a) Above-chance accuracy at u=32 against gain g (log axis; g=0 ablation, g=1 intact, g>1 amplify), one curve per model with 95% Wilson bands (n=250 per point). Amplifying Qwen3-1.7B lifts u=32 retrieval $0.06 \rightarrow 0.85$ at g=8; turning the same knob down on GPT-2-large pushes it toward the dashed regime boundary. (b) Qwen3-1.7B's accuracy-versus-load curve lifts off the $1/u$ chance line at every load as g grows. The gain edits only the final query token, so the stored order code is untouched by construction.

version also holds on base models across three families, Pythia-6.9B ($0.00 \rightarrow 0.07$), Pythia-12B ($0.02 \rightarrow 0.07$), and Qwen2.5-7B ($0.03 \rightarrow 0.09$), each a significant lift from at or below chance (0.03; $\approx 3\sigma$, $\geq 2.6\sigma$ after Bonferroni over the gain grid). The magnitude is small on these weaker bases, as the deconfounded probe predicts (§4.1).

Scope and magnitude. The same frozen heads at the same fixed gain convert latest-binding on two natural-language surface forms (person-location $0.12 \rightarrow 0.61$, box-contents $0.07 \rightarrow 0.44$; u=32) and, with English-screened heads untouched, in Spanish ($0.13 \rightarrow 0.61$), French ($0.14 \rightarrow 0.68$), and Chinese (free generation $0.33 \rightarrow 0.50$; App. D), so the route is content-, form-, and language-agnostic. Under a model's own chat template, the gain lifts the smallest chat-binding-limited model, Qwen3-0.6B, $0.02 \rightarrow 0.16$ ($0.03 \rightarrow 0.07$ at nd=2) but is null on Llama-3.2-1B-Instruct; the strong conversions belong to base models under completion. Across 25 models the amplification is positive in 24, and how much a model converts is predicted by a low-dose causal probe of its copy heads (ρ=0.68) but not by their attention to the latest value ($\rho = -0.18$; App. Fig. 10). At a single fixed setting (k=4, gain 16) applied to all 25 models with no tuning, mean conversion is +0.22 (positive in 20/25) and the flagship numbers are unchanged, so they are not a grid-search artifact. Prompting does not substitute: instruction and few-shot arms sit at baseline, and chain-of-thought closes ≈60% of the gap at a large token cost, only where the route exists (App. D).

One knob, both directions. The same continuous knob lowers GPT-2-large as $g \rightarrow 0$ and lifts Qwen3-1.7B as g grows (Fig. 5a); across binding-limited (Qwen2.5/3, Llama) and recency-copy (GPT-2, Pythia) models, amplification raises high-load retrieval and ablation lowers it, each in its baseline regime's direction, with the context fixed throughout (App. Fig. 33). Copy-route strength causally controls retrieve-latest capacity.

4.8 Benchmarks and naturalistic tasks

The same failure, a model silently returning a stale value after an in-view update, appears on three published benchmarks and on templated code, dialogue, and document streams. Here the repair is input-level eviction rather than the query-token gain, which acts only in the middle interference regime and is null where the baseline is already ≈0 (PI-LLM) or the interference is mild (MultiWOZ, $0.63 \rightarrow 0.51$–0.64 across gains, controls at baseline). **PI-LLM and bAbI.** Under PI-LLM's proactive-interference configuration (Wang & Sun, 2025) (46 interleaved keys, each rebound n_{updates} times in prose), retrieve-current accuracy falls from 0.57 at two updates to ≤0.05 by six, with a stale value returned on up to 99% of queries under both constrained choice and the benchmark's own generation metric. An oracle eviction that deletes only the queried key's stale entries lifts generation accuracy to 0.97–1.00: interference, not a capacity wall, defeats retrieval. The collapse and restore hold across six models from five families (App. Figs. 17, 18) and reproduce on narrative bAbI (Weston et al., 2015) (App. Fig. 19). **MultiWOZ 2.2.** In real task-oriented dialogues (Budzianowski

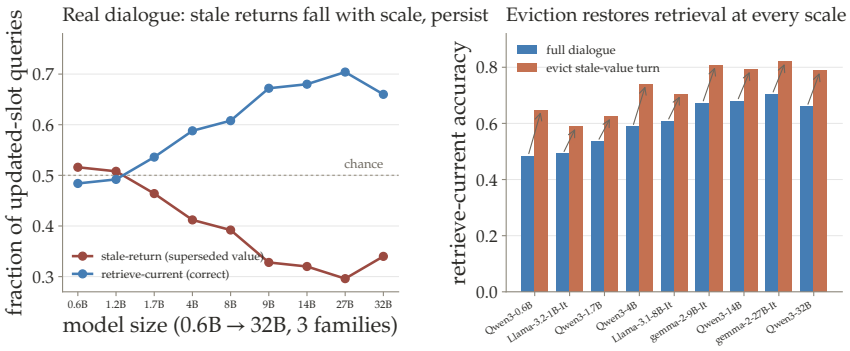


Figure 6: The stale-value effect on real MultiWOZ 2.2 dialogues (Budzianowski et al., 2018) where the user updates a slot (constrained choice over the slot's values; chance ≈0.5; 250 probes per model; nine models, three families). (a) Retrieve-current is at chance for the smallest models and plateaus with scale, so even a frontier-scale model returns the superseded value on about a third of queries. (b) Deleting the turn that introduced the stale value raises retrieval at every scale, the same input-level eviction that repairs PI-LLM and bAbI.

et al., 2018), users routinely revise a slot ("a hotel in the east" . . . "actually, the centre"), leaving the superseded value in view. Scoring the current value by constrained choice over the values the slot took (teacher-forced, length-normalized log-probability; chance ≈0.5; 250 queries per model), nine Qwen, Llama, and gemma models (0.6B–32B) sit at chance at the smallest sizes (Qwen3-0.6B 0.48) and plateau with scale (Qwen3-14B 0.68, gemma-2-27B 0.70, Qwen3-32B 0.66): even at frontier scale, about a third of updated-slot queries return the stale value (Fig. 6a). Deleting the single turn that introduced it lifts retrieval at every scale (+9 to +16 points, to 0.82; Fig. 6b).

Naturalistic forms and actions. The collapse appears on a code assistant, a multi-turn dialogue, and a document stream (superseded values on 72–94% of high-load queries; Fig. 16), and the copy heads frozen from the synthetic task repair the binding-limited code form with no code-specific labels (0.08 → 0.78). The failure also corrupts actions: a RAG-style gatekeeper that acts correctly at u=1 (0.97) falls to chance under interference, a confidently wrong action on 22–39% of trials (Chao et al., 2026), which stale-binding eviction drives to ≈0.

5 Discussion

What the result means. The number of simultaneously resolvable bindings is the packing capacity of a fixed-extent code under retrieval noise, resource-like rather than a slot count (Ma et al., 2014; Gurnee et al., 2026). (i) For copy-route models, capacity is the price of random access: their only path to a non-final binding is the crowding order route. (ii) Interference benchmarks measure the route, not the store: every sink-controlled model writes the same crowding form, yet models sit at opposite ends of the task depending on which circuit is engaged and which interface elicits it. (iii) Latent capability can exceed elicited capability: a scalar gain on four of a model's own heads elicits near-ceiling retrieval where prompting does not. Memory benchmarks should therefore be scored under fair elicitation with non-final probes, and a behavioral floor read as unrouted computation, not absent information. **Limitations.** (1) The strong conversion (0.06 → 0.85) is a single-model existence proof; the population claim is its direction (20 of 25 models at a fixed setting), not its magnitude. (2) The geometry is a sink-controlled description, not a predictive law: the closed-form u^* and direct order-code interventions carry no causal weight (App. E.6); the routing interventions do. (3) Items are single-token, the mechanism is characterized on Qwen and Llama, and the three regimes discretize a continuous axis (the binding-limited count moves 17 → 23 as the threshold shifts 0.05 → 0.08), so the argument rests on curves, not counts. (4) The metric is constrained choice, verified under free generation and interleaved distractors for every headline. **Conclusion.** Apparent retrieve-latest "working-memory capacity" is largely controlled by retrieval routing. A query-token edit to a few copy heads flips the regime in both directions with the context provably untouched, so a benchmark failure need not mean the value is absent. The copy route buys robustness with a fixed address; genuine random access is rare, not bought by scale, and carries a distinctive query-conditioned pointer signature. Language models address their context by recency and content, not by ordinal position.

AI USE STATEMENT

In this work, we used generative AI tools to generate datasets, formulate mathematical claims, provide suppor for mathematical proofs, assist in writing and refining proofs, and implement methods through code generation and debugging. We did not use generative AI tools for the other tasks requiring disclosure, and the remaining required-disclosure tasks were not applicable to this work.

Additionally, we used generative AI tools to create or modify scientific figures and research artifacts, create or edit software code, and suggest possible titles and keywords for the paper.

All AI-assisted outputs were reviewed and validated by the authors. Synthetic data, mathematical claims, proofs, code, figures, and artifacts were checked for correctness and consistency with the reported methodology and results. Title, keyword, and presentation suggestions were used only as editorial assistance.

We take responsibility for the final content of this work, including text, claims, code, figures, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies open-weight language models on synthetic and publicly released benchmark data (PI-LLM, bAbI, MultiWOZ 2.2) and involves no human subjects or private data. The interventions are diagnostic edits at inference time on a single query token and do not produce new capabilities beyond retrieving values already present in the context. The practical findings (stale-value errors in agents, retrieval pipelines, and dialogue systems) are safety-relevant, and we release the screening protocol so that practitioners can detect the failure before deployment.

REPRODUCIBILITY STATEMENT

All code will be released with the paper; every number and figure regenerates from our shared results. Verbatim prompts, the scoring rule, the head-screening and intervention protocol, and the train/validation/test load split are in Apps. A and B; proofs and the assumptions behind Prop. 1 are in App. C; the model set behind each claim is listed in App. H; and every reported negative result, with its control, is in App. E.6.

REFERENCES

Micah Adler and Nir Shavit. On the complexity of neural computation in superposition. 2026. URL https://arxiv.org/abs/2409.15318.

Nicholas Barnfield, Juno Kim, Eshaan Nichani, Jason D. Lee, and Yue M. Lu. Sharp capacity thresholds in linear associative memory: From top-1 retrieval to tail-average learning. 2026. URL https://arxiv.org/abs/2605.05189.

Saksham Bassi and Sharvi Tomar. Geometry of ordinal representations in language models. 2026. URL https://openreview.net/forum?id=Hle1jgzG0f.

Trenton Bricken and Cengiz Pehlevan. Attention approximates sparse distributed memory. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=WVYzd7GvaOM.

Gordon D. A. Brown, Ian Neath, and Nick Chater. A temporal ratio model of memory (simple). *Psychological Review*, 114(3):539–576, 2007.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi

Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5016–5026, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1547. URL https://aclanthology.org/D18-1547/.

Vivien Cabannes, Elvis Dohmatob, and Alberto Bietti. Scaling laws for associative memories. 2024. URL https://arxiv.org/abs/2310.02984.

Declan Campbell, Sunayana Rane, Tyler Giallanza, Nicolò De Sabbata, Kia Ghods, Amogh Joshi, Alexander Ku, Steven M. Frankland, Thomas L. Griffiths, Jonathan D. Cohen, and Taylor W. Webb. Understanding the limits of vision language models through the lens of the binding problem. 2025. URL https://arxiv.org/abs/2411.00238.

Violet Cane. Individual choice behavior: A theoretical analysis. *Royal Statistical Society. Journal. Series A: General*, 123(4):486–488, 07 1960. ISSN 0035-9238. doi: 10.2307/2343282. URL https://doi.org/10.2307/2343282.

Hanxiang Chao, Yihan Bai, Rui Sheng, Tianle Li, and Yushi Sun. Stale: Can llm agents know when their memories are no longer valid?, 2026. URL https://arxiv.org/abs/2605.06527.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. 2023. URL https://arxiv.org/abs/2304.14997.

Nelson Cowan. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24:87–114, 10 2001. doi: 10.1017/S0140525X01003922.

Qin Dai, Benjamin Heinzerling, and Kentaro Inui. Representational analysis of binding in language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17468–17493, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.967. URL https://aclanthology.org/2024.emnlp-main.967/.

Qin Dai, Benjamin Heinzerling, and Kentaro Inui. Cell-based representation of relational binding in language models. 2026. URL https://arxiv.org/abs/2604.19052.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. 2021. URL https://arxiv.org/abs/2006.00995.

Jiahai Feng and Jacob Steinhardt. How do language models bind entities in context? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zb3b6oKO77.

Jiahai Feng, Stuart Russell, and Jacob Steinhardt. Monitoring latent world states in language models with propositional probes. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=0yvZm2AjUr.

E. Paxon Frady, Denis Kleyko, and Friedrich T. Sommer. Theory of the superposition principle for randomized connectionist representations in neural networks. 2017. URL https://arxiv.org/abs/1707.01429.

Dongyu Gong and Hantao Zhang. Self-attention limits working memory capacity of transformer-based models. 2024. URL https://arxiv.org/abs/2409.10715.

Dongyu Gong, Xingchen Wan, and Dingmin Wang. Working memory capacity of chatgpt: An empirical study. 2024. URL https://arxiv.org/abs/2305.03731.

Rhys Gould, Euan Ong, George Ogden, and Arthur Conmy. Successor heads: Recurring, interpretable attention heads in the wild. 2023. URL https://arxiv.org/abs/2312.09230.

Xiaobo Guo and Soroush Vosoughi. Serial position effects of large language models. 2024. URL https://arxiv.org/abs/2406.15981.

Yoav Gur-Arieh, Mor Geva, and Atticus Geiger. Mixing mechanisms: How language models retrieve bound entities in-context. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=UJ2UUjT2ko.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models. 2026. URL https://arxiv.org/abs/2607.15495.

Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long T. Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. Found in the middle: Calibrating positional attention bias improves long context utilization. 2024. URL https://arxiv.org/abs/2406.16008.

Najoung Kim and Sebastian Schuster. Entity tracking in language models. 2023. URL https://arxiv.org/abs/2305.02363.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. Snapkv: Llm knows what you are looking for before generation, 2024. URL https://arxiv.org/abs/2404.14469.

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. volume 12, pp. 157–173, Cambridge, MA, 2024. MIT Press. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Steven J. Luck and Edward K. Vogel. The capacity of visual working memory for features and conjunctions. *Nature*, 390:279–281, 1997.

Wei Ji Ma, Masud Husain, and Paul M. Bays. Changing concepts of working memory. *Nature Neuroscience*, 17(3):347–356, 2014.

Callum McDougall, Arthur Conmy, Cody Rushing, Thomas McGrath, and Neel Nanda. Copy suppression: Comprehensively understanding an attention head. 2023. URL https://arxiv.org/abs/2310.04625.

Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=-h6WAS6eE4.

Klaus Oberauer. Working memory capacity limits memory for bindings. *Journal of Cognition*, 2 (1):40, 2019.

Soyoung Oh and Vera Demberg. A retrieval conditioned rebinding circuit for dynamic entity tracking in large language models, 2026. URL https://arxiv.org/abs/2606.08644.

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

Francesco Ortu, Zhijing Jin, Diego Doimo, Mrinmaya Sachan, Alberto Cazzaniga, and Bernhard Schölkopf. Competition of mechanisms: Tracing how language models handle facts and counterfactuals. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8420–8436, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.458. URL https://aclanthology.org/2024.acl-long.458/.

Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=8sKcAWOf2D.

Nikhil Prakash, Natalie Shapira, Arnab Sen Sharma, Christoph Riedl, Yonatan Belinkov, Tamar Rott Shaham, David Bau, and Atticus Geiger. Language models use lookbacks to track beliefs. 2026. URL https://arxiv.org/abs/2505.14685.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. 2024. URL https://arxiv.org/abs/2310.11324.

Roger N. Shepard. Toward a universal law of generalization for psychological science. *Science*, 237 (4820):1317–1323, 1987.

Nishit Singh. Emergent ordinal geometry in transformers trained on local comparisons, 2026. URL https://arxiv.org/abs/2606.01269.

Paul Smolensky. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46(1–2):159–216, 1990.

Shangwen Sun, Alfredo Canziani, Yann LeCun, and Jiachen Zhu. The spike, the sparse and the sink: Anatomy of massive activations and attention sinks, 2026. URL https://arxiv.org/abs/2603.05498.

Zilu Tang, Qiao Zhao, Gabriel Franco, Derry Wijaya, Aaron Mueller, Sebastian Schuster, and Najoung Kim. Do language models track entities across state changes? 2026. URL https://arxiv.org/abs/2605.30233.

Pavlo Vasylenko, Hugo Pitorro, Andre Martins, and Marcos Vinicius Treviso. Long-context generalization with sparse attention. 2026. URL https://openreview.net/forum?id=PsB6Lynznk.

Petar Veličković, Christos Perivolaropoulos, Federico Barbero, and Razvan Pascanu. Softmax is not enough (for sharp size generalisation). 2025. URL https://openreview.net/forum?id=S4JmmpnSPy.

Boshi Wang and Huan Sun. Is the reversal curse a binding problem? uncovering limitations of transformers from a basic generalization failure. 2026. URL https://arxiv.org/abs/2504.01928.

Chupei Wang and Jiaqiu Vince Sun. Unable to forget: Proactive interference reveals working memory limits in llms beyond context length. 2025. URL https://arxiv.org/abs/2506.08184.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. 2015. URL https://arxiv.org/abs/1502.05698.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=EytBpUGB1Z.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. 2024. URL https://arxiv.org/abs/2309.17453.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. 2024. URL https://arxiv.org/abs/2309.16042.

Qingru Zhang, Chandan Singh, Liyuan Liu, Xiaodong Liu, Bin Yu, Jianfeng Gao, and Tuo Zhao. Tell your model where to attend: Post-hoc attention steering for LLMs. 2024. URL https://openreview.net/forum?id=xZDW0oejD.

A EXPERIMENTAL SETUP

Full prediction protocol. We fit the Luce parameters (β, λ) on train loads $u \in \{2, 3, 4\}$; select the retrieval layer ℓ by accuracy RMSE of the geometry-only model on validation loads $u \in \{5, 6, 7, 8\}$; fit the single gain κ of Eqs. equation 7 and equation 8 on train and validation loads ($u \leq 8$); and report all held-out numbers on test loads $u \in \{10, 12, 14, 16, 20, 24, 28, 32\}$. No test load informs any parameter.

Capacity summaries (definitions). From an observed accuracy curve $A(u)$ over the measured loads ($u_{\min}$=2 to $u_{\max}$=32) the interference-endurance score is the chance-referenced area under the curve per octave of load,

$$\text{IES} \;=\; \frac{1}{\log_2 u_{\max} - \log_2 u_{\min}} \int_{u_{\min}}^{u_{\max}} \left[A(u) - \tfrac{1}{u}\right] d(\log_2 u), \tag{6}$$

evaluated by the trapezoid rule on the measured loads (IES $= 0$ means at chance everywhere). The capacity number u^* is the load at which $A(u)$ first crosses 0.5, log-linearly interpolated between measured loads (u^*=∞ if never crossed). Since 0.5 equals chance at u=2, u^* is floor-compressed and treated as suggestive. Regimes are classified from the curve: incapable if $\max_{u\in\{2,3,4\}} \left[A(u) - \frac{1}{u}\right] \leq 0.06$; otherwise binding-limited if $A(32) - \frac{1}{32} < 0.05$, else recency-copy.

Models are run in bfloat16 with eager attention and forward hooks (no TransformerLens or nnsight dependency). Entities are single-token names; values are single-token words filtered per tokenizer. For each load u we generate prompts with the tracked key rebound u times in temporal order. In the default setting (nd=0) the updates are consecutive, so the final line is the answer; in the nd$\geq$1 companion, neutral distractor lines are interleaved so the tracked key's latest value is not the final line and must be matched by content (see the task-description caveat in §3). Behavioral accuracy is the constrained choice among the u presented values; the geometry is the mean key-token residual per slot over 300 content-randomized prompts. The geometric-Luce model (Eq. 3) and baselines are fit by maximum likelihood on the train loads, and the retrieval layer is chosen by validation accuracy RMSE. The packing law (Eq. 7, D=u) and its measured-spacing variant (Eq. 8) are each fit by their single SNR gain κ on the train and validation loads.

B TASK PROMPTS AND FULL PROTOCOL

This appendix gives the verbatim prompts and the intervention protocol, complementing App. A. Names (72) and values (129) are drawn from single-token word pools and sampled without replacement, so the u competing values are distinct.

Proactive-interference prompt (base models, completion). One tracked key is rebound u times; the model completes the final line with the latest value. Example (u=4, nd=0, updates consecutive):

```
Emma:  camel
Emma:  Moscow
Emma:  sword
Emma:  miner
Emma:
```

The correct completion is `miner`.

Interleaved-distractor variant (nd=2). Distractor keys are interleaved between the tracked key's rebindings, so the answer is not the final line and blind copying fails (the $0.02 \to 0.48$ condition):

```
Dave:  pink
Frank:  mint
Alice:  ruby
Dave:  plum
...
Alice:  blue
Frank:  olive
Dave:
```

Sink-free positional variant (geometry and random access). A neutral header line is prepended so no measured token sits at sequence position 0 (the sink control, §D):

```
Kira:  birch
Bob:  gold
Bob:  teal
Bob:  green
Bob:  lime
Bob:
```

Chat and reasoning models (fair elicitation). The identical task is wrapped in the model's own chat template (generation prompt added; thinking disabled for hybrid reasoners), with system message

```
You are tracking variables that are
reassigned over time.  When asked for a
variable's current value, reply with ONLY
that one value, nothing else.
```

and a user turn holding the key:value block (plus, for random access, a cued query).

Cued-recall queries (random access). Appended to the user turn; the answer is the value at the cued occurrence:

```
cued-latest: {k}'s most recent value was:
cued-first:  {k}'s very first value was:
cued-middle: Counting assignments from 1,
             {k}'s value number {n} was:
```

The middle slot ($n{=}\lfloor u/2 \rfloor{+}1$) is the true random-access probe, neither primacy (unlike first) nor recency (unlike latest).

Scoring. At the final token we compare the logits on the first token of each of the u presented values (the better-scoring of the space-prefixed and bare forms) and count a prediction correct if and only if its argmax is the latest value or, for cued recall, the cued value. Chance is $1/u$. Behavioral accuracy uses 500 content-randomized prompts per load for the primary model and 200–300 for the cross-model sweep, and the geometry uses 300.

Copy-route screening and interventions. Copy heads are identified by ranking all heads by mean attention from the final query token to the latest value token (top-24 candidates), then keeping the k with the largest single-head causal response, the rise in retrieve-latest accuracy when that one head is amplified by a fixed gain of 4 at load $u{=}16$ (150 prompts). The constructive conversion applies a scalar gain g on the top-k copy heads' output at the final query token only. Controls are k random heads from the same layers, the same gain at the wrong token position, and a norm-matched random vector (all null). The cross-model moderator (§4.7) grid-searches $k{\leq}8$, $g{\leq}32$ at high load for the best achievable conversion, and correlates it with a low-dose causal probe (summed retrieve-latest gain from amplifying the top-4 copy heads) and with those heads' summed attention to the latest value. To rule out per-model grid optimism we also report the conversion at one fixed setting ($k{=}4$, $g{=}16$) applied to every model (script `67`); it reproduces the flagship magnitudes exactly and averages $+0.22$ over 25 models. Ablation sets the gain on the screened heads to 0.

Geometry and order interventions. Geometry is the mean key-token residual per slot at the validation-selected layer (extent, participation ratio, median inter-slot spacing, and the crowding exponent from spacing against $(u{-}1)$), always with the position-0 sink controlled. Occurrence-versus-semantic is the extent along the occurrence-order axis against the semantic-value axis (ratio > 1 means occurrence-indexed). The RoPE-aware order test screens the heads that most attend to the cued non-final slot, ablates them, and applies a key-side positional swap (rotating the target occurrence's post-RoPE key to another slot's position); all are null (§E.6).

C THE TWO-REGIME PROPOSITION AND THE PACKING FORM (ORGANIZING ACCOUNT)

This appendix states the formal version of the route-logit decomposition of §3. Its status is as scoped there: the quantitative u^* prediction of part (i) is not supported out of sample (§E.6), while the qualitative route-gating corollary of part (ii) holds ($8/8$ recency-copy models above their geometry-implied capacity, sign test p=0.0039).

The packing form. When codes are random and equidistant, the noisy-argmax retrieval has a classical closed form. Eq. equation 3 is its Gumbel-noise case (argmax over Gumbel-perturbed scores is exactly the softmax). With Gaussian retrieval noise one obtains the superposition-retrieval law (Frady et al., 2017): for $M{=}u$ bundled items, the retrieval signal-to-noise ratio is $s(u) = \kappa/\sqrt{u}$, and the winner-take-all accuracy over a codebook of D candidates is

$$A_{\text{pack}}(u) = \int_{-\infty}^{\infty} \varphi(h)\,\big[\Phi\big(h + s(u)\big)\big]^{D-1}\,dh, \qquad s(u) = \kappa/\sqrt{u}, \tag{7}$$

with φ, Φ the standard normal pdf and cdf, $\kappa > 0$ a single fitted SNR gain (absorbing $\sqrt{d}$ and code statistics), and $D{=}u$ throughout: the codebook is the u presented options, matching the constrained-choice score and its $1/u$ chance floor. Real codes are neither random nor equidistant, and we relax the random-code assumption in two ways. Eq. equation 3 replaces the uniform-similarity assumption with the measured similarities $s_u[k]$, which is what lets it capture model-specific structure, similarity dependence, and confusion gradients. A complementary measured-geometry variant of Eq. equation 7 instead drives the SNR with the measured median spacing,

$$s(u) \;=\; \kappa\,\Delta(u)/\Delta(2), \tag{8}$$

so the decay is set by the observed crowding rather than an assumed $1/\sqrt{u}$ noise growth. Both are descriptive summaries of the measured geometry, not predictive laws (§4.2).

An attention-level account of the two regimes. We model the retrieval softmax of Eq. equation 3 as a retrieval head's attention distribution over the u occurrence tokens, with query–key logit $z_k = \beta\, s_u[k] + \lambda\rho_k$: $\beta\, s_u[k]$ is the order-code component of $\mathbf{q}\mathbf{K}_k$ and $\lambda\rho_k$ a positional/recency component ($\rho_k = k - (u{-}1)$). Under three assumptions we can check, the capacity dichotomy follows. **A1** (crowding margin): the manifold is low-dimensional and the query aligns with the latest code, so the top order margin, in logit units, scales with the median spacing normalized by the extent, $g(u) \equiv \beta(s_u[u{-}1] - s_u[u{-}2]) = g_0\,(u{-}1)^{-p}$ with $g_0 = \beta\,\Delta(2)/E = \beta$ (two codes span the extent, so $\Delta(2){=}E$), by Eq. equation 2. **A2** (translation-invariant recency): the recency prior depends only on the occurrence gap, so consecutive occurrences carry a constant margin λ. **A3** (Gumbel noise) of scale σ, making the noisy argmax exactly $\mathrm{softmax}(z)$. All three are dimensionally consistent (g and λ are logits) and empirically testable (§4.2).

Proposition 2 (Two capacity regimes from the query–key structure) *Under A1–A3:* (i) Order route ($\lambda{=}0$). $\Pr[\textit{report}{=}u{-}1]$ *is decreasing in* u *and crosses a fixed accuracy criterion at a finite load. In the crowded regime the competing slots are near-equidistant from the query, so the* $\frac{1}{2}$*-crossing for the* $m_{\text{eff}}(u) \le u{-}1$ *effective competitors is* $g(u) = \sigma \log m_{\text{eff}}(u)$*; inverting* $g(u) = g_0(u{-}1)^{-p}$ *gives*

$$u^* - 1 \;\approx\; \Big(\tfrac{g_0}{\sigma\,\log m_{\text{eff}}}\Big)^{1/p}, \qquad g_0 = \beta, \tag{9}$$

so the geometry gain β *sits inside the exponent (raised to* $1/p$*), not as a prefactor:* $u^*{-}1$ *is of order* $(\beta/\sigma)^{1/p}$ *up to the slowly varying competitor factor* $(\log m_{\text{eff}})^{-1/p}$*. That factor is itself the same order as the effect over our load range (*$\log(u{-}1)$ *runs* $0 \to 3.4$*), one reason this closed form is descriptive rather than predictive (§E.6). Retrieve-latest is capacity-limited, set by the gain-to-noise ratio* β/σ *and the crowding exponent* p*.* (ii) Copy route ($\lambda{>}0$). *By A2 the latest occurrence keeps a margin* λ *over its predecessor that does not vanish with* u*, so accuracy is bounded below by a positive constant. For a linear recency prior,* $\Pr[\textit{report}{=}u{-}1] = \big(1 + \sum_{m=1}^{u-1} e^{-\lambda m/\sigma}\big)^{-1} \to 1{-}e^{-\lambda/\sigma}$*, decreasing in* u *but never toward chance. (A strongly saturating recency kernel, whose top gap itself shrank with load, would forfeit this floor; the observed recency-copy plateau is evidence*

that the top gap does not saturate.) Over the load range we test, this reads as "no limit". (iii) Random access. *For a non-final target $k^* < u{-}1$ the recency prior ranks the latest above k^* ($\rho_{k^*} < \rho_{u-1}$), so within this two-term family no λ selects k^*; its accuracy rests solely on the order margin and obeys (i) under every route in the family. (A non-recency address, a primacy/attention-sink term, lies outside the family. That random access is nonetheless bounded by the order code rests on the structural fact that a copy route cannot address a non-final slot and on the behavioral evidence of §4.3 and the frontier random-access dissociation of §4.5, not on this proposition.)*

Proof. A3 makes the report an argmax over $\{z_k + \varepsilon_k\}$ with ε_k i.i.d. Gumbel, so $\Pr[\text{report}{=}u{-}1] = \big(1 + \sum_{j \neq u-1} e^{-(z_{u-1}-z_j)/\sigma}\big)^{-1}$, which is increasing in every gap $z_{u-1} - z_j$ and hence decreasing in u as the gaps shrink. (i) With $\lambda{=}0$ the nearest gap is $g(u) = g_0(u{-}1)^{-p}$ (A1); in the crowded regime the $m_{\text{eff}}(u)$ near-equidistant competitors make the sum reach 1 (the $\frac{1}{2}$-crossing) at $g(u) = \sigma \log m_{\text{eff}}(u)$, and inverting gives Eq. equation 9, with β/σ inside the exponent. (ii) A2 gives $z_{u-1} - z_{u-2} \geq \lambda > 0$ for all u; summing the geometric competitor series $\sum_m e^{-\lambda m/\sigma}$ bounds $\Pr[\text{report}{=}u{-}1]$ below by $1 - e^{-\lambda/\sigma} > 0$, independent of u. (iii) $\rho_{k^*} - \rho_{u-1} = k^* - (u{-}1) < 0$, so the recency term favours the latest over k^*; only βs_u separates k^* from its neighbours, reducing to (i). □

Most ingredients are measured, not assumed: A1's power-law margin and the per-head (β, λ) decomposition (order-reader $\lambda{\approx}0$, copy head $\lambda{>}0$), the crowding exponent p (§4.2), and the spacing-to-noise crossing at the capacity load (Fig. 8). The one property we assume rather than measure is discrete convexity of the position kernels. Under RoPE the per-slot logits can be non-convex in relative position, which would place a head outside the family, so the endpoint prediction is best read as a statement about the recency/primacy family, borne out by the endpoint-structured behavioral profile of Fig. 3, rather than a proven property of every head. The quantitative test of (i)'s u^* scaling (sink-free (a, p) from the header-controlled crowding fit, retrieval noise from a paired within-prompt estimator, thresholds fixed in advance) returned out-of-sample Spearman 0.56 (permutation $p{=}0.21$, $n{=}5$), with the fitted $\kappa{=}15.6$ outside its plausible range: not supported (§E.6).

Proof of Prop. 1 (endpoint access). We argue in three steps.

Step 1 (the maximum is at an endpoint). Set $\beta{=}0$, so $z_k = w_k = \lambda\, r(u{-}1{-}k) + \mu\, \pi(k)$. The map $k \mapsto u{-}1{-}k$ is affine and integer-valued, and discrete convexity is preserved under it and under nonnegative combination, so w is discretely convex on $\{0, \ldots, u{-}1\}$; its first differences $d_k = w_{k+1} - w_k$ are non-decreasing. An interior maximizer k would require $d_{k-1} \geq 0 \geq d_k$, which with $d_{k-1} \leq d_k$ forces $d_{k-1} = d_k = 0$ and propagates the maximum outward to an endpoint. Hence

$$\max_{0 \leq k \leq u-1} w_k \;=\; \max\{\, w_0,\; w_{u-1} \,\},$$

so a position-only head strictly prefers only the first or the latest slot.

Step 2 (the interior margin is the order contrast; Eq. equation 5). For $\beta > 0$ and an interior k, discrete convexity gives $w_{k-1} + w_{k+1} \geq 2w_k$, so at least one neighbour $j \in \{k{-}1, k{+}1\}$ has $w_j \geq w_k$. Using $z = \beta s_u + w$,

$$z_k - \max_{i \neq k} z_i \;\leq\; z_k - z_j \;=\; \beta\big(s_u[k] - s_u[j]\big) + \underbrace{(w_k - w_j)}_{\leq 0} \;\leq\; \beta\big(s_u[k] - s_u[j]\big).$$

Step 3 (the rate, under the two regularity assumptions). Write $\widehat{\mathbf{v}} = \mathbf{v}/\|\mathbf{v}\|$, so $s_u[k] = \widehat{\mathbf{q} - \bar{\mathbf{c}}} \cdot \widehat{\mathbf{c}_k - \bar{\mathbf{c}}}$. The elementary two-point bound $\|\widehat{\mathbf{x}} - \widehat{\mathbf{y}}\| \leq 2\|\mathbf{x} - \mathbf{y}\|/(\|\mathbf{x}\| + \|\mathbf{y}\|)$ (equivalently $\|\mathbf{x} - \mathbf{y}\|^2 - \big(\frac{\|\mathbf{x}\| + \|\mathbf{y}\|}{2}\big)^2 \|\widehat{\mathbf{x}} - \widehat{\mathbf{y}}\|^2 = \frac{1}{2}(1 + \cos\theta)(\|\mathbf{x}\| - \|\mathbf{y}\|)^2 \geq 0$) gives, under non-degenerate codes ($\|\mathbf{c}_k - \bar{\mathbf{c}}\| \geq r_0$),

$$s_u[k] - s_u[j] \;=\; \widehat{\mathbf{q} - \bar{\mathbf{c}}} \cdot \big(\widehat{\mathbf{c}_k - \bar{\mathbf{c}}} - \widehat{\mathbf{c}_j - \bar{\mathbf{c}}}\big) \;\leq\; \big\|\widehat{\mathbf{c}_k - \bar{\mathbf{c}}} - \widehat{\mathbf{c}_j - \bar{\mathbf{c}}}\big\| \;\leq\; \tfrac{1}{r_0}\, \|\mathbf{c}_k - \mathbf{c}_j\|.$$

Under quasi-uniform packing ($\|\mathbf{c}_k - \mathbf{c}_j\| \leq C\, \Delta(u)$ for adjacent j), combining with Step 2,

$$z_k - \max_{i \neq k} z_i \;\leq\; \tfrac{\beta C}{r_0}\, \Delta(u) \;=\; O\big((u{-}1)^{-p}\big) \quad \text{by Eq. equation 2,}$$

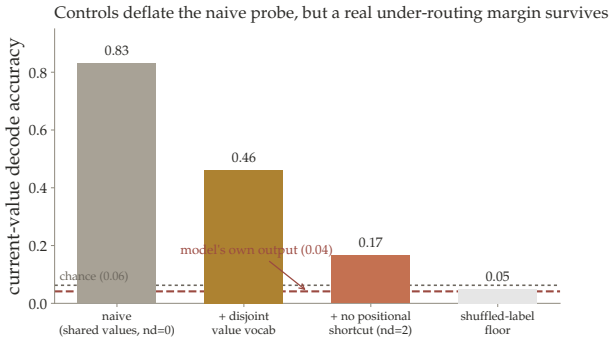


Figure 7: The decoding probe's control decomposition (Qwen3-1.7B, §4.1). The naive probe reads the current value at 0.83 (versus the model's 0.05), but disjoint-value training drops it to 0.46 (partial value memorization) and removing the positional shortcut (nd=2, u=16) drops it to 0.17; a shuffled-label probe sits at chance. A real under-routing margin survives every control, but the naive gap is inflated.

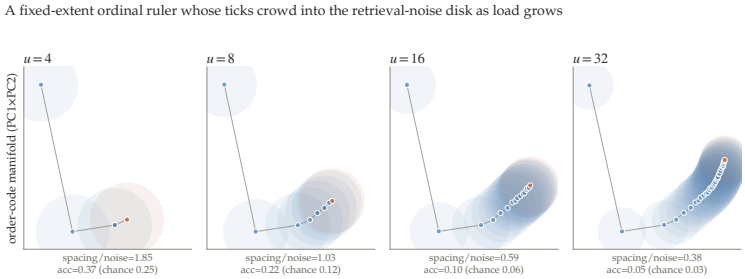


Figure 8: The mechanism in one picture: a fixed-extent ordinal ruler whose ticks fall inside the retrieval noise as load grows (§4.2). Qwen3-1.7B's u content-averaged order codes projected into one shared 2-D PCA plane (light-to-dark blue orders the slots, orange marks the latest); shaded disks are the per-prompt retrieval noise. The extent barely changes, so the spacing-to-noise ratio falls $1.85 \to 0.38$ from u=4 to 32, crossing 1 at $u \approx 8$, the range where above-chance accuracy fades.

where the last step requires C and r_0 to be bounded uniformly in u (not merely at each fixed load). We report the measured ranges of both constants on the sink-controlled geometry (the max-to-median adjacent-gap ratio and $\min_k \|\mathbf{c}_k - \bar{\mathbf{c}}\|$; App. Fig. 11). They stay bounded over the tested loads, and since the codes are only mildly non-collinear (PR 1.1–2.1) we treat the rate as conditional on $r_0 > 0$ rather than as a law. On the uncontrolled geometry quasi-uniform packing fails outright, because a single position-0-sink gap can carry almost the entire extent (Fig. 12); this is why the rate and the sink control it needs are stated together. □

D Additional experiments and controls

The analyses here support the main-text claims off the critical path of the argument. Each keeps its original label, so a main-text cross-reference resolves to this appendix. Figure 7 gives the decoding probe's full control decomposition (0.83→0.46→0.17; its deconfounded cross-model margin is Fig. 2 in the main text), and Figure 8 draws the order manifold whose spacing crowds into the retrieval noise as load grows (§4.2). The remaining figures accompany the controls below.

The query-conditioned pointer head (§4.6). Among models clearing retrieve-latest across eight families and up to 70B, only gemma-4-31B recovers the interior slot under the matched cue (middle 0.86 at u=16). Llama-3.3-70B answers "most recent" at 0.99 but the middle at 0.08, and Qwen3-32B/3.5-27B/3.6-27B (0.11–0.39) and gemma-4-31B's own siblings (gemma-4-E2B/E4B/12B, 0.07–0.24) all fail even though they clear retrieve-latest and count to the first slot perfectly (Table 1). For a head (ℓ, h) define its pointer selectivity as the change in query-token attention to occurrence k

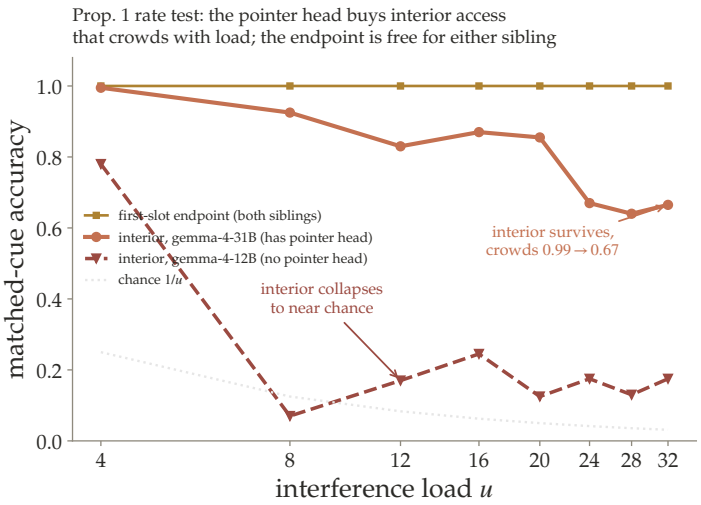


Figure 9: Prop. 1's rate, and a gemma-4 sibling contrast for the pointer signature. Matched ordinal-cue accuracy at the interior (middle) slot against load. The model with the pointer signature (gemma-4-31B) survives at the interior and merely crowds $0.99 \to 0.67$, the $O(\Delta(u))$ rate of Eq. equation 5; the sibling without one (gemma-4-12B) collapses to 0.07–0.25 at every load. Both keep the first-slot endpoint pinned near 1.0. The contrast is correlational (the siblings differ in many weights) and is consistent with the pointer enabling interior access. Either way, interior access pays the crowding tax while the endpoint is free.

when the cue asks for slot k rather than for the latest,

$$\pi_{\ell,h} \;=\; \bar{a}_{\ell,h}(k \mid \text{cue}{=}k) \;-\; \bar{a}_{\ell,h}(k \mid \text{cue}{=}\text{latest}), \qquad k = \lfloor u/2 \rfloor, \tag{10}$$

where $\bar{a}_{\ell,h}(k \mid c)$ is the head's final-token attention to occurrence k averaged over prompts under cue c. A pure copy head attends to the latest regardless of the cue ($\pi \approx 0$); a pointer head that dereferences the ordinal address moves onto the cued slot (π large and positive). In gemma-4-31B this is sharply concentrated: layer 47 head 17 has $\pi{=}0.81$ (0.82 of its query attention on the middle occurrence when cued middle, 0.003 when cued latest), with seven more heads in layers 47–59 above 0.38, all in gemma-4's full-attention layers, the only ones that can reach an arbitrary past occurrence. This realizes the one escape Prop. 1 leaves from endpoint access, a query-conditioned content read that supplies $\beta\, s_u[k] > 0$ at an interior slot. The complementary induction/copy head is the λ recency term of §4.7, which is why strengthening it converts a model toward recency-copy but never grants random access (it moves along the λ axis, not the β axis). This sharpens the correlational route-mixture picture of Gur-Arieh et al. (2026); Prakash et al. (2026) into a single cue-selective head whose presence predicts a discrete capability, and suggests a candidate mechanism for "lost in the middle" (Liu et al., 2024) that we do not test here: the interior may be lost where no pointer can reach it.

Prop. 1's rate, and a sibling contrast for the pointer signature. gemma-4-31B (with the pointer signature) and gemma-4-12B (without a comparable one) let us test the interior-access rate. Scoring the matched ordinal cue at the first, middle, and last slot across loads $u{=}4$–32 (Fig. 9), gemma-4-31B's interior slot pays exactly the geometry cost Eq. equation 5 predicts: middle accuracy crowds $0.99 \to 0.67$ while the first-slot endpoint stays pinned at 1.00 (reachable by a primacy read with $\beta{=}0$) and semantic "most recent" stays at ceiling (0.98–1.00, the copy route). Its sibling gemma-4-12B, which lacks a comparable pointer head, keeps the same pinned endpoint and semantic-latest, but its interior collapses far below, to 0.07–0.25 across $u{=}8$–32 (a few multiples of chance, versus gemma-4-31B's 0.67–0.99). The two models differ in many weights, so this is a correlational contrast, not an isolation. It is consistent with the pointer head being what enables interior access, and it shows that even a model that reaches the interior still pays the crowding tax there while its endpoint stays free.

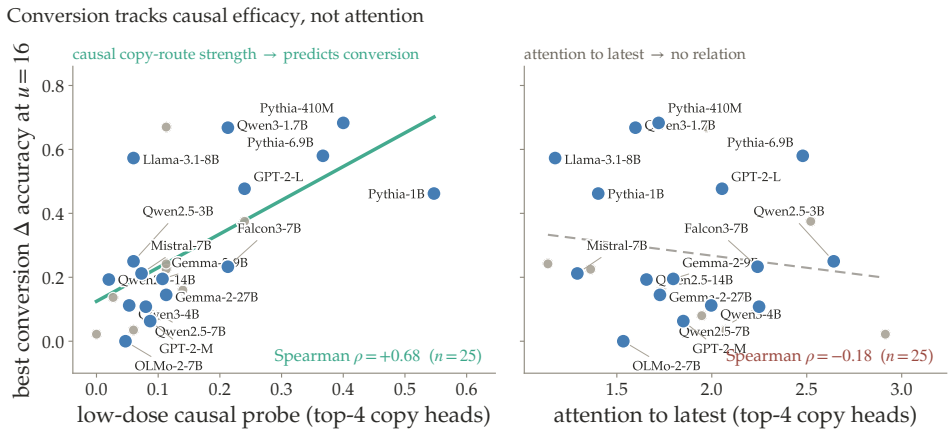


Figure 10: Conversion magnitude is set by the copy route's causal efficacy, not its attention (n=25 models, eleven families; §4.7). Best-achievable conversion (rise in retrieve-latest at u=16, matched budget $k\leq 8$, $g\leq 32$) against two candidate moderators. Left: a low-dose causal probe of the top four copy heads tracks convertibility (Spearman ρ=0.68, permutation $p=2\times10^{-4}$; leave-one-family-out $\rho\in[+0.56,+0.74]$, family-mean ρ=+0.72, block-permutation p=0.0008). Right: the same heads' summed attention to the latest value does not ($\rho=-0.18$; single top head $\rho=-0.14$); the model with the strongest latest-attention converts near zero. Frontier chat models are absent by construction (already copy-routed at ceiling). Probe and outcome share the gain, so this is a low-dose-predicts-high-dose consistency, not an independent law.

Pointer-head causal forensics (§4.6). Two verified reasons explain why the attentional pointer identification does not convert to a clean head ablation on gemma-4-31B. First, zeroing the top pointer heads' output at the query token (the intervention that converts Qwen, GPT-2, and Pythia; the hook fires correctly, with a 22-logit swing when zeroing a full projection) leaves middle retrieval essentially unchanged (u=8 here, $0.92\rightarrow0.91$ even for the top 64 heads, matching a random-head control). This is the expected self-repair signature: the pointer function is written redundantly, not absent. Second, severing the query-to-middle attention edge is confounded on gemma-4, whose QK-normalized, logit-softcapped attention our biased-attention path does not reproduce (passing middle retrieval through that path with zero bias already collapses it, a control we report), while robust latest retrieval survives the same path.

Free generation, and the chat interface (§4.7). Two skeptical readings: that the constrained-choice metric merely reshuffles logits within the presented values, and that raw completion is a deliberately weak interface. We test both (Table 2). *Free generation:* on base models under their fair completion interface, the gain converts the full-vocabulary argmax of the emitted token (values are single-token, so this is greedy exact match) across three models and two families at u=16: Pythia-6.9B $0.05\rightarrow0.29$, Pythia-12B (a $>$ 10B base) $0.03\rightarrow0.42$, and Llama-3.1-8B $0.00\rightarrow0.35$. The free-generation random-head and wrong-position controls are null on every one, so the lift is head- and position-specific, not a scoring artifact. *Chat interface:* copy heads screened on raw completion and applied under the chat template convert Qwen3-0.6B on constrained choice ($0.02\rightarrow0.16$, surviving interleaved distractors, controls null) but are null on the smaller Llama-3.2-1B-Instruct, and chat-mode free generation does not convert on these tiny models. The models that are binding-limited under chat are the smallest available, with the least amplifiable latent route; the strong conversions belong to base models under completion, where they survive free generation.

Cross-surface circuit identity (§4.7). To test whether the copy route is specific to the key:value surface, we take the heads screened once on the key:value task, freeze them, and apply the same fixed gain (k=4, g=16) to two semantically distinct natural-language forms of latest-binding on Qwen3-1.7B (u=32): *lives* ("{name} moved to {place}"; retrieve the current place) and *boxes* ("Box {X} now holds {value}"). The frozen key:value heads convert both: key:value $0.07\rightarrow0.74$ (free generation $0.01\rightarrow0.61$), lives $0.12\rightarrow0.61$ (free generation $0.02\rightarrow0.15$), boxes $0.07\rightarrow0.44$ (free generation $0.02\rightarrow0.17$), each far above the same frozen heads' random-head (0.12–0.21)

model	interface	condition (u=16, chance 0.06)	baseline	+copy-head gain
Pythia-6.9B (base)	completion	**free greedy generation**	0.05	**0.29**
Pythia-12B (>10B)	completion	**free greedy generation**	0.03	**0.42**
Llama-3.1-8B (base)	completion	**free greedy generation**	0.00	**0.35**
Pythia-12B	completion	constrained choice	0.15	0.82
	completion	free-gen random-head / wrong-position	n/a	0.03 / 0.07
Qwen3-0.6B	own chat	constrained choice	0.02	0.16
	own chat	constrained, nd2 (no last-line cue)	0.03	0.07
	own chat	free greedy generation	0.02	0.04
Llama-3.2-1B-Inst.	own chat	constrained choice	0.05	0.05

Table 2: The conversion survives free generation on three base models under completion (Pythia-6.9B, Pythia-12B, Llama-3.1-8B; controls null), and operates under a model's own chat template (Qwen3-0.6B, constrained), but is bounded on the tiniest chat-binding-limited models (Llama-3.2-1B-Instruct null; chat free generation null). Apparent capacity is controlled by routing most clearly where an amplifiable copy route exists.

and wrong-position (0.08–0.12) controls. Heads screened independently on each form overlap the key:value set only partially (Jaccard 0.14), yet the key:value heads suffice for all three, so the same query-token route governs latest-binding across surface forms rather than being tied to one format.

Cross-language transfer (§4.7). With the English-screened heads left untouched and the same fixed gain (u=16), a Spanish binding task converts $0.13 \rightarrow 0.61$ (free generation $0.01 \rightarrow 0.46$), a French one $0.14 \rightarrow 0.68$ ($0.00 \rightarrow 0.52$), and a Chinese one raises free generation of the correct hanzi $0.33 \rightarrow 0.50$ (constrained choice is confounded by CJK tokenization), each well above the same frozen heads' random-head and wrong-position controls. The query-token route is therefore content-, form-, and language-agnostic. An evals corollary: an interference benchmark scored in another language mismeasures capacity for the same routing reason, and the English-derived fix transfers to it for free.

Further controls on the query-token gain (§4.7). *Multi-token values.* Rebuilding the task with two-token value phrases and gaining only the query token lifts constrained retrieve-latest $0.15 \rightarrow 0.40$ and greedy exact match of both emitted tokens $0.03 \rightarrow 0.11$ (Qwen3-1.7B, u=8), with random-head and wrong-position controls near baseline, so the recency route copies multi-token content and the conversion is not an artifact of single-token scoring. *Interface patching.* The copy heads compute the latest under both the model's chat and completion interfaces: injecting the completion heads' query-token output into the chat run lifts retrieval $0.25 \rightarrow 0.56$, with random-head, shuffled-source, and wrong-position controls flat at ≈ 0.24. But a patched head output cannot by itself restore the failing completion interface (≤ 0.11); that gap is a downstream read-out that only amplifying the model's own heads repairs, so we report this control as equivocal. *Address-side knob.* A knob on the address side that carries zero bits about which occurrence is latest also lifts accuracy, so the model's own machinery performs the latest-selection. *Prompting.* Instruction and few-shot arms sit at baseline, and chain-of-thought closes $\approx 60\%$ of the gap at a large token cost, and only in models where the route exists; prompting does not substitute for the route.

Matched-cue random-access control. To remove the semantic-versus-ordinal instruction confound between cued-latest ("most recent value") and cued-middle ("value number n"), we re-score the first, middle, and last slots under a single ordinal template ("Counting assignments from 1, k's value number n was:"), chat-templated and sink-free, with constrained choice over the u tracked values (Table 1). The control is decisive because the counterexample answers the identical template: gemma-4-31B recovers the middle slot at 0.86 (u=16), so the ordinal address is constructible, and the collapse of the copy-route models (gemma-4-12B middle 0.24, Qwen3-14B 0.09) is therefore a retrieval failure, not an address-construction one. Every model is endpoint-structured (the primacy slot stays addressable throughout, consistent with Prop. 1); on smaller counters (Qwen3-4B, Qwen3.5-2B) ordinal addressing is noisier and we do not rely on them. Separately, across copy-route models the semantic "most recent" cue exceeds the ordinal "value number u" cue for the same latest slot by 0.25–0.7, so these models lack a usable ordinal index and read by recency and content.

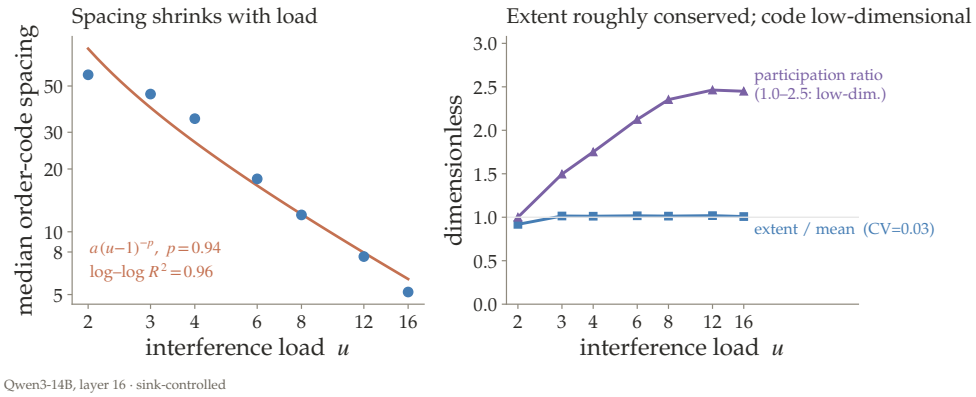


Figure 11: Order-code crowding (sink-controlled). Left: order-code spacing shrinks with load. Right: manifold extent is roughly conserved, and the codes are low-dimensional (PR$\approx$ 1.1–2 across the 30 sink-controlled models). Spacing shrinking within a bounded extent is the resource regime.

SINK CONTROL (CAUTIONARY METHODS RESULT)

A causal-compression test, and why we do not rely on it (a reported null). We attempted to establish causal necessity of the geometry by editing the order-code spacing at the key tokens, rescaling the in-subspace deviation by $\alpha \geq 0$,

$$\mathbf{h}' \;=\; \mathbf{h} + (\alpha - 1)\, P_B(\mathbf{h} - \mathbf{m}), \qquad P_B = BB^\top, \tag{11}$$

with B the top-r order-code basis and $\mathbf{m}$ the cross-slot mean, and re-running the model.

A massive-activation sink can manufacture a causal result (a cautionary methods finding). An apparent large, layer-localized, "subspace-specific" drop at early layers proved to be an artifact of a position-0 activation sink. The anatomy of this sink, a first-token slot of very large norm, is by now well characterized (Xiao et al., 2024; Sun et al., 2026). Our contribution here is not to describe it but to show how easily it fabricates a positive interpretability result. At nd=0 the tracked key's first occurrence is the first token of the sequence, whose residual norm is $\sim$100$\times$ the others and which dominates the extracted subspace, so the edit is really a large perturbation of the sink direction (a random subspace misses it, which is why the "specificity" control looked clean). With the sink controlled by a neutral header line so that no measured token sits at position 0, compressing the sink-free order code does not damage retrieval: the drop is $\lesssim 0.05$ across all layers on capable models with full headroom (gemma-4-E2B-it, baseline 0.79; Qwen3-1.7B), for both latest-cued and non-final-cued recall. We therefore do not claim causal necessity of the geometry via compression. The order code's role rests on the structural and behavioral evidence above, and the routing interventions (§4.7), which are sink-free by construction, are the paper's causal core. We flag this artifact explicitly because it also inflates raw nd=0 geometry statistics, which we therefore measure sink-free throughout (§4.2); Fig. 12 shows the diagnostic and what the one-line control corrects, per model.

Further controls and cross-model context. Three more results sit here. Figure 13 re-runs the cross-model regime map under interleaved distractors (nd=2), where blind last-line copying scores at chance, and confirms that recency-copy is genuine latest-value retrieval (§4.4). Figures 14a–14b then give cognitive-science context: the load u^* at which retrieve-latest crosses 0.5 (against Cowan's human $\approx$4), and how the binding-limited regime emerges over pretraining.

E PRACTICAL MITIGATIONS AND PRACTITIONER GUIDANCE

This appendix gives the real-world evidence of §4.8 in full, then the guidance that follows from it. The same collapse-and-restore signature appears on the published PI-LLM benchmark, on a single model (Fig. 17) and across six families (Fig. 18); on bAbI's narrative surface form (Fig. 19); and on the naturalistic code, dialogue, and document forms of the main text (Fig. 16). An oracle eviction

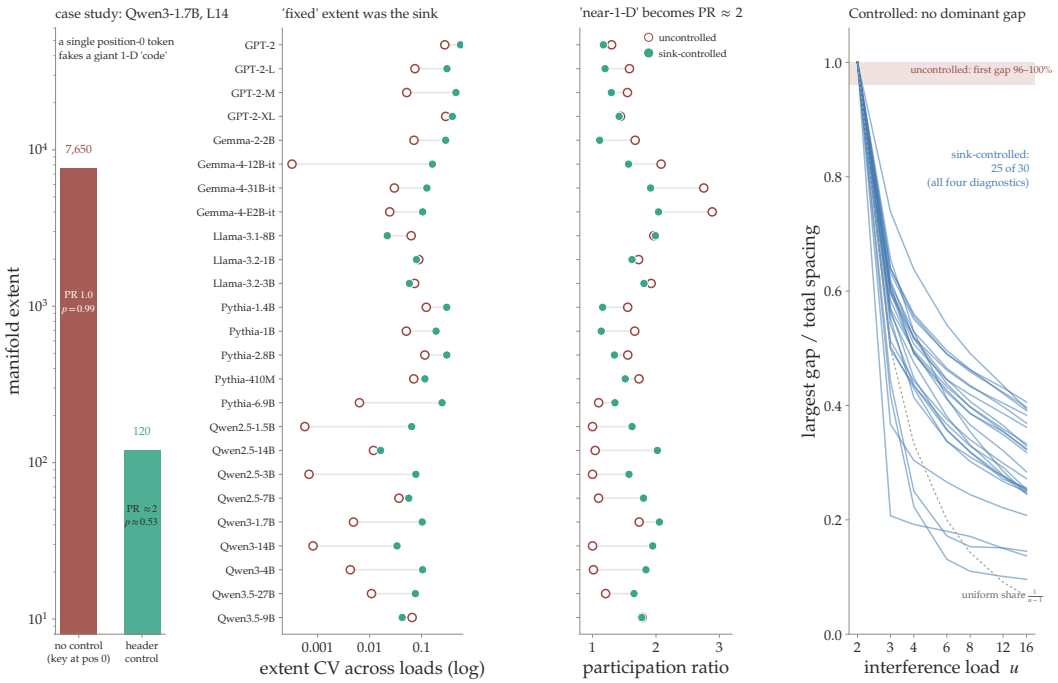


Figure 12: The position-0 sink control (methods). Far left: in Qwen3-1.7B at L14, a single position-0 token inflates the measured manifold extent $\sim 60\times$ ($7650 \to 120$) and fakes a near-1-D, near-perfect power law (PR 1.0, p=0.99); one neutral header line removes it (PR $\approx$2, $p \approx 0.53$). Middle: across the 25 (of 30) models measured both ways, the control corrects the "perfectly fixed" extent (CV) and the "near-1-D" participation ratio. Right: uncontrolled, the first inter-slot gap is 96–100% of the total spacing at every load, so the "spacing law" was one outlier gap; sink-controlled, no single gap dominates.

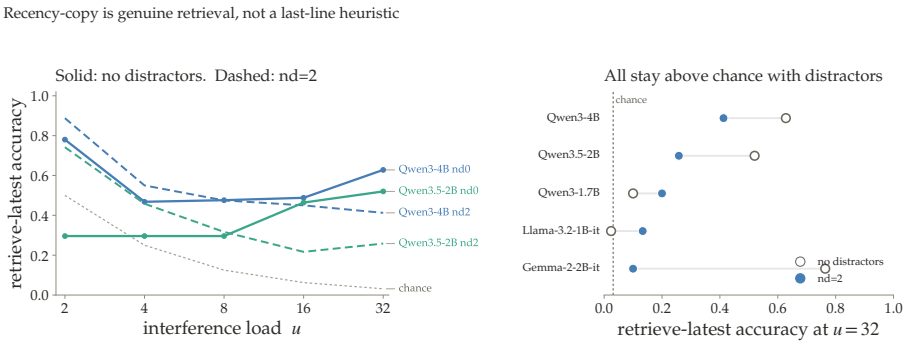
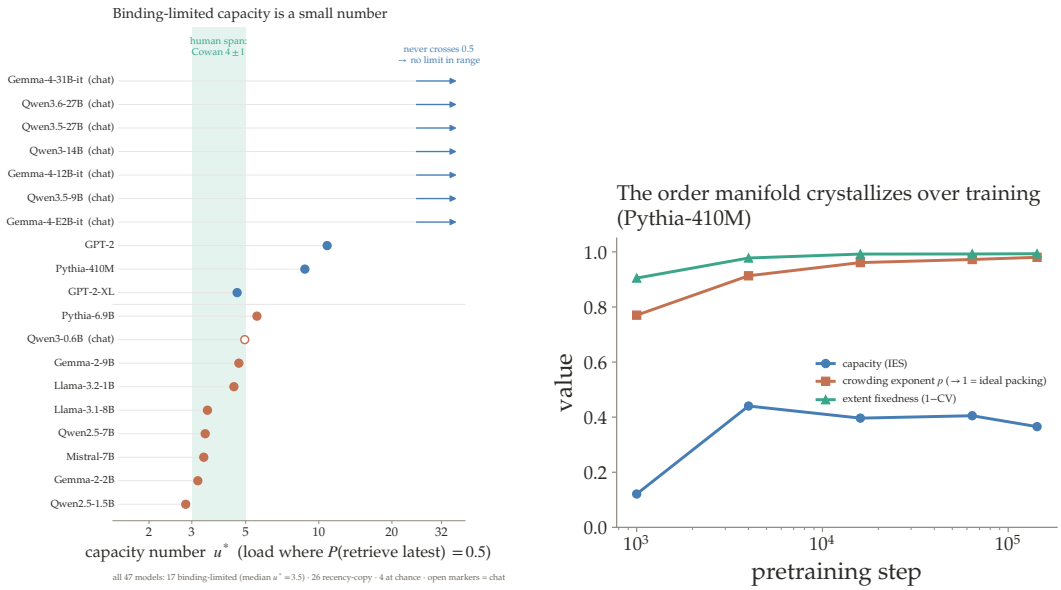


Figure 13: Recency-copy is genuine latest-value retrieval, not a last-line heuristic. Chat capacity re-measured with nd=2 interleaved distractors, so the latest value is not the final line (blind last-line copying would score at chance). Left: for two exemplars the with-distractor curves (dashed) stay above chance at every load. Right: at u=32 all five locally runnable chat recency-copy models remain above chance under distractors (filled) though below their no-distractor value (open); the drop is the final-line convenience, not the regime.

that removes only the queried key's stale occurrences repairs retrieval at negative KV cost (Fig. 20), and we close with five concrete practices.

Mitigation by eviction (an oracle bound). Since the limit is crowding of the competing occurrences, masking attention to the queried key's stale occurrences (keeping only the latest) removes the interference. On Qwen3-1.7B it restores retrieve-latest accuracy from 0.05 to 0.84 at u=32 (and



(a) Capacity number vs. Cowan's 4

(b) Emergence over pretraining

Figure 14: Left: the load u^* at which $P(\text{retrieve latest})$ crosses 0.5 (exemplars shown; all 47 measured models: 17 binding-limited, 26 recency-copy, 4 at chance). Binding-limited models (orange) sit at a median u^*=3.5 (range 2.0–6.0), near the human 4 ± 1 band (shaded); recency-copy models (blue) mostly cross late or never (a few cross early, at the low-load accuracy floor; Fig. 1). Because chance equals 0.5 at u=2, u^* is floor-compressed and, worse, non-monotone in interference robustness: a curve flat below 0.5 (e.g. Qwen3-1.7B chat, 0.44 at u=2) registers as u^*=2 despite being maximally interference-robust. We therefore report u^* for descriptive color only and rest no claim on the tie to Cowan's number. Right: across Pythia-410M checkpoints the crowding geometry organizes early (exponent and extent-fixedness saturate by ∼4k steps) while behavioral capacity (IES) is non-monotone. The substrate is learned early; capacity is not claimed to emerge with it.

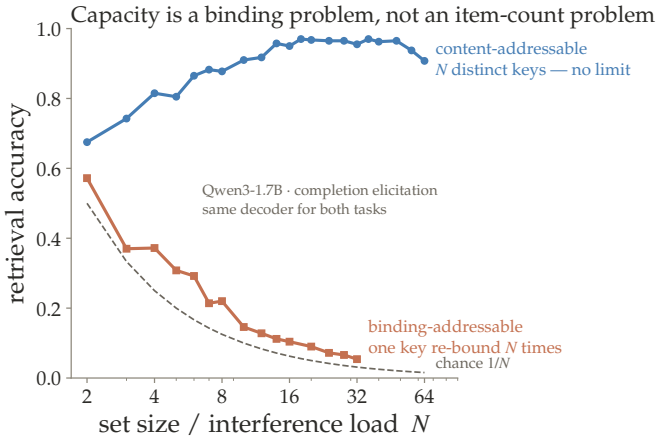


Figure 15: The limit is a binding problem, not an item-count problem. Content-addressable retrieval (distinct keys) shows no capacity limit up to $N = 64$, while binding-addressable retrieval (one key rebound N times) collapses toward chance, with the same model and decoder (Qwen3-1.7B, completion).

to ≥ 0.91 for $u \leq 16$; Fig. 20), with the content-addressable control unharmed, while shrinking the KV footprint. This is the interpretability-side analogue of KV-cache compression (zha; Li et al., 2024), but keyed on binding identity rather than attention mass. It is an oracle demonstration that assumes the stale occurrences of the queried key are already identified, which in real agentic and

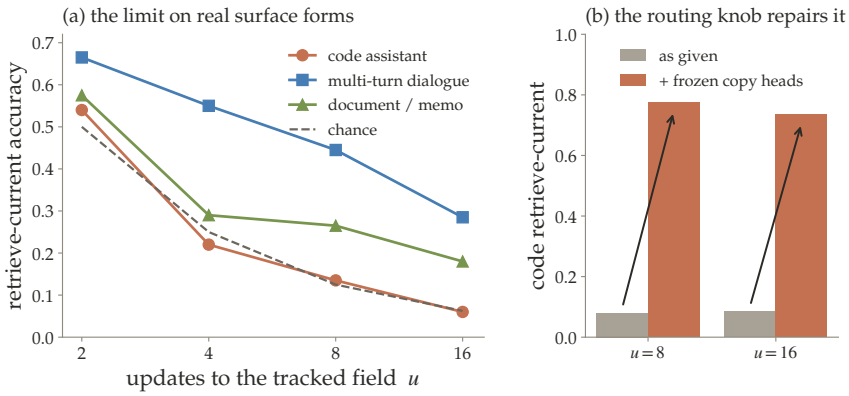


Figure 16: The limit, and the routing repair, on naturalistic surface forms (Qwen3-1.7B, chat elicitation). (a) Retrieve-current accuracy against the number of field updates for a code assistant, a dialogue, and a document stream; all decline. The code form is binding-limited; dialogue and document decline more slowly because their "now"/"today" cue engages the copy route. (b) On the code form, the copy heads frozen from the synthetic task restore retrieval $0.08 \rightarrow 0.78$.

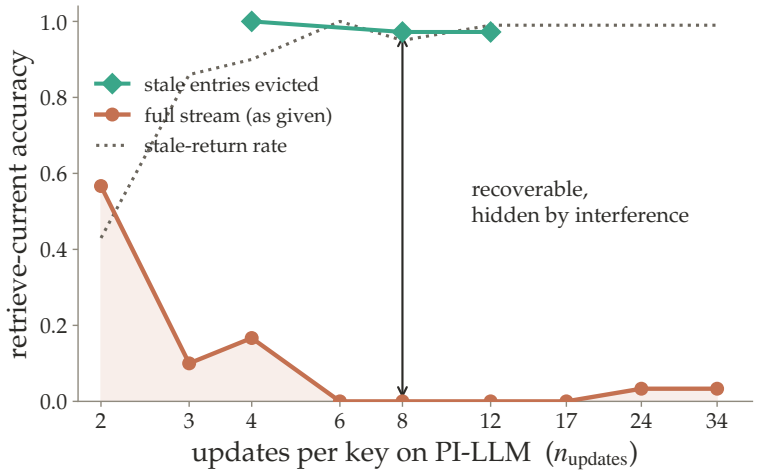


Figure 17: The phenomenon on the real PI-LLM benchmark (Wang & Sun, 2025) (Qwen3-1.7B). On the benchmark's own interference axis n_{updates}, retrieve-current accuracy on the stream as given (clay) collapses to near zero while the stale-return rate (dotted) rises to ≈ 0.99. Deleting only the queried key's stale entries (teal; its latest kept, every other key untouched) restores accuracy to 0.97–1.00. The current value was recoverable throughout, so the failure is interference, not lost information.

RAG text is itself the hard problem (and is what KV-eviction heuristics approximate). As a trigger for when to evict, the model's own output confidence (top-1 minus top-2 option logit) is usable (AUROC up to 0.75); a white-box probe from our RoPE-free Q-space did not beat it (AUROC $\approx$ 0.3–0.6).

Practitioner takeaways. For engineers building agents, RAG, or code assistants, the mechanism implies five concrete practices. (i) Evaluate for rebinding, not just recall, and include non-final probes: a memory eval whose answer is always the last mention is passed by blind copying, so add probes that rebind one key 4–16 times under distractors and cue a non-final value. (ii) Screen before deploying: a $\sim$250-prompt interference test (minutes, API-only) classifies a model as binding-limited, recency-copy, or incapable; binding-limited models begin silently returning stale state at $\approx$2–3 rebindings of one entity. (iii) Canonicalize mutable state upstream: summarizing each entity to its latest binding before the model reads removes the interference and shrinks the KV cache. (iv)

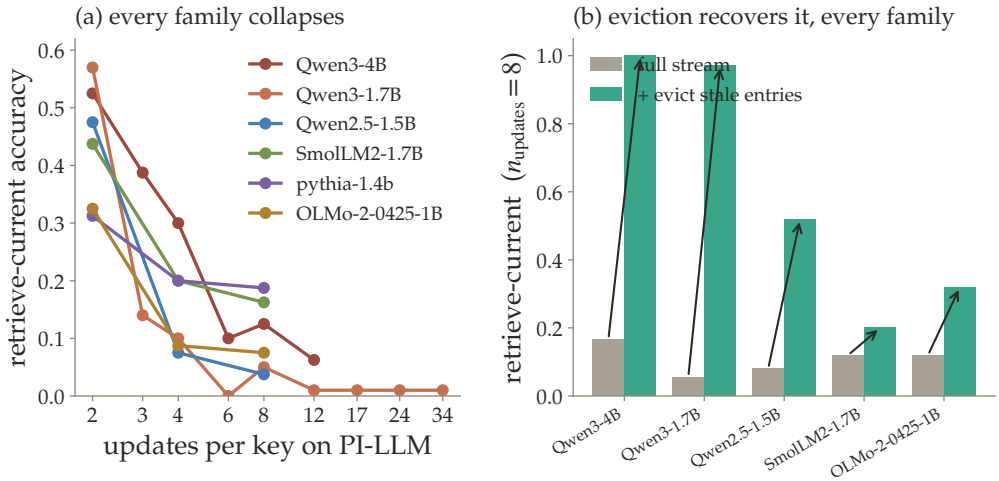


Figure 18: The PI-LLM collapse and eviction-restore generalize across model families (§4.8). (a) Retrieve-current (constrained choice) on the benchmark's update axis collapses in all six models tested (Qwen3-4B, Qwen3-1.7B, Qwen2.5-1.5B, SmolLM2-1.7B, Pythia-1.4B, OLMo-2-1B; base models under raw completion, chat models under their template). (b) Deleting the queried key's stale entries recovers a large fraction of the lost accuracy in every model able to emit the value, most fully in the more capable ones (n_{updates}=8). Pythia-1.4B is omitted from (b) because it cannot emit the value in the requested format even with a single occurrence, a base-model limitation orthogonal to interference.

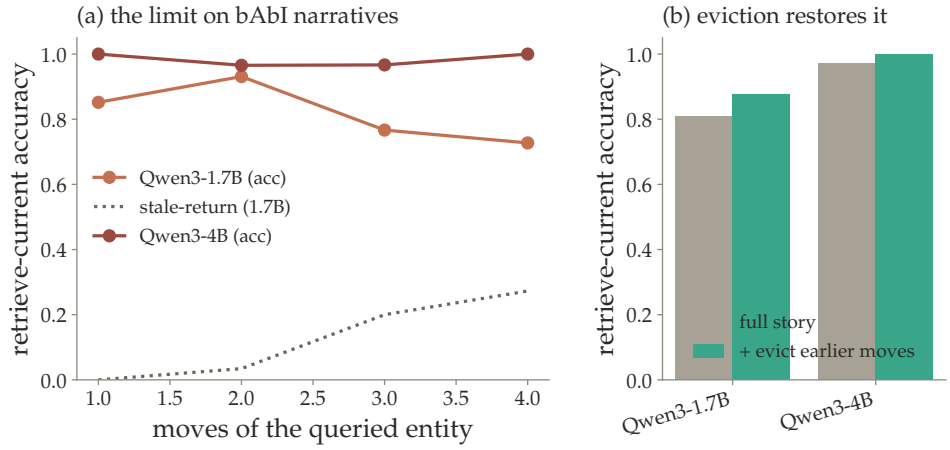


Figure 19: The same retrieve-latest failure on the bAbI benchmark (Weston et al., 2015), a natural-language narrative surface form (§4.8). (a) On task 1, as the queried entity is moved more times, Qwen3-1.7B's retrieve-current accuracy falls ($0.89 \rightarrow 0.73$) and its stale-return rate rises ($0 \rightarrow 0.27$); the stronger Qwen3-4B handles bAbI's milder interference (few moves, short passages). (b) Removing the entity's earlier moves restores accuracy in both. bAbI is easier than PI-LLM because its interference load is small, which is why the effect is milder and appears mainly in the smaller model, consistent with the severity growing with the number of updates.

Sort updates into textual order: no route reads semantic time, so out-of-order logs must be pre-sorted. (v) Do not trust confidence or fluency to catch stale outputs: the retrieval head reads a stale entry cleanly and confidently, so guard with programmatic state checks.

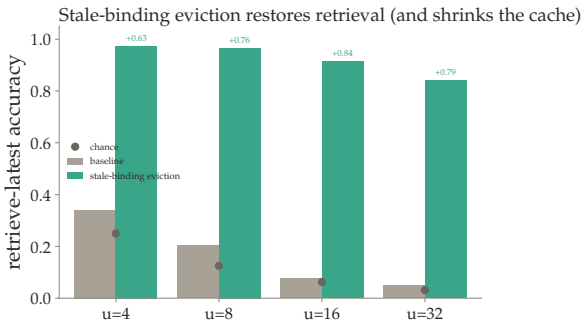


Figure 20: Stale-binding eviction. Masking the queried key's stale occurrences restores retrieve-latest accuracy to near ceiling at every load while reducing the KV cache, a mechanism-grounded, negative-cost mitigation for interference failures.

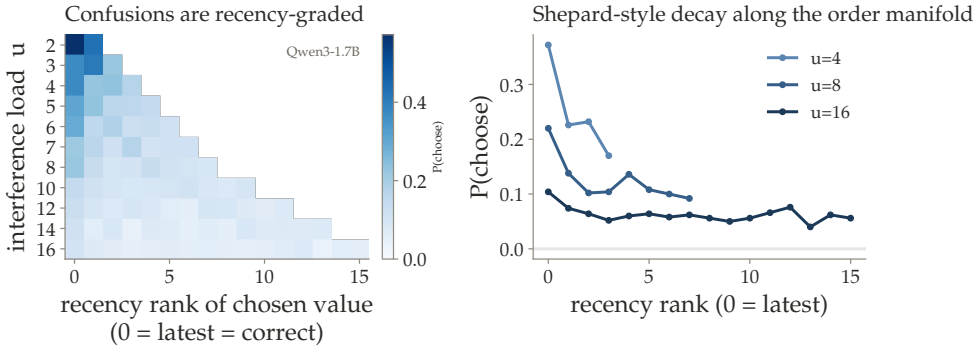


Figure 21: Confusions are order-graded. Left: choice probability by recency rank of the chosen value and load. Right: profiles at three loads, a Shepard decay along the order manifold that flattens as codes crowd.

E.1 Error geometry: a Shepard gradient along the order manifold

If failures are crowding-driven, they should concentrate on order-adjacent competitors. Figure 21 shows exactly this: the probability of reporting slot k decays smoothly with its order-distance from the latest slot, at every load, and flattens toward uniform as load grows. This is Shepard's exponential generalization law (Shepard, 1987) with distance measured along the model's own order manifold, and it is the transformer analogue of transposition gradients in human serial recall and of Oberauer's within-list binding swaps (Oberauer, 2019). The link is quantitative: the fitted decay constant of the Shepard gradient at each load tracks the measured order-code spacing $\Delta(u)$ (Fig. 22), so the behavioral confusion gradient and the stored geometry are one ruler. The error structure is read off the same crowding the geometry measures.

E.2 A load-invariant primacy anchor (within-load, confound-free)

Because the accuracy-versus-load curve is entangled with the $1/u$ chance trend, we test the geometry with a within-load probe that the trend cannot explain. At a fixed load we cue the model to recall a specific serial position of the tracked key's history, the first (primacy), the last (recency), or the middle value, and compare accuracies at that same u (Fig. 23). Crucially, cueing a non-final position (first or middle) is a task the induction/copy route cannot solve, since copying the most recent occurrence is wrong, so this probe isolates order-based reading. Recall of the first value is load-invariant at ≈ 0.67 from $u{=}4$ to $u{=}12$ (far above chance 0.08–0.25): the first binding is an anchor that does not crowd. Recall of the last value instead degrades with load ($0.45 \rightarrow 0.13$), and the middle sits at chance throughout, the classic primacy-plus-recency serial-position curve (Guo & Vosoughi, 2024; Wang & Sun, 2025), here primacy-dominant. The mechanism is not mere separability: a naive "endpoints are more separable" geometric argument predicts that the first and the last should both

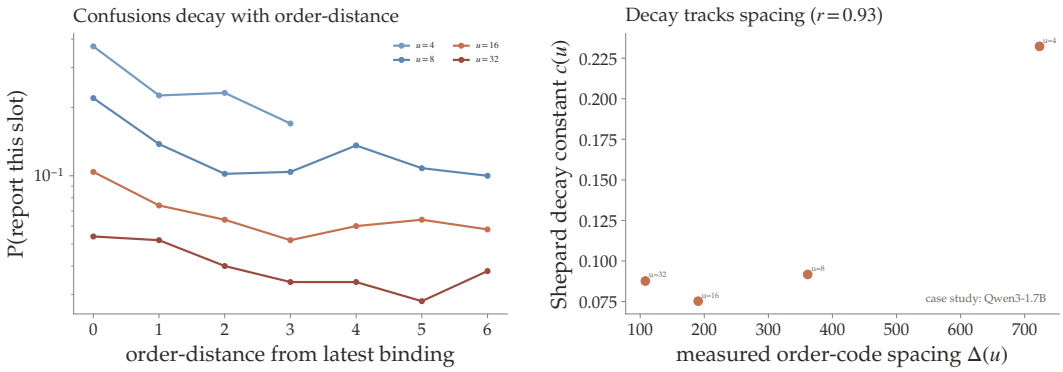


Figure 22: Behavior and geometry share one ruler. Left: the probability of reporting a stale slot falls off exponentially with its order-distance from the latest (Shepard's law, measured inside the model), flattening with load. Right: the fitted behavioral Shepard decay constant $c(u)$ (the rate of $\eta_k = \exp(-c\,\delta_k)$) against the measured median order-code spacing $\Delta(u)$. The confusion gradient's length scale is set by the stored geometry.

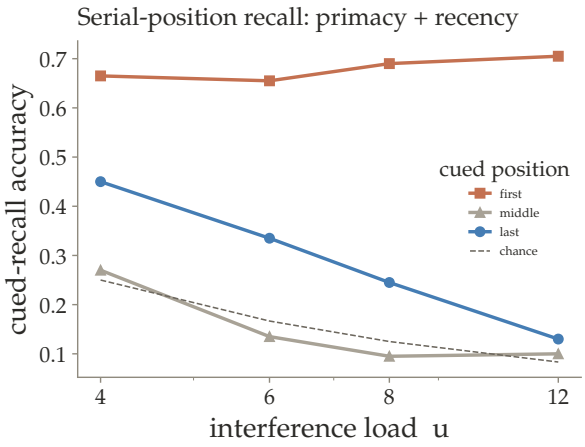
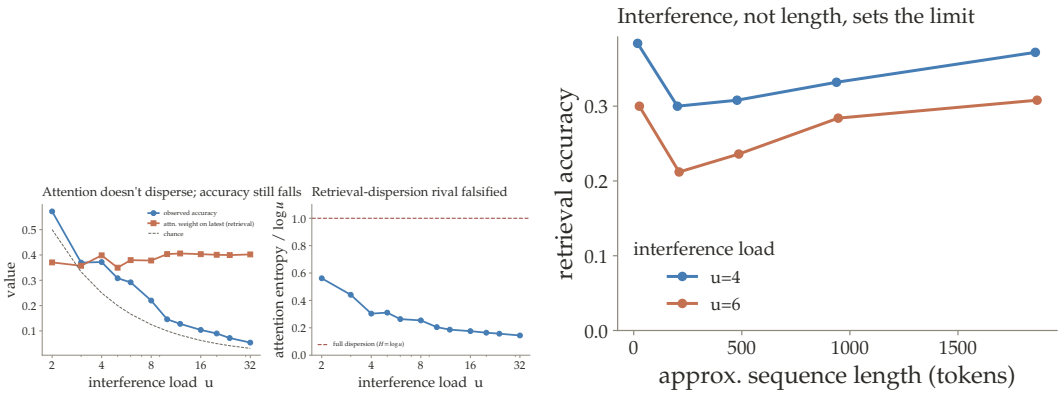


Figure 23: A load-invariant primacy anchor. Cued recall of the first value is flat in load (≈ 0.67); recency degrades with load; the middle sits at chance. Compared at fixed u, so the ordering (first $\gg$ last $\gg$ middle) is unconfounded by the $1/u$ trend. Cueing a non-final position is unsolvable by copying, so this isolates order-based reading; the first-over-last asymmetry implicates an attention-sink component beyond pure order-code separability.

be robust, yet only the first is load-invariant. The asymmetry indicates that the primacy anchor is not pure order-code separability but is reinforced by a first-token/attention-sink effect (Xiao et al., 2024); a control that decouples update-history position from absolute context position is needed to fully attribute it. What the probe does establish, confound-free, is that order-based recall of a non-final binding is possible and structured. The retrieval can address the code by order, not only copy its end.

E.3 Crowding, not attention dispersion: adjudicating the rival

The retrieval-limited rival predicts that the failure is softmax dispersion: the retrieval head's attention spreads toward uniform as the number of competing occurrences grows, with entropy $\Theta(\log u)$ (Veličković et al., 2025; Vasylenko et al., 2026), a mechanism Gong & Zhang (2024) argue directly sets the transformer's working-memory-capacity limit. We test their hypothesis on the identified retrieval head with its own metric (Fig. 24). It is false here: the head's attention does not disperse. Its Shannon entropy over the u (renormalized) occurrence weights, normalized by $\log u$ (the growth rate the dispersion rival predicts), falls from 0.56 at $u{=}2$ to 0.14 at $u{=}32$, so entropy grows strictly



(a) Attention does not disperse

(b) Length is not the driver

Figure 24: Crowding, not attention dispersion (a within-route question). Given a model reading through the order route, its failure is code crowding, not softmax dispersion: (a) the retrieval head's attention entropy (over the u occurrences, normalized by $\log u$, averaged over prompts) falls while its weight on the latest occurrence stays flat and accuracy collapses, so the dispersion rival is falsified; (b) at fixed load, sequence length barely changes accuracy.

slower than $\log u$, while the weight the head places on the latest occurrence stays roughly constant (≈ 0.39) rather than decaying as $1/u$. (Both are per-prompt averages; the per-distribution bound $H \geq H_b(p_{\text{latest}})$ does not constrain these averages, because prompts on which the head locks onto a single occurrence contribute near-zero entropy, pulling the mean entropy below H_b of the mean weight.) Yet accuracy collapses from 0.57 to 0.05. Attention delivers a stable, concentrated read; what degrades is the value it reads, since the order codes it superposes crowd with load. (This is a correlational match to the crowding geometry, on one head. We do not claim it is the causal bottleneck: the sink-controlled compression test of that geometry is null, §E.6, and whether crowding bottlenecks behavior is itself route-dependent.) A context-length control confirms the same conclusion from the behavioral side: padding the prompt with neutral filler at fixed u barely moves accuracy, whereas raising u collapses it (Fig. 24b). The bottleneck is occurrence-code crowding, not retrieval dispersion; the retrieval's only role is a fixed, concentrated window that caps the causal upside.

E.4 Order, not absolute position

A natural worry is that the "order code" is really a generic absolute- or RoPE-position code (update k always sits near token $4k$). We separate the two by shifting the entire update block to widely varying absolute positions (inserting 8–96 neutral filler lines before it, 70–533 tokens) while its ordinal structure is fixed. The order-code spacing is invariant to this $\sim 7\times$ shift (coefficient of variation 3.5%; Fig. 36), and likewise invariant to varying the token gap between updates (CV 15%). An absolute-position code's geometry would move with position; this one does not, so the code is ordinal. *Limitation.* We could not turn this into a clean behavioral position-transfer. Inserting filler flips the model into a primacy-dominated read (latest-retrieval falls below chance, consistent with the primacy anchor of §E.2), so retrieval accuracy is at floor in the shifted conditions and the causal-transfer arm is uninformative. We therefore establish position-invariance of the geometry (Fig. 36, App. L), and leave a filler-robust behavioral decoupling to future work.

What the code indexes: occurrence order, not semantic time. Prior work names the substrate an "ordinal" code (Dai et al., 2024; Bassi & Tomar, 2026) without asking ordinal with respect to what. We answer this directly, and the answer identifies the object rather than limiting it. We decouple occurrence order from semantic order with explicit timestamps whose numeric order is shuffled relative to the textual line order ("At time 7, Alice became blue. At time 2, Alice became red. ... Alice's value at the latest time was:"), so the maximum-timestamp value is not the last line. The result generalizes across all 30 models we test on this task (eleven families, 70M–32B; Fig. 25). *(i) Geometry:* averaging the key-token residual by textual position yields a clean, high-extent manifold,

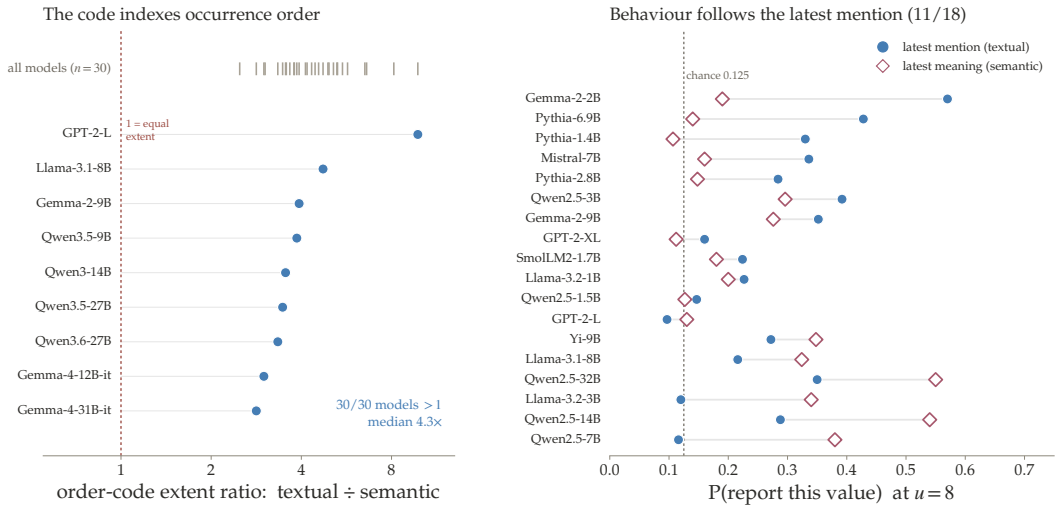


Figure 25: The stored code indexes occurrence order, not semantic recency. Left: the occurrence-order (textual) manifold is larger in extent than the semantic-order manifold in all $30/30$ models measured (median $4.3\times$, range 2.5–$9.8\times$). Right: at $u{=}8$, behavior places more mass on the latest mention than on the latest meaning in $11/18$ base models; the exceptions (largest Qwen2.5/Llama models, Yi-9B) show an emerging above-chance semantic-time read.

while averaging the same residuals by semantic (timestamp) rank collapses it. The occurrence-order manifold is 2.5–$9.8\times$ larger in extent than the semantic-order manifold in every one of the 30 models (median $4.3\times$; e.g. 4.3:1 Qwen3-1.7B, 4.7:1 Llama-3.1-8B, 9.8:1 GPT-2-large). *(ii) Behavior:* at a load every model can still serve ($u{=}4$), all 18 base models retrieve the textual-latest value above chance (median 0.46 versus chance 0.25), and the mass placed on the textual-latest exceeds that on the semantic-latest in $13/18$ at $u{=}4$ and $11/18$ at $u{=}8$ (Fig. 25). Behavior follows occurrence order, as the geometry predicts. The exceptions are informative and we show them: the largest base models tested (Qwen2.5-7B/14B, Llama-3.1-8B, Llama-3.2-3B, Yi-9B) place more mass on the semantic-latest value (up to 0.54, Qwen2.5-14B), so an above-chance semantic-time read does emerge with scale, even while their stored code remains occurrence-ordered (their extent ratios stay 3.0–$4.7\times$). The conserved substrate is therefore a surface-order ledger of the key's mentions: abstract enough to be invariant to absolute position and token gap (above), yet syntactic, not semantic. The falsifiable consequence follows and the same data confirm it in most models. A task whose answer requires integrating updates by semantic order rather than presentation order (out-of-order logs, corrections, non-chronological retrieval) fails in the majority of models at every load, because their routes read the mention ledger, not semantic time; where a partial semantic read exists (the largest models above), it competes with, rather than replaces, the occurrence-order read.[1] This sharpens, rather than weakens, the causal results, which act on exactly this occurrence-order code: it is what crowds, and what the two routes read differently.

What the code is specific to (sink-controlled). A deflationary reading is that we have merely rediscovered "transformers index repeated tokens." An earlier version of this test, run without the position-0 control, appeared to confirm the deflation dramatically: binding-free filler repetition seemed to occupy the rebound key's subspace with variance captured ≈ 1.0. Those numbers were the sink artifact of §D (the "shared subspace" was the sink direction), and we retract them. Re-run sink-controlled (a neutral header line; six occurrence conditions from rebinding through key-randomized and binding-free filler repetition; Qwen3-1.7B and gemma-4-E2B-it), the honest picture is graded. Every condition that repeats the token does induce an occurrence manifold of similar modest ex-

[1] We also attempted a causal double dissociation on these timestamp prompts (collapse the textual- versus semantic-order subspace and predict the behavior each drives); it is confounded and we do not rely on it. At nd=0 the textual-latest value is the last line, so a recency read, not the order code, can serve it, and collapsing the order subspace leaves behavior intact. The identification above rests on the geometry (30 models) and the behavioral occurrence-versus-semantic dissociation, not on this null.

tent (8–19× the label-shuffled floor; PR 1.3–2.4), so an occurrence code is present whenever the token recurs, bound or not. But the subspaces only partially coincide: generic repetition captures 34–38% of the rebound key's code variance in Qwen3-1.7B (leaving roughly two-thirds binding-associated) and 72–80% in gemma-4-E2B-it, with the shared fraction varying by layer (0.17–0.70). A causal split of the binding-specific from the shared component is inconclusive at nd=0 (collapsing either costs ≤ 0.13 of $P(\text{latest})$, consistent with the compression null of §D). We therefore reserve "binding-limited" for the behavioral regime, describe the substrate conservatively as an occurrence-order code of the queried key that is partially specialized by binding, and rest no causal claim on this decomposition.

E.5 Capacity is similarity-dependent

The geometry predicts that crowding the content of the competing values should lower capacity through the leakage channel. Holding load fixed, we draw the tracked key's values either from one semantic category (HIGH-similarity, crowded content codes) or across categories (LOW-similarity). Figure 37 shows that HIGH-similarity values yield lower accuracy at every matched load, a leftward capacity shift of $+0.09$ to $+0.16$, and we verify the manipulation directly: the HIGH-sim values have higher in-context representational similarity ($\cos = 0.85$) than the LOW-sim values ($\cos = 0.79$). Capacity is not a fixed number of items; it is set by how crowded the relevant codes are, as the geometry predicts.

E.6 What did not survive our controls (reported negatives)

Repairing retrieval does not repair the downstream decision (partial null). On the RAG-style decision task of §4.8, oracle eviction restores value retrieval to 1.00, yet decision accuracy recovers only partially (0.50–0.68), and the copy-head gain does not repair the decision at all (0.46 at u=16). The decision evidently consumes state through a path our retrieval-level interventions do not reach, which is why the main text claims failure propagation, not that the routing knob fixes downstream behavior.

What survives all of this. The paper's causal core is untouched by every item below: the routing interventions (§4.7) are sink-free by construction, edit only the query token, and none of the negatives here bears on them. We collect the failed results in one place, each as claim, control, verdict, and surviving scope, so the surviving claims are read against them.

(i) The predictive capacity law. Claim: each binding-limited model's u^* is predictable from its sink-free crowding coefficients and measured retrieval noise (Prop. 2(i)). Control: out-of-sample test with acceptance thresholds fixed before the run, n=5 models with complete inputs. Verdict: not supported (Spearman 0.56, permutation p=0.21, fitted SNR gain κ=15.6 outside its plausible range); with n=5 this test is underpowered and uninformative rather than a positive refutation, which is why we present the point law only in bound form. Surviving scope: the proposition organizes measured quantities, and its qualitative corollary holds, since all $8/8$ recency-copy models sit above their geometry-implied capacity (sign test p=0.0039).

(ii) Causal necessity of the geometry via compression. Claim: compressing the order-code subspace should abolish retrieval. Control: the position-0 activation-sink control (a neutral header line). Verdict: the apparent layer-localized effect was a sink artifact; sink-free, compression costs $\lesssim 0.05$ accuracy at every layer (§D, Fig. 12). Surviving scope: the geometry's role rests on structural and behavioral evidence, not compression; the routing interventions are sink-free by construction.

(iii) Slot-code swap and the RoPE-aware order intervention. Claim: the occurrence-order geometry is the causal address that random-access retrieval reads, so ablating the heads that read a cued past slot, or moving an occurrence's positional code, should induce the matching stale-slot error. Control: (a) a random-subspace slot swap; and (b), on a model that genuinely random-accesses (gemma-4-31B, cued-middle 0.86), we screened the heads that most attend to the cued (non-final) slot (attention 0.45–0.71, cleanly disjoint from the copy heads), ablated them at the query token, and separately applied a RoPE-aware key-side positional swap (moving the target occurrence's key to slot j's position; verified on a purely positional query to re-address the retrieved slot exactly). Verdict: null throughout. The residual-space swap has no effect ($\Delta P(\text{target}) = -0.02$); ablating the order heads changes cued-middle accuracy by 0.00 (even the head attending 0.71 to the cued slot is

causally unnecessary); and swapping an occurrence's key position does not move the retrieved slot (slope ≈ 0 on both gemma-4-31B and gemma-4-12B). Surviving scope: random access in capable models is distributed and robust, not a localizable positional address. We therefore use the order-code geometry descriptively and explicitly do not claim it is the causal substrate a route reads. This is the RoPE-aware test we had flagged as the next step, now run and reported.

(iv) The white-box screen as a staleness predictor. Claim: a synthetic interference screen predicts naturalistic stale-output rates better than a task-external induction-head inventory. Control: four naturalistic target models. Verdict: uninformative, because the target set is degenerate (all four are majority-stale at every load). Surviving scope: we claim nothing from it.

(v) The knob does not flip downstream decisions. Claim: repairing retrieval repairs the RAG-style decision built on it. Control: oracle eviction (retrieval $\rightarrow 1.00$) and the copy-head gain. Verdict: decision accuracy recovers only partially (0.50–0.68 versus a 0.97 ceiling; gain 0.46). Surviving scope: we claim failure propagation only (§4.8).

(vi) Behavioral position-transfer and template transfer. Claim: the geometry supports a clean behavioral position-transfer, and the fitted retrieval parameters transfer across surface templates. Control: filler-shifted prompts; three surface templates. Verdict: filler flips the model to a primacy-dominated read (position-transfer uninformative, §E.4), and the fitted parameters do not transfer (§4.4). Surviving scope: the crowding phenomenon is position-invariant and template-robust even though its retrieval parameterization is not.

F Retrieval-head identification and the routing dissociation

The routing dissociation (§4.4). For each base model we fit the geometry-only retrieval to its low-load behavior and compare the interference endurance it implies to the observed value (Fig. 27; the 14 base models with geometric fits). Out of sample the fit has only marginal skill, so this is a dissociation, not a law. All $8/8$ recency-copy models sit above the geometry-implied capacity (sign test p=0.0039, median $\approx 2.4\times$), reading through a copy route the geometry does not describe, while incapable and weak-copy models sit on the identity line. We call this route-gated capacity: a conserved representation bounds behavior only where retrieval routes through it, the capacity analogue of decodable-versus-used (Elazar et al., 2021), except that here the gating variable is causally switchable in both directions (§4.7).

Head identity. Fig. 26 shows the causal retrieval-head localization and the Q-space prediction referenced in §4.2. A caveat on head identity: the top heads are selected as the maximum ablation drop over ~ 450 heads at a single screening seed, so the individual drop magnitudes (top three $0.125/0.108/0.100$) are winner's-cursed and mutually within noise. The specific label L17H6 should be read as "a top order-reading head," not a unique one. The claims that do not depend on the exact head identity, the order-reader-versus-copy distinction and the bidirectional causal conversion (§4.7), are established on fresh, disjoint evaluation seeds and do not inherit this selection noise.

G Regime is a property of the (model, interface) pair

The regime labels of §4.4 attach to a model under an elicitation interface, not to weights alone. The sharpest case is our own primary model. Under raw completion Qwen3-1.7B is binding-limited (decay to chance by u=32, the curve of Fig. 5b at g=1), while under its chat template the same weights read recency-copy, with accuracy $0.44, 0.54, 0.53, 0.35, 0.33, 0.27, 0.23, 0.19, 0.10$ at u=2, 3, 4, 6, 8, 12, 16, 24, 32, i.e. 0.07 above chance at u=32, just over the 0.05 regime boundary (boundary-adjacent, and binding-limited under a 0.08 threshold). All mechanism experiments (conversion, moderator, ablations) are therefore within-interface contrasts: baseline and intervened runs share the same completion prompts, so any harness artifact is common to both arms, and the completion "limit" of a reasoning model is itself an instance of the interface effect we quantify in §4.5. Regime counts are robust to the boundary threshold: at decay thresholds $0.03/0.05/0.08$ the (binding-limited, recency-copy, incapable) counts are $(17, 26, 4)$, $(17, 26, 4)$, and $(23, 20, 4)$; the six models that flip at 0.08 are the boundary-adjacent recency-copy ones (GPT-2-M/L/XL, Pythia-1.4B/2.8B, and chat-elicited Qwen3-1.7B). The four figures below give the interface evidence in

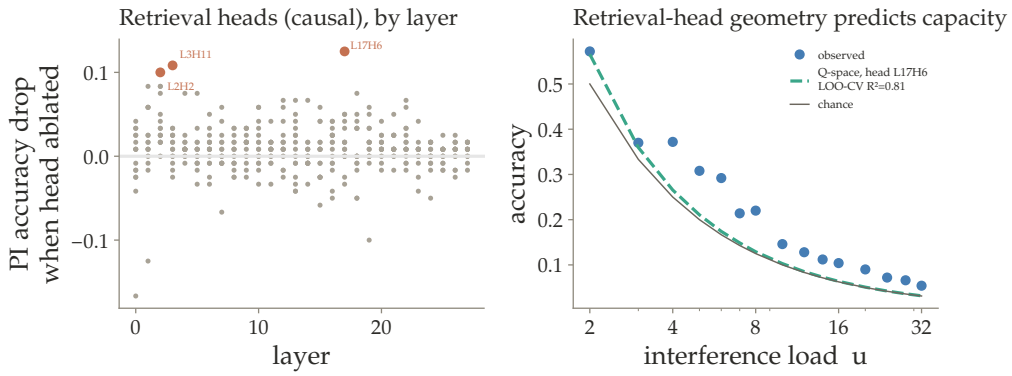


Figure 26: Left: PI accuracy drop when each attention head is ablated; the top causal heads (orange) sit at early and intermediate layers, not the output layer. Right: the top head's Q-space order-code geometry predicts the held-out choice distribution beyond chance at resolvable loads; see §4.2 for the chance-referenced evaluation.

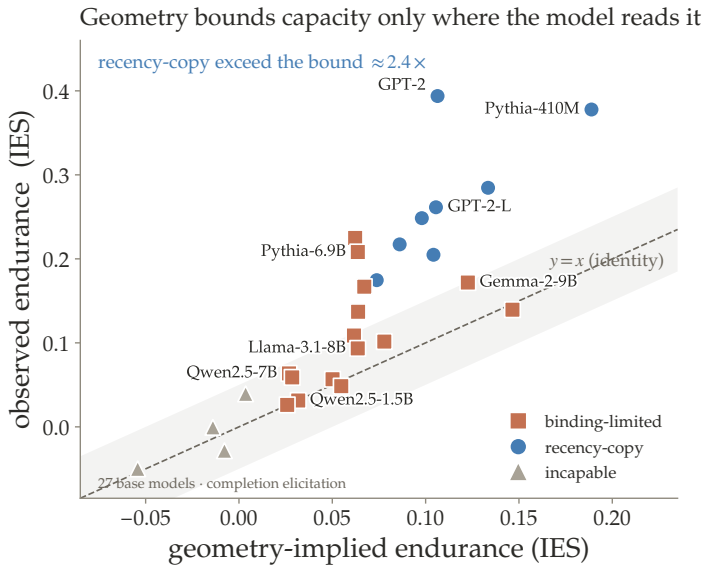


Figure 27: The routing dissociation (14 base models with geometric fits, completion elicitation; chat-elicited models excluded, §4.4). Observed interference endurance (IES, the above-chance area under the accuracy-versus-load curve) against the endurance implied by the order-code geometry (a similarity retrieval fit to each model's low-load behavior and extrapolated; the x-axis is a fit reference, not an out-of-sample prediction, and its held-out skill is marginal, so we report no R^2). Incapable models (grey) and weakly convertible binding-limited models (orange) lie near the identity line; every recency-copy model (blue) sits above it ($8/8$, sign test p=0.0039; median excess $\approx 2.4\times$), because at high load it copies the latest occurrence, a route the geometry does not describe. The exceedance tracks latent copy-route strength, so a few strongly convertible base models (e.g. Pythia-6.9B, Llama-3.1-8B) also rise above the line. A dissociation, not a scaling law; the causal support is §4.7.

full: the scale trend within versus across eras (Fig. 29), the matched-backbone base/instruct by completion/chat decomposition (Fig. 30), the raw-versus-fair evals artifact (Fig. 31), and the elicitation-fairness check itself (Fig. 32).

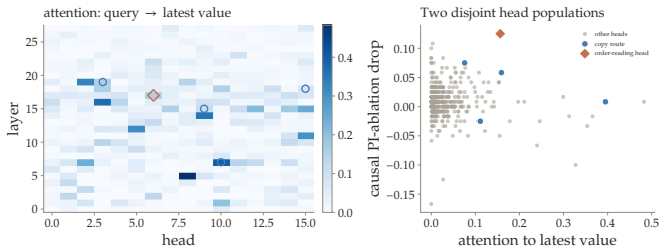


Figure 28: Where the copy route lives. Left: attention from the final query to the latest value token, per (layer, head) in Qwen3-1.7B at u=32; a compact set of heads is concentrated on the latest value (the amplified set L19H3, L18H15, L7H10, L15H9, blue rings). Right, the informative view: causal PI-ablation drop against attention-to-latest. The order-reading retrieval head L17H6 (orange diamond) has the largest causal impact on the task but only moderate attention-to-latest, whereas the heads with the highest attention-to-latest have near-zero causal drop (a redundant copy route). The two roles separate on causal importance, not on attention magnitude alone.

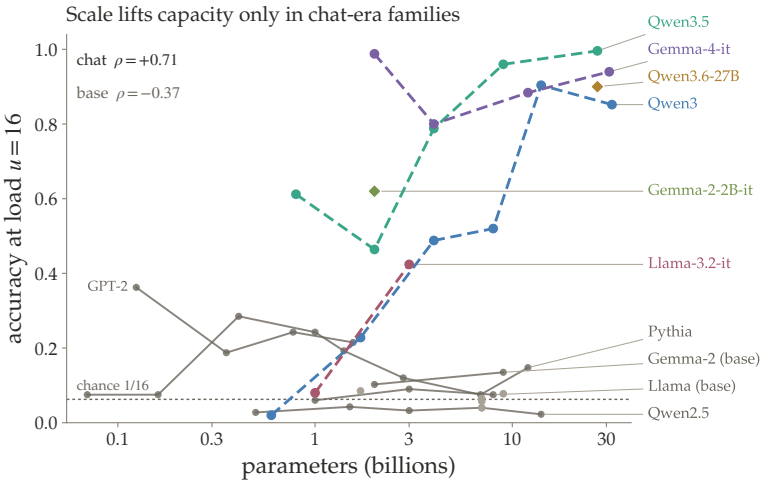


Figure 29: Scale escapes the interference limit only in chat-era families, and only through the chat interface (§4.5). High-interference accuracy (u=16) against parameters. Inside reasoning-era families (colored, chat-elicited) capacity climbs steeply with scale (pooled ρ=+0.71); base-era families (grey) stay near chance at every size (pooled ρ=−0.37). Scale buys capacity only where the copy route is engaged.

H SCOPE OF EVIDENCE: WHICH N SUPPORTS WHICH CLAIM

Different claims rest on different model sets, and "every model" in the main text always refers to the N of the claim at hand. The behavioral regime map spans 47 models (27 completion-elicited base plus 20 chat-elicited; 70M–32B, 11 families); the sink-controlled crowding geometry is measured on 30 of them, and the occurrence-versus-semantic geometry contrast on 30. The routing dissociation uses the 14 base models with geometric-Luce fits. The scale and era-by-interface analysis uses 46 models with known parameter counts (19 chat, 27 base), with the weights-by-interface decomposition on 3 backbones ($\times$4 cells each), and the evals-artifact comparison on the 19 chat models measured under both harnesses. The conversion moderator spans 25 models; the fully controlled conversion is a single model per direction (Qwen3-1.7B constructive, GPT-2-large ablation) with 25 directional replications; the elicitation ladder uses 2 models; and the naturalistic code, eviction, and RAG-decision results are on Qwen3-1.7B. The predictive-law test is on 5 binding-limited models with complete sink-free inputs, plus 8 recency-copy exceedance checks.

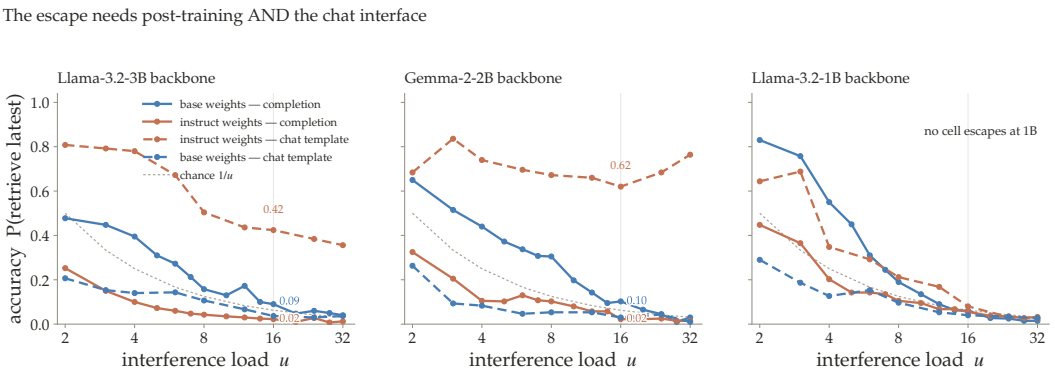


Figure 30: Only instruct weights under the chat interface escape; neither instruction tuning nor the template alone suffices. Matched-backbone cells at u=16 (base-completion / instruct-completion / instruct-chat / base-chat-template): Llama-3.2-3B 0.09/0.02/0.42/0.04; gemma-2-2B 0.10/0.02/0.62/0.03; Llama-3.2-1B 0.06/0.06/0.08/0.04 (no cell escapes at 1B). The base weights under the instruct chat template stay at chance, so the template's token structure alone does not enable the escape. n=3 backbones, every complete set in the campaign, an existence-level decomposition.

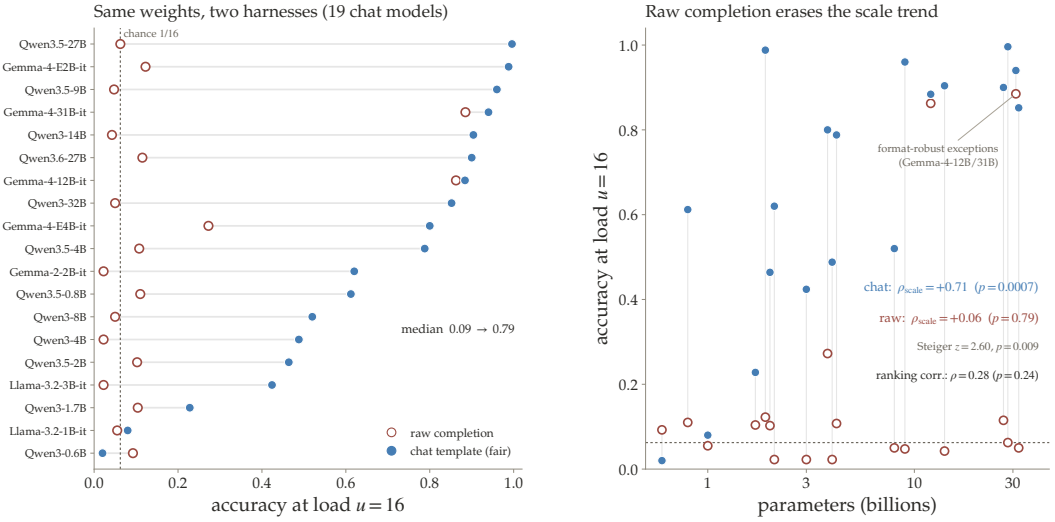


Figure 31: Raw-completion interference benchmarks mismeasure chat models (19 models measured under both harnesses). Left: same weights, two harnesses, median u=16 accuracy 0.09 raw versus 0.79 chat; the raw ranking is uncorrelated with the fair one (ρ=0.28, p=0.24). Right: the raw harness erases the scale trend (ρ=+0.06 versus ρ=+0.71; Steiger z=2.60, p=0.009); 9/19 models score below chance at u=2 raw. Format-robust exceptions: gemma-4-12B/31B.

I PER-FAMILY RESULTS

Table 3 lists every model; here we read it off family by family, under each family's fair elicitation. The recurring pattern: base families measured under raw completion are binding-limited or (at small scale) incapable, and their latent copy route can be amplified into a conversion; chat and reasoning families measured under their chat template are recency-copy and climb toward ceiling with scale; and every family, base or chat, writes the crowded occurrence-order code (textual extent far above semantic, ratios 2.5–9.8×).

Base families (completion is fair; accuracies quoted at u=32). *GPT-2* (4 models, 0.12–1.5B) are all recency-copy, capable at low load and declining slowly, and GPT-2-large is the strongest

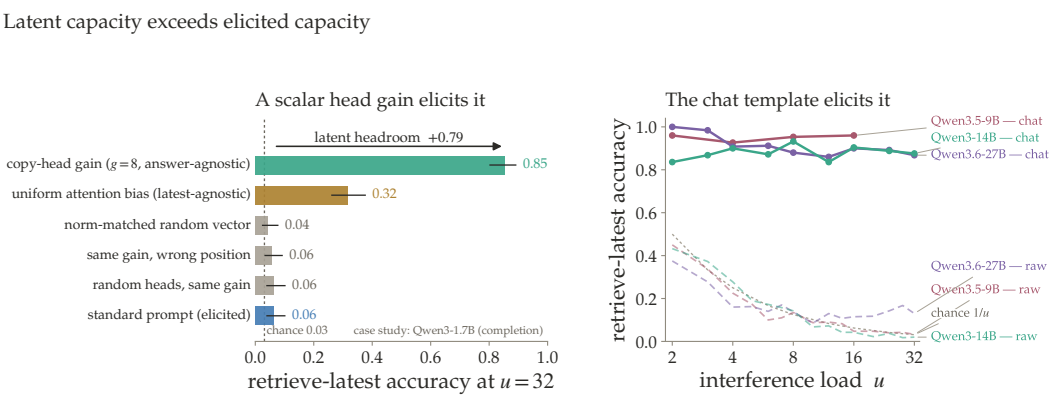


Figure 32: Latent capacity exceeds elicited capacity. (a) Qwen3-1.7B, u=32, completion elicitation (n=250, Wilson 95% CIs): the standard prompt scores 0.06; the answer-agnostic copy-head gain elicits 0.85 and the latest-agnostic uniform attention bias 0.32, while norm-matched, wrong-position, and random-head controls stay at baseline. (b) The other elicitation knob is the prompt format: the same weights that decay to chance under raw completion (dashed) retrieve at 0.8–0.96 across loads under their chat template (solid; Qwen3.5-9B, Qwen3-14B), the harness artifact that motivates chat-elicited measurement of chat and reasoning models (§4.4).

converter of the four (Δ=+0.48). *Pythia* (8, 70M–12B) is the cleanest scale sweep: 70M/160M are incapable, 410M–2.8B recency-copy, and 6.9B/12B binding-limited, with among the largest conversions in the paper (Pythia-410M Δ=+0.68, Pythia-6.9B +0.49). Bigger base models are not more capable on this task, only convertible, and convertibility does not itself grow with scale (410M exceeds 6.9B). *Qwen2.5* (5, 0.5–14B) are binding-limited throughout under raw completion (14B sits at chance, i.e. incapable, a raw-format-parsing failure rather than a capacity one), and convert Δ=+0.08–0.25. *Mistral-7B* (Δ=+0.21), *Falcon3-7B* (+0.23), and *SmolLM2-1.7B* (+0.67) are binding-limited and convertible; *OLMo-2-7B* is binding-limited and the one clean non-converter (Δ=+0.00; its copy route is not wired to the retrieval); *Yi-9B* is at chance under raw completion.

Chat and reasoning families (the chat template is fair). *Qwen3* (6, 0.6–32B): the 0.6B is binding-limited and the rest are recency-copy, climbing 0.10 $\rightarrow$ 0.88 (u=32) from 1.7B to 14B; Qwen3-1.7B is the completion-elicited conversion exemplar of §4.7 (0.06 $\rightarrow$ 0.85). *Qwen3.5* (5, 0.8–27B) are recency-copy at every size and show the sharpest within-family scaling (0.53 $\rightarrow$ 1.00 at u=32, 0.8B$\rightarrow$27B); *Qwen3.6-27B* matches (0.87). *Gemma* (7): the two gemma-2 base models are binding-limited (gemma-2-9B converts +0.20), but the gemma-2-2B instruct model is already recency-copy (0.76), and every gemma-4 instruct model (E2B/E4B/12B/31B) is recency-copy and near ceiling (0.80–0.97); gemma-4-31B is the one model that also random-accesses a non-final slot (§4.5). *Llama* (5): the base models are binding-limited and convert strongly (Llama-3.1-8B Δ=+0.57, the fair-interface conversion of §4.7), while the instruct models are weaker recency-copy (Llama-3.2-3B-it 0.36) or, at 1B, still binding-limited even under chat, one of the two chat-elicited binding-limited models (with Qwen3-0.6B) that show the limit is not a raw-completion artifact. *Phi-3.5-mini* (3.8B) is a mid recency-copy (0.36).

J FULL RESULTS TABLE

Table 3 reports every model under its fair elicitation: parameter count, the elicitation harness (chat template for chat, instruct, and reasoning models; raw completion for base models), the retrieval regime, the capacity number u^* ("–" when accuracy never falls to 0.5 over our range, i.e. no limit), and the sink-controlled crowding exponent p and participation ratio PR.

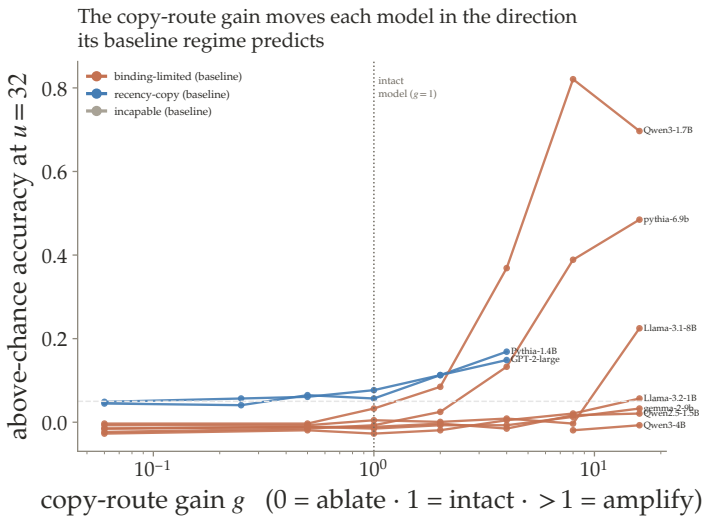


Figure 33: The copy-route gain moves each model in the direction its baseline regime predicts (a complete crossing only in Qwen3-1.7B). Above-chance accuracy at u=32 against gain g (log axis; 0=ablate, 1=intact, >1=amplify), one curve per model, coloured by baseline regime; the dashed line marks the 0.05 regime boundary. Binding-limited models (orange) rise under amplification, strongly for Qwen3-1.7B, weakly for Qwen2.5-1.5B and Llama-3.2-1B, negligibly for Qwen3-4B (§4.7), and recency-copy models (blue) fall as their copy heads are removed. The claim is the direction of every trajectory, not a boundary crossing by every model.

K SUPPORTING FIGURES: THE PACKING-FORM FIT AND ITS HONEST EVALUATION

These figures support §4.2. The packing form fits the decay and beats a recency-only plateau (Fig. 34), but, as we stress in the main text, its out-of-sample skill over the $1/u$ chance trend is marginal. We therefore rest the geometric claim on causation (§4.3), not on this fit.

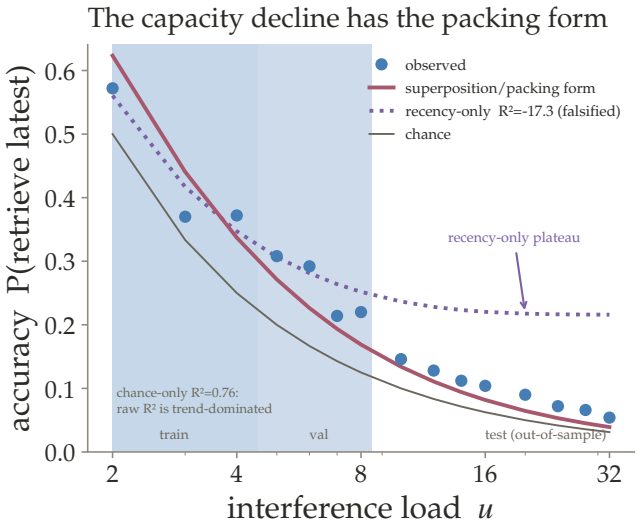


Figure 34: The packing form fits the capacity decay. Parameters fit on $u \leq 8$ (shaded), evaluation on held-out $u \geq 10$. The superposition law and the measured-geometry retrieval both track the decay to chance, and a recency-only account predicts a plateau the data break through. But the fit's advantage over the $1/u$ chance trend is small out of sample (Fig. 35), so we do not claim it as a predictive law.

model	params	elic.	regime	u^*	p / PR
Pythia-12B	12B	compl.	binding	6.03	– / –
gemma-2-9B	9B	compl.	binding	4.69	0.39 / 1.6
Llama-3.1-8B	8B	compl.	binding	3.48	0.94 / 2.0
Qwen2.5-7B	7B	compl.	binding	3.41	1.01 / 1.8
OLMo-2-7B	7B	compl.	binding	2.82	0.49 / 1.8
Mistral-7B	7B	compl.	binding	3.36	1.18 / 1.8
Falcon3-7B	7B	compl.	binding	5.52	0.90 / 1.8
Pythia-6.9B	7B	compl.	binding	5.56	0.88 / 1.4
Qwen2.5-3B	3B	compl.	binding	2.94	1.15 / 1.6
Llama-3.2-3B	3B	compl.	binding	2.00	1.06 / 1.8
gemma-2-2B	2B	compl.	binding	3.18	0.90 / 1.1
SmolLM2-1.7B	2B	compl.	binding	4.71	1.31 / 1.7
Qwen2.5-1.5B	2B	compl.	binding	2.83	1.19 / 1.6
Llama-3.2-1B-it	1B	chat	binding	3.52	– / –
Llama-3.2-1B	1B	compl.	binding	4.47	1.19 / 1.6
Qwen3-0.6B	600M	chat	binding	4.96	– / –
Qwen2.5-0.5B	500M	compl.	binding	3.41	– / –
Qwen3-32B	32B	chat	recency	–	– / –
gemma-4-31B-it	31B	chat	recency	–	0.84 / 1.9
Qwen3.5-27B	27B	chat	recency	–	1.16 / 1.7
Qwen3.6-27B	27B	chat	recency	–	– / –
Qwen3-14B	14B	chat	recency	–	0.94 / 1.9
gemma-4-12B-it	12B	chat	recency	–	0.80 / 1.6
Qwen3.5-9B	9B	chat	recency	–	1.13 / 1.8
Qwen3-8B	8B	chat	recency	2.00	– / –
Qwen3-4B	4B	chat	recency	2.57	0.85 / 1.8
Qwen3.5-4B	4B	chat	recency	–	– / –
gemma-4-E4B-it	4B	chat	recency	–	– / –
Phi-3.5-mini	4B	chat	recency	11.30	– / –
Llama-3.2-3B-it	3B	chat	recency	8.19	– / –
Pythia-2.8B	3B	compl.	recency	5.04	0.84 / 1.3
Qwen3.5-2B	2B	chat	recency	2.00	– / –
gemma-2-2B-it	2B	chat	recency	–	– / –
gemma-4-E2B-it	2B	chat	recency	–	0.80 / 2.0
Qwen3-1.7B	2B	chat	recency	2.00	0.88 / 2.1
GPT-2-XL	2B	compl.	recency	4.61	0.76 / 1.4
Pythia-1.4B	1B	compl.	recency	3.48	0.93 / 1.2
Pythia-1B	1B	compl.	recency	7.09	0.99 / 1.1
Qwen3.5-0.8B	800M	chat	recency	–	– / –
GPT-2-L	774M	compl.	recency	7.54	0.67 / 1.2
Pythia-410M	410M	compl.	recency	8.76	1.19 / 1.5
GPT-2-M	355M	compl.	recency	5.57	0.52 / 1.3
GPT-2	124M	compl.	recency	10.82	0.26 / 1.2
Qwen2.5-14B	14B	compl.	incapable	2.18	1.03 / 2.0
Yi-9B	9B	compl.	incapable	2.24	– / –
Pythia-160M	160M	compl.	incapable	2.00	– / –
Pythia-70M	70M	compl.	incapable	2.15	– / –

Table 3: Per-model results under each model's fair elicitation. The regime and u^* use the chat template for chat, instruct, and reasoning models and raw completion for base models (§4.4, App. G); this is why no strongly post-trained model is mislabeled "incapable." The geometry columns (p, PR) are the sink-controlled measurements (§4.2, Fig. 12); "–" marks models for which a given metric was not measured. The four incapable rows are base models at chance under raw completion. For the two smallest (Pythia-70M/160M) this is a genuine capacity floor, whereas Qwen2.5-14B and Yi-9B show a raw-format-parsing failure rather than a capacity one (App. G), so "incapable" here is a measurement label, not a claim that the task is unreachable for those weights.

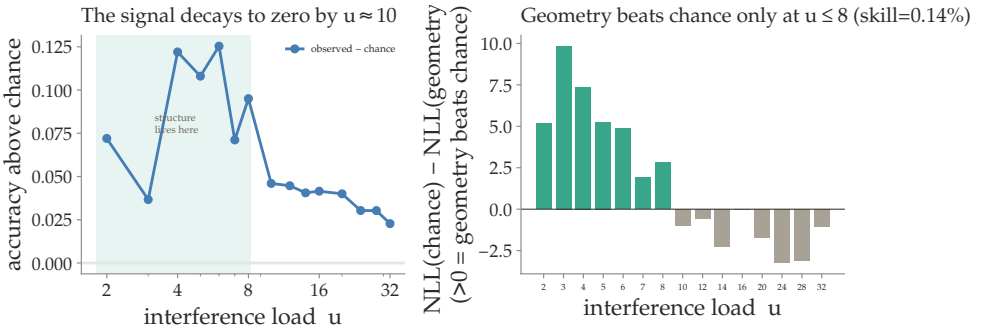


Figure 35: Chance-referenced evaluation. Left: accuracy above chance is 0.07–0.12 at low load and decays to ≈ 0 by $u \approx 10$, so there is above-chance structure to predict only at resolvable loads. Right: per-load log-likelihood skill of the retrieval-head geometry over the uniform (chance) baseline; the geometry beats chance at $u \leq 8$ and ties or loses beyond, for an aggregate skill of 0.14%. We report this rather than a trend-inflated R^2.

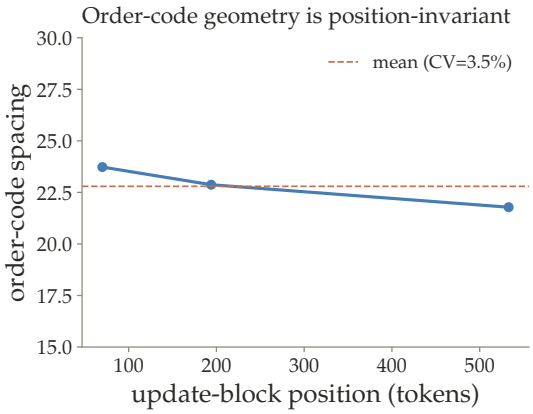


Figure 36: The order code is ordinal, not absolute position (§E.4). Order-code spacing is flat (CV$=$ 3.5%) as the update block is shifted across a $\sim 7\times$ range of absolute token positions; a position code's geometry would change with position.

L Supporting controls: position-invariance and similarity

M Matched random-head ablation control

To confirm that the forward conversion (§4.7) is specific to the copy/induction route rather than generic capacity removal, we ablate a matched number of random heads (excluding the identified copy heads) in GPT-2-large and re-measure the full accuracy-versus-load curve over five independent random draws. Ablating six matched random heads leaves the model in the recency-copy regime in all five draws (u=32 accuracy 0.15–0.20), whereas ablating the six identified copy heads flips it to binding-limited (u=32 above-chance 0.049). The conversion is thus specific to the copy/induction route, not a consequence of removing any six heads' worth of capacity.

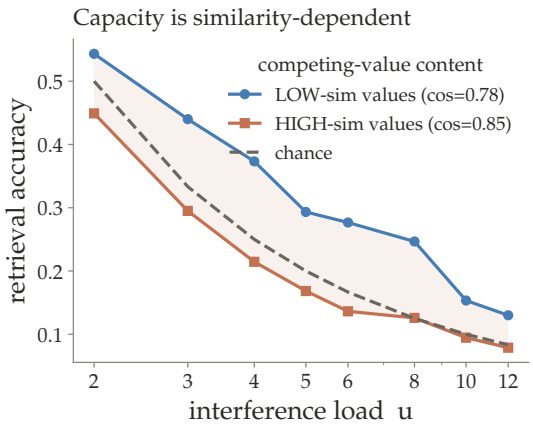


Figure 37: Similarity-dependent capacity (§E.5). Crowding value content (HIGH-sim) shifts the capacity curve left relative to spread content (LOW-sim).

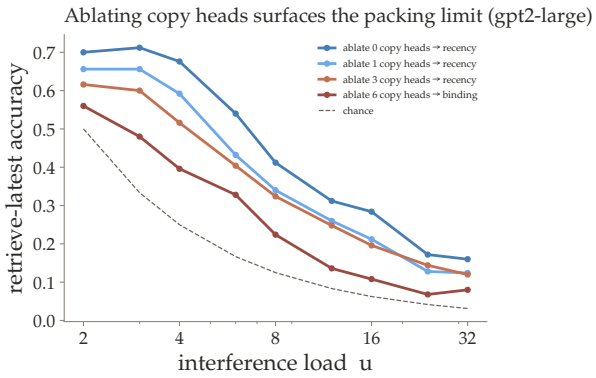


Figure 38: Editing the retrieval converts the regime. Cumulatively ablating GPT-2-large's copy/induction heads turns a recency-copy model (escapes the limit, stays above chance) into a binding-limited one (the tail steepens toward chance). A matched random-head ablation does not flip the regime (App. M), so the effect is specific to the copy route. (This ablation acts at all positions; the query-only edit that provably leaves storage untouched is the constructive direction, Fig. 5.) Apparent capacity is controlled by routing, demonstrated causally in a non-Qwen model.