Retrieval-Induced Forgetting in LLMs: A Reproducible Failure Mode and a Verified Fix

Shuzhen Li [1] **Mengdi Wang** [1]

Abstract

We document a content-controlled cue-timing dissociation in instruction-tuned LLMs: the same item-name list, delivered post-hoc versus pre-elaboration, opens a **103 pp** recovery gap at Qwen-2.5-72B (98% vs. -5%). Cue content is held fixed across the contrast; only its position in the KV cache differs. The underlying failure is elaboration-induced suppression of in-context recall: when the model elaborates on one of several related items present in its context, its later recall of the others drops by $+11.2$ pp on Qwen-2.5-7B (judge-scored, $N{=}150$, $p{<}10^{-6}$) and by $+24.9$ pp on its 72B sibling. A one-turn post-hoc "list every item by name" prompt reverses the deficit (100% recovery at 7B, 98% at 72B); the same text delivered *before* the elaboration recovers only half as much at 7B and goes negative at 72B. Per-layer attention extraction shows that the cure lowers late-layer attention to the original target tokens to $0.42\times$ baseline rather than restoring it, and an attention-mask intervention on the elaboration positions yields -21% recovery. Together these favour a *bypass-not-restore* reading: the cure writes a fresh reachable copy of the items into recent KV-cache positions while the originals stay suppressed.

1. Introduction

An instruction-tuned LLM is given a list of item names. Delivered *after* the model has elaborated on a competing topic in the same context, that list restores 98% of suppressed recall. Delivered *before* the elaboration, the same words — same items, same length, same prompt — recover -5%. Cue content does not separate the two outcomes; cache position does. The underlying failure is large, systematic, reversible, and the reversal is sensitive to where in the cache the cue lands. We document the phenomenon, measure attention layer by layer, and contrast the resulting picture against three accounts of how it works.

The phenomenon is, to our knowledge, an under-described in-context recall failure: when the context contains multiple semantically adjacent items and the model elaborates on one of them, recall of the unretrieved items in the same context drops. It mirrors retrieval-induced forgetting (RIF) in cognitive psychology (Anderson et al., 1994); whether the same dynamics emerge from transformer attention has been an open question. Section 2 positions our results against the human RIF literature, adjacent LLM long-context failure modes, and recent mechanistic-interpretability work.

We adapt the human paradigm to a three-phase recall task on 15 fictional multi-sector worlds. The four behavioural properties of Anderson–Bjork–Bjork RIF reproduce on Qwen-2.5-7B-Instruct at publication-grade sample sizes (§4); the effect strengthens at 72B (§7).

The cure works through bypass rather than repair. The post-hoc cue that recovers 98% of the behavioural failure lowers late-layer B-attention to $0.42\times$ baseline (§6); its own response enters the KV cache after the suppressing elaboration, producing a second reachable copy of the target tokens at a recent position the test-time softmax can attend to, while the original B-tokens stay suppressed. An attention-mask knockout of the elaboration positions yields -21% recovery, running counter to the "remove the competitor" account (which predicts $\sim 100\%$ recovery). The content-controlled post-vs-pre-cue contrast (identical text, different cache position) yields a wide separation at 72B and rules out cue content as the operative variable.

Contributions. (1) A content-controlled cue-timing dissociation. The same item-name list opens a 103 pp gap between post-cue (98%) and pre-cue (-5%) recovery at 72B (Table 4); the gap is 56 pp at 7B. With cue content held fixed, only cache position separates the two outcomes. **(2) A reproducible failure mode and a verified fix.** The behavioural deficit on Qwen-2.5-7B sits at $N{=}150$, $p{<}10^{-6}$, with all four Anderson–Bjork–Bjork properties reproducing at $N{\geq}60$ per cell (Table 1); $5.9\times$ amplification at 72B (§7).

[1]Princeton University. Correspondence to: Mengdi Wang <mengdiw@princeton.edu>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

A one-turn post-hoc item-name cue recovers 100% at 7B and 98% at 72B, inside a 92–110% band across six convergent measurements (Figure 3); the cure adds ~ 50 tokens of inference and requires no retraining. **(3) Bypass-not-restore mechanism.** The successful cure lowers late-layer B-attention to $0.42\times$ baseline while recovering 98% of recall (Figure 5); an attention-mask intervention on the elaboration positions yields -21% recovery (Table 6). The cure adds fresh reachable tokens rather than restoring access to the suppressed originals. All materials, prompts, code, pertrial JSONs for every reported number, and a reproducibility recipe are released at `https://github.com/lszshu/Retrieval-Induced-Forgetting`.

2. Related work

Retrieval-induced forgetting in human memory. Anderson et al. (1994) introduced the retrieval-practice paradigm in which repeated retrieval of a subset of related items suppresses recall of the unretrieved items, replicated in over 200 studies (Murayama et al., 2014). The mechanism debate divides into inhibition (Anderson & Spellman, 1995) and context-driven blocking (Jonker et al., 2013); both accounts predict that the suppressed items remain in storage and could in principle be re-accessed, but in human studies that prediction can only be tested behaviourally. Tulving's distinction between memory *availability* (stored at all) and *accessibility* (reachable via the current cue) (Tulving & Pearlstone, 1966; Tulving, 1974) frames our cure: post-hoc cueing recovers 98% of recall without restoring attention on the suppressed tokens, an accessibility intervention rather than a repair of availability. The LLM setting lets us probe the same distinction at the attention level rather than only behaviourally.

Long-context recall failures in LLMs. RIF as we measure it is elaboration-induced, in-context, a recall deficit, and conditional on phase-2 generation. These four properties separate it from positional lost-in-the-middle (Liu et al., 2024), training-data memorization (Carlini et al., 2021), attention-sink phenomena (Xiao et al., 2024), and generic distractor-induced degradation (Shi et al., 2023); a copy control replicates the latter without producing RIF (§4). The multi-turn information loss documented by Laban et al. (2025) is plausibly downstream of in-context RIF; we provide a mechanistic substrate (KV-cache attention competition) and a verified repair (post-hoc cueing). The same accessibility deficit is consequential for retrieval-augmented generation (Lewis et al., 2020), where the model routinely elaborates on retrieved passages before being asked to recall distinct items from the same context.

Mechanistic interpretability of attention. Mechanistic-interpretability work has identified attention-level structures such as induction heads (Olsson et al., 2022) and the IOI circuit (Wang et al., 2022). Wu et al. (2025) isolate a sparse set of *retrieval heads* whose ablation selectively breaks needle-in-a-haystack recall, establishing that long-context retrieval is carried by a structured, head-level circuit. RIF, in this view, is the behavioural signature of competition over the same retrieval machinery: when one item monopolises the heads' attention budget, recall of related items in the same KV cache degrades. Geva et al. (2023) dissect factual recall via attention knockouts and trace it to a multi-step flow in which late-layer attention heads extract enriched subject representations to produce the answer token; we extend this picture to the competitive regime, where the attention edges that would normally extract a related subject's representation are reweighted away and an explicit cue can fully reverse the deficit.

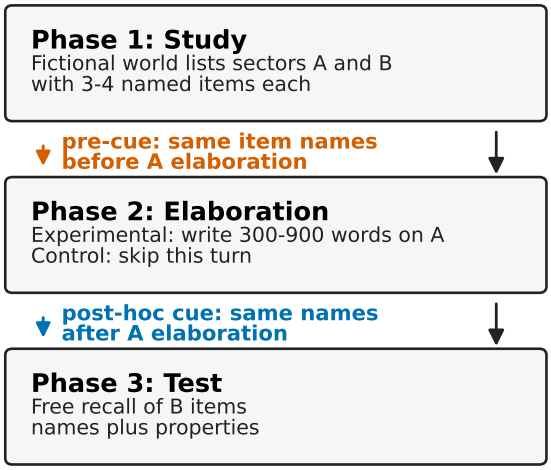

Figure 1. **The three-phase recall paradigm.** The model studies a fictional multi-sector world (phase 1), elaborates on sector A for 300–900 words (phase 2), then is asked to list the items in sector B (phase 3). The control condition skips phase 2. Cure conditions insert a one-turn prompt between phases 2 and 3; prevention conditions modify phase 2 itself.

3. Paradigm and methods

Three-phase recall paradigm. The classical retrieval-practice paradigm of Anderson et al. (1994) adapts to LLM inference as follows. Figure 1 summarises the protocol.

- **Phase 1 (study).** The model is given a 200-word fictional world description listing several sectors $A, B, \ldots$, each populated by 3–4 named items with brief property descriptions.
- **Phase 2 (elaboration).** In the experimental condition, the model is asked to elaborate on sector A for 300–900 words. In the control condition, phase 2 is skipped.

- **Phase 3 (test).** The model is asked to list every item in sector B with its properties: free recall, no multiple choice, no retrieval cue beyond the sector name.

Materials. Fifteen invented worlds (Meridian Federation, Zalorian Empire, Kenthara Alliance, . . .; full list in Appendix A) constructed so that no item name or property string returns hits in pre-training corpus audits. Phase-3 recall can therefore succeed only by retrieval from the in-context KV cache, controlling for parametric-knowledge contamination. The A/B sector pairings (Appendix A) are hand-designed to produce measurable semantic overlap; two controls discriminate overlap-dependent from materials-fragile readings. The `filler`-elaboration control (§4) shows that the effect requires *related* A/B rather than any phase-2 elaboration. The high-overlap vs. low-overlap contrast (§4; $+13.7$ pp interaction, p=0.039) shows that the effect tracks A/B semantic overlap rather than world identity.

Worked example. A representative trial illustrates the failure and the cure. The *Ashkari Collective Transport* world describes an Aerial Transit sector with three items: Skydrift Shuttles, Cargo Zeppelins, and Emergency Drones, each with five concrete property details (capacity, range, technology, etc.). In the control condition the model recovers 12 of 13 (88%) details. After elaborating on a sister sector (Maritime Transit), the same model in the `none` condition still names all three items but its recovery on details drops to 4 of 13 (31%): properties become hedged or absent ("does not specify a maximum passenger capacity"). A one-turn "list every item from both sectors by name" cue, inserted between elaboration and test, restores recovery to 12 of 13 (88%) on the same trial—each property is now named with the original concrete value. The content of the cue is two short item lists; what changes between `none` and `explicit` is what sits in the cache at test time. Figure 2 shows the trial.

Scoring. Each phase-3 response is rated independently for each ground-truth B-item by an LLM judge given the world description, the item, and the response, and asked whether the response retrieves the item by name or by a unique property. The primary quantity is per-condition recovery percentage aggregated across trials.

Models. Qwen-2.5-7B-Instruct (primary) and Qwen-2.5-72B-Instruct (scale) (Hui et al., 2024). Inference uses sampling temperature T=0.7 and top-p=0.95. Compute budget in Appendix D.

Statistics. Position counterbalancing (A-before-B vs B-before-A in phase 1) across all main results. Paired t-tests on matched control / experimental trials within world; per-world means used for the N=15-world clustered analysis. We report point estimates with p values in the main tables; per-trial standard errors and 95% confidence intervals on every reported recovery percentage are computed by bootstrap over trials.

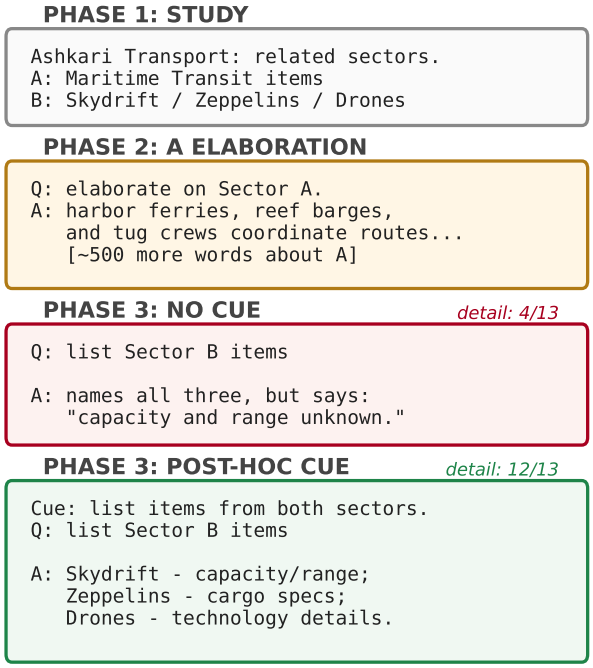

Figure 2. **Worked example: Ashkari Collective Transport.** Same trial, same model, three conditions. Control (top): all three items recovered with full property detail (88% detail rate). `none` after elaboration on a sister sector (middle): items named but properties hedged or absent (31%). `explicit` post-hoc cue (bottom): properties restored to control level (88%). The cue text is two short item lists; what changes is what sits in the cache at test time.

Why this paradigm. Three design choices isolate the failure mode from confounds common in long-context evaluations. Fictional materials prevent the model from succeeding via parametric knowledge: an item named "Skydrift Shuttles" returns no pre-training-corpus hits, so phase-3 recall must come from in-context retrieval. Position-balanced ordering (A-before-B and B-before-A) controls for primacy / recency artefacts that would otherwise confound the elaboration manipulation. Free recall (rather than multiple choice) requires the model to actively retrieve property details rather than recognise them, giving us the multi-item detail-rate metric that judge scoring evaluates.

4. Existence: four classic RIF properties transfer

Existence is established on Qwen-2.5-7B-Instruct under the position-balanced design with N=150 trials per condition (15 worlds × 5 seeds × 2 orderings).

Table 1. Main effect on Qwen-2.5-7B-Instruct (judge-scored), N=150 trials, position-balanced.

	Ctrl	Exp	RIF (pp)	p
Recall (judge)	0.824	0.712	+11.2	$<10^{-6}$

At elevated dose (∼600 w phase-2 elaboration, normal position) the judge-scored effect grows to +26.0 pp. The four classic Anderson–Bjork–Bjork properties all transfer:

- **Existence.** +11.2 pp judge / +26.0 pp at high dose, N=150.
- **Retrieval specificity** (Anderson et al., 2000). A copy control matched on KV-content, token-count, and semantics, where the model reads the same A-elaboration text without generating it, produces no detectable drop (−1.1 pp, p=0.71). Generation produces −15.4 pp (p=0.0003 vs. copy). The deficit therefore tracks generation, not context length or repeated exposure.
- **Competition dependence.** High-overlap A/B (same super-category) gives +17.4 pp (p=0.002); low-overlap gives +3.7 pp (p=0.33, ns). Interaction +13.7 pp (p=0.039).
- **Cue independence** (Anderson & Spellman, 1995). Probing phase-3 with a paraphrased descriptor not seen during phase 1 gives identical RIF (+11.5 pp under both standard and novel cues).

Two further patterns specific to LLMs accompany the classical four. The dose-response is a step function near 300 words: near-zero below 200 w, +27 pp by 355 w, saturating above 500 w (Appendix B). The human part-set cuing effect (partial B-items as pre-cues suppress recall of the rest in humans) is absent on Qwen (p=0.83); partial cues complement B-tokens in attention, consistent with the cache-position account in §6.

5. Cure vs. prevention: timing is the mechanism

Is the failure correctable, and if so by what class of intervention? We test two intervention families: cures applied after the A-elaboration, and preventions applied before or during it. The two dissociate, and the dissociation is mediated by cue timing rather than cue content.

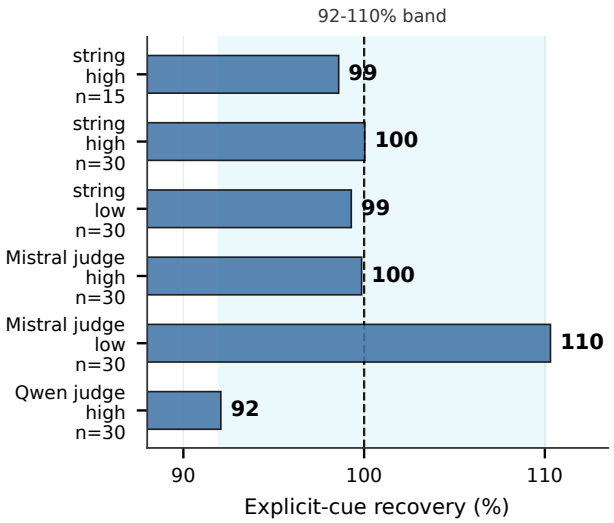

Figure 3. **The explicit-name cue restores recall on every measurement.** Six convergent measurements of `explicit`-cure recovery on Qwen-2.5-7B (judge-scored), varying sample size (n=15 and n=30) and dose (∼200 w and ∼500 w phase-2 elaboration). Each row labels the varied conditions and the trial count; recovery sits within the shaded [92, 110]% band on every row.

5.1. Cure family

Four post-hoc cure variants are tested at high dose (∼500 w) on Qwen-2.5-7B, N=30 per condition × 6 conditions, position-balanced. Each cure inserts a one-turn "reminder" between phase 2 and phase 3, varying only what the reminder asks the model to write:

Table 2. Cure-family recovery (judge-scored). Recovery = (cured − failed)/(control − failed). Only the explicit-name cue reaches full recovery.

Cure	Judge
`reset`	70%
`summarize`	59%
`explicit`	**100%**
`detailed`	89%

The 100% point estimate of Table 2 sits inside a 92–110% band across six measurements that vary sample size (N=15, 30) and dose (∼200 w, ∼500 w) (Figure 3). The recovery band is what we report; the point estimate is only the band's centre. The six measurements share materials, architecture, and explicit-cue text; they are convergent rather than statistically independent.

5.2. Prevention family

Preventions modify phase 2 itself, cueing the model before or during the elaboration that drives the suppression. We test four variants on Qwen-7B at matched dose. None reaches the floor of the cure family (`reset`, 70%). Warnings about

Algorithm 1 Post-hoc explicit-cue procedure (the verified fix). The cue inserts one turn between phase 2 and phase 3, costs ∼50 tokens of inference, and requires no retraining.

1: **Phase 1 (study).** Model receives world description W containing sectors $A, B, \ldots$, each with 3–4 named items.
2: **Phase 2 (elaboration).** Model elaborates on sector A for $d \in [300, 900]$ words.
3: **Cue (the fix).** Insert one turn: *"Before answering the next question, please list every specific item from BOTH sectors in the original information, by their specific names."*
4: **Phase 3 (test).** Ask the model to list items in sector B with their properties.

sector B (`soft`, `balanced`) recover 25–30% and are not significant. Forcing alternation between A- and B-items inside phase 2 (`interleave`) reaches 63%. Prepending the explicit B-item-name list before A-elaboration (`pre-cue`) reaches only 44% (p=0.10; Table 3). The `pre-cue` variant is content-identical to the explicit cure; only its position differs (§5.3).

Table 3. Prevention recovery. No prevention matches even the weakest cure (`reset`, 70%).

Prevention	Recovery (judge)
`soft warning`	25% (ns)
`balanced warning`	27–30% (ns)
`interleave`	63% (trend, p=0.08)
`pre-cue`	44% (trend, p=0.10)

5.3. The critical control: same content, different timing

The cure-vs-prevention gap admits two interpretations: cue content (cures name specific items, preventions do not) or cue position (cures intervene at retrieval, preventions at encoding). The `pre-cue` control isolates position by holding content fixed: it lists the same item names as the explicit cure, only earlier in the sequence.

Table 4. **Cue-timing dissociation: same content, same length, different position.** The item-name list delivered post-hoc gives full recovery; delivered pre-hoc, it recovers half at 7B and goes negative at 72B. The **103 pp gap at 72B** is robust to the small-N and judge-track concerns that bear on other tables.

Scale	post-cue	pre-cue	Gap
Qwen-2.5-7B (N=60)	100%	44%	**56** pp
Qwen-2.5-72B (N=60)	98%	**−5%**	**103** pp

Holding cue content fixed, only its position varies, and recovery flips with it. The 72B gap is inconsistent with two simple accounts. Content-sufficiency, on which any prompt with B-item names should work, fails: the same words at the wrong position recover −5%. The warning-as-encoding account fails for the same reason: telling the model what to expect should improve recall, but here the warning is what accompanies the negative recovery.

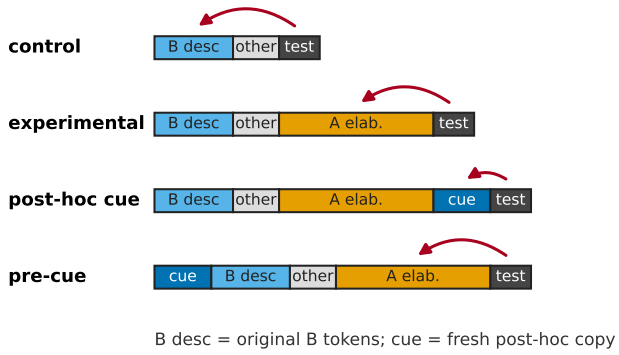

Figure 4. **Where test-query attention lands across conditions.** Each row is a schematic KV-cache layout for one condition; the curved arrow marks the dominant attended cache region (read off Figure 5). In the experimental condition the A-elaboration captures the budget that would otherwise reach the original B-tokens. The post-hoc cue inserts a fresh B-name copy *after* the elaboration where the test-time softmax can still reach it; the pre-cue's B-names sit before A and are suppressed along with the originals.

Only an accessibility-at-retrieval account fits: the cue works when, and only when, its content lands in the KV cache after the suppressing A-elaboration, where phase-3 attention can still reach it. §6 shows this is what happens.

6. Mechanism: a fresh reachable copy, not a repaired one

Per-step logits, per-layer attention, and a direct attention-mask intervention together localise the failure (Figure 4 sketches where test-time attention lands across conditions).

6.1. The suppression signal at the logit and attention levels

We trace the suppression signal first at the per-step logit level. For each trial in 15 matched control/experimental pairs we record the per-step logit assigned to the first token of each B-item name across phase 3. Averaged across steps and items, B-token logits are suppressed by −1.27 in the experimental condition (t=8.98 paired across trials, $p<10^{-4}$), a 3.6× probability reduction at every position. The suppression is constant across the response, which rules out a transient cue-conflict explanation.

Per-layer attention extraction extends the picture. A subset re-run with `attn_implementation="eager"` extracts attention from the last test-phase token to every non-

generated position, layer by layer. B-attention is reduced 4–19× in the experimental condition across all 27 decoder layers of Qwen-2.5-7B; A-attention is correspondingly increased 3–6× in mid-to-late layers (L12–L25). The attention budget has shifted from B toward A, starving the items phase 3 asks about.

6.2. Cure mechanism

If RIF is attention redistribution away from B, the obvious cure should work by restoring B-attention. It does not. We replicate the attention extraction in three matched conditions on 15 trials, measuring attention from the last test-phase token to the original B-description tokens in the context block (excluding cue-response tokens to keep denominators comparable). Table 5 reports the per-layer ratios; the late-layer L15–L27 average is 0.42× baseline (paired $p<10^{-3}$). The `explicit` cure recovers 98% of behavioural RIF while lowering late-layer B-attention by 58%. Sequence length is matched between `none` (865 tokens) and `explicit` (852); the ratio is not a dilution artefact.

Table 5. Per-layer attention to the original B-tokens, ratio of `explicit`-cure to `none`. The cure recovers 98% of behavioural recall while *lowering* late-layer attention to the original tokens to 0.42× baseline (paired $p<10^{-3}$). Sequence lengths matched to within 1.5%.

Layer band	explicit/none ratio
Early (L0–L4)	1.00×
Mid (L5–L14)	0.43×
Late (L15–L27)	**0.42**×

Mechanistic reframe. The cure response (the model's enumeration of B-item names when prompted) itself enters the KV cache at positions after the A-elaboration wedge, where it is not shadowed by A's attention dominance. At test time the model attends to this fresh enumeration, recent and unsuppressed, instead of to the original early B-description. The cure does not repair the suppressed B-entries; it writes a second, unsuppressed copy at a cache position the test-time softmax can reach. This is the bypass mechanism named in §1.

A hierarchy across the cure variants of Table 2 follows: item-name cues recover 92–110%, topic-level cues 53–88% under judge, generic prompts 70–84%, and pre-elaboration cues only 44% at 7B and −5% at 72B (§5.3). The more item-specific content the cue writes into the recent cache, the better the recovery. The 47 pp `summarize`-vs-`explicit` gap under judge scoring is the gap between "topic tokens in recent cache" and "item tokens in recent cache."

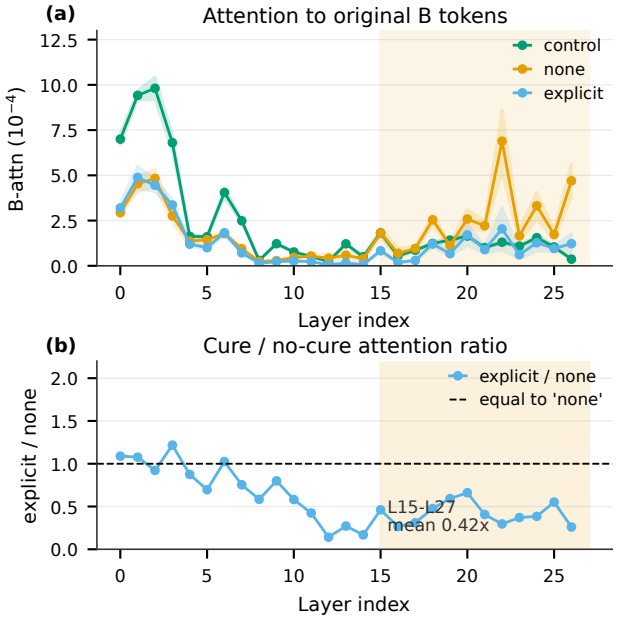

Figure 5. **The successful cure lowers late-layer B-attention rather than restoring it.** (a) Mean per-layer B-attention in the `control` (green), `none` (orange, the failed condition), and `explicit`-cure (purple) conditions on Qwen-2.5-7B; shading is ±1 SEM across 15 matched trials. The cure curve sits at or below `none` from mid-layers onward despite recovering 98% of behavioural recall. (b) The `explicit`/`none` ratio per layer; from layer 15 onward (orange shading) the ratio averages 0.42 (−58% B-attention).

6.3. Direct intervention: attention-mask knockout

Bypass-not-restore predicts that removing A-elaboration from the phase-3 attention distribution, without adding any fresh cue, should fail to repair the failure. The remove-the-competitor account predicts ∼100% recovery. The contrast is tested via the `attention_mask` parameter in `generate()`, which zeroes out the elaboration positions during phase 3. Token-span localisation uses `return_offsets_mapping` and a first-80-character locator on the phase-2 response; full code in Appendix C. Setting the mask to zero at those positions prevents them from being attended to during decode.

Table 6. Attention-mask knockout on Qwen-2.5-7B, 30 trials × 15 worlds. Recovery = (masked_exp − experimental)/(control − experimental) × 100%. The mask not only fails to abolish RIF, it marginally worsens it.

Condition	Recall	Δ ctrl	p
control	0.916	—	—
experimental	0.708	−20.8 pp	$<10^{-4}$
masked_exp	0.665	−25.2 pp	$<10^{-4}$
Recovery = −21% (vs. experimental)			

A +50% recovery here would have supported restore. The

observed result runs in the opposite direction.

6.4. Alternative accounts

Three alternative readings are inconsistent with the combined evidence of §§ 5–6. Representational damage (availability loss in Tulving's sense) is hard to reconcile with 100% post-hoc recovery. Naive attention-restore at the original B-position is contradicted by the per-layer measurement (Figure 5): the cure lowers late-layer B-attention by 58% while still recovering 98% of behavioural recall. Remove-A (A's KV presence as the sufficient cause) predicts $\sim 100\%$ recovery from the attention-mask intervention, while the observed value (Table 6) points the other way, at -21%. The remaining account, attention redistribution at test time routed by cache duplication, is what §5.3's cue-timing dissociation tests most directly.

Together, the per-layer attention pattern (the cure lowers late-layer B-attention to $0.42\times$ baseline) and the attention-mask knockout (-21% recovery) are convergent with bypass-not-restore.

7. Scaling

Figure 6 summarises the main scaling patterns. Under identical $N{=}150$ protocol, Qwen-2.5-72B amplifies the trigger to $+\mathbf{24.9}$ **pp** (control 0.926, experimental 0.678, $p{<}10^{-6}$, $d{=}1.56$). The four-property signature and the dose threshold persist; the threshold shifts slightly left (~ 250 w). The cure hierarchy is preserved and sharpened: `reset` 92%, `summarize` 71%, `explicit` 98%, `detailed` 116% (all $p{<}10^{-6}$).

The larger 72B effect weighs against a small-model-curiosity reading. The post-vs-pre-cue timing gap widens too: from 56 pp at 7B to 103 pp at 72B. Post-cue recovery holds at 98%; pre-cue recovery collapses from 44% to -5%. Forced-interleave prevention also lifts with scale ($47\% \rightarrow 75\%$, 7B to 72B) but does not match the post-hoc cure on either model.

Two scaling signatures support bypass. First, the cue-timing gap widens with scale (56 pp $\rightarrow$ 103 pp). Bypass predicts the widening because pre-cue B-name tokens and the original B-description sit in the same early cache region; deeper KV-cache accumulation at 72B suppresses both more thoroughly, while the post-hoc cue's tokens are written *after* the elaboration and remain reachable. Second, post-cue recovery is scale-invariant ($100\% \rightarrow 98\%$): the recent-position channel does not depend on the depth of suppression at earlier positions. A cure that worked by attention restoration should weaken at 72B where the suppression to reverse is deeper; we observe the opposite. Both signatures are quantitative consequences of the cache-position account rather than independent stipulations.

The within-Qwen 7B vs. 72B contrast is the empirical anchor for the mechanistic claims of §6; generality across other architectures is out of scope for this paper.

8. Discussion

Tulving & Pearlstone (1966); Tulving (1974) distinguished memory *availability* (stored at all) from *accessibility* (reachable via the current cue). LLM RIF sits in the accessibility bucket: post-hoc cues fully restore recall (§5), logit-level suppression is reversible (§6), and pre-hoc cues with identical content fail (§5.3). Tulving's framework does not anticipate the repair mechanism we observe. Accessibility, in a write-capable cache, can be restored by appending fresh reachable tokens at a later position, rather than by lifting suppression on the suppressed tokens themselves.

FAGEN scope, delivered. This paper instantiates FAGEN's four-output structure on a single failure mode: (a) an operational definition of the deficit on a controlled three-phase paradigm (§3); (b) a reproducible trigger at $N{=}150$ on Qwen-2.5-7B with a clean $\sim$300-word dose threshold (§4) and a $5.9\times$ amplification at 72B (§7); (c) trace-level diagnostics that go beyond final-score reporting, locating the failure at the per-layer attention level and intervening on it via attention-mask knockout (§6); (d) a verified low-cost fix (Algorithm 1) whose cost is explicitly admitted ($\sim$50 tokens, one turn, no retraining) and whose efficacy sits inside a 92–110% recovery band across six convergent measurements (Figure 3). The cue-timing dissociation is the load-bearing causal control: with content held fixed, only cache position separates 98% from -5% recovery.

Bypass-not-restore as a principle. The mechanistic finding generalises beyond the present paradigm in a specific sense: in any write-capable memory architecture where retrieval queries the most recent reachable cache positions, accessibility deficits can be repaired by *adding* fresh retrieval-reachable tokens, even when the original trace is left suppressed. The contribution is empirical at the substrate level (transformer KV cache, instruction-tuned LLM at the dose and materials scope of §3) rather than architectural in the abstract. Whether the same accessibility-via-duplication principle holds in other architectures, scales, or modalities is an open empirical question we leave to follow-up work; we do not claim it transfers automatically beyond what the present experiments demonstrate.

Within-Qwen scaling is the empirical anchor for the mechanistic claims; further architectures are out of scope for this paper.

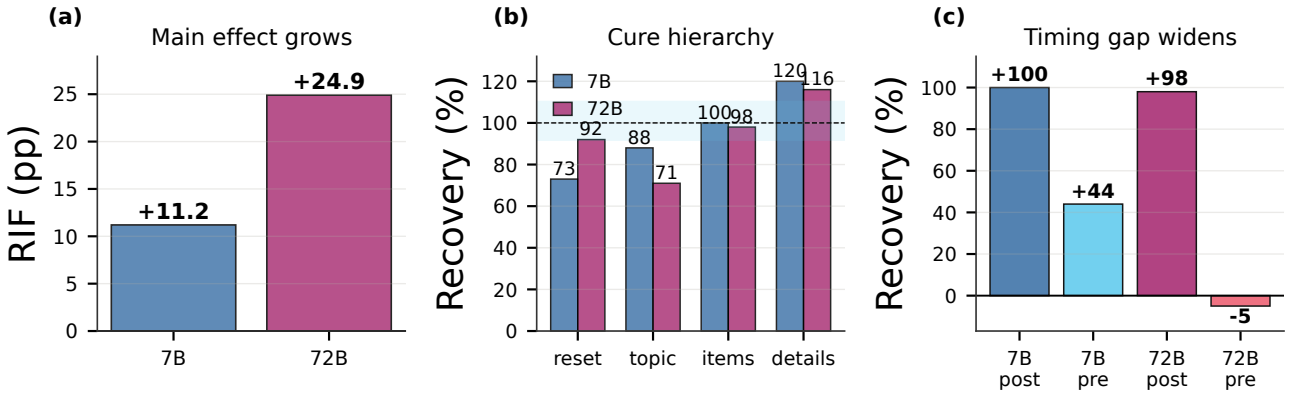

Figure 6. **Scaling within the Qwen-2.5 family (7B vs. 72B).** (a) Main effect: $+11.2$ pp at 7B and $+24.9$ pp at 72B ($N{=}150$, $p{<}10^{-6}$ for both). (b) The cure hierarchy is preserved at scale: `reset` $73\% \rightarrow 92\%$, `summarize` $59\% \rightarrow 71\%$, `explicit` $100\% \rightarrow 98\%$, `detailed` $89\% \rightarrow 116\%$. (c) The cue-timing gap WIDENS with scale: post-cue 98% vs. pre-cue -5% at 72B opens a 103 pp gap, up from 56 pp at 7B.

9. Limitations

Three scoping notes bound the present claims. **(i) Materials.** The three-phase paradigm uses fictional multi-sector worlds (§3), trading materials breadth for parametric-knowledge control; we make no claims about naturalistic retrieval-augmented generation in the wild, and explicitly do not extrapolate the within-document repair to cross-document RAG settings. **(ii) Mechanism granularity.** The per-layer attention measurement and the `attention_mask` intervention operate at the layer-band level (early / mid / late) rather than per-head; per-head causal patching of a sparse retrieval-head circuit is the natural follow-up and is left for future work. **(iii) Operating regime.** The cure efficacy is anchored on the canonical dose range ($\sim$200–500 w of phase-2 elaboration) and six convergent measurements at $N{=}15$–30 per cell; out-of- distribution behaviour — extreme doses, longer contexts beyond the studied range, or very different cue surface forms — is not characterised here.

10. Conclusion

The headline result is that cache position, not cue content, separates 98% recovery from -5% recovery at 72B (§5.3). The supporting picture is a reproducible in-context recall failure that mirrors human retrieval-induced forgetting (§4), curable by a one-turn item-name prompt costing $\sim$50 tokens (§5, Algorithm 1), with the cure working through bypass rather than attention restoration (§6, Figure 5). At 72B the trigger is $5.9\times$ stronger and the cue-timing gap widens from 56 pp to 103 pp; the cure remains scale-invariant (98% recovery at 72B). **In sum:** with cue content fixed, cache position alone separates 98% from -5% at 72B — the verified fix is an addition, not a restoration.

The principle we draw is bypass-not-restore as accessibility repair via fresh-token addition in write-capable caches; we offer it grounded at the substrate level studied here, and release all materials, prompts, code, and per-trial JSONs at `https://github.com/lszshu/Retrieval-Induced-Forgetting` so that any reader can verify each headline number and re-run the experiments.

References

Anderson, M. C. and Spellman, B. A. On the status of inhibitory mechanisms in cognition: memory retrieval as a model case. *Psychological review*, 102(1):68, 1995.

Anderson, M. C., Bjork, R. A., and Bjork, E. L. Remembering can cause forgetting: retrieval dynamics in long-term memory. *Journal of experimental psychology: Learning, memory, and cognition*, 20(5):1063, 1994.

Anderson, M. C., Bjork, E. L., and Bjork, R. A. Retrieval-induced forgetting: Evidence for a recall-specific mechanism. *Psychonomic bulletin & review*, 7(3):522–530, 2000.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Geva, M., Bastings, J., Filippova, K., and Globerson, A. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235, 2023.

Hui, B., Yang, J., Cui, Z., Yang, J., Liu, D., Zhang, L., Liu, T., Zhang, J., Yu, B., Lu, K., et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.

Jonker, T. R., Seli, P., and MacLeod, C. M. Putting retrieval-induced forgetting in context: an inhibition-free, context-based account. *Psychological review*, 120(4):852, 2013.

Laban, P., Hayashi, H., Zhou, Y., and Neville, J. Llms get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120*, 2025.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173, 2024.

Murayama, K., Miyatsu, T., Buchli, D., and Storm, B. C. Forgetting as a consequence of retrieval: a meta-analytic review of retrieval-induced forgetting. *Psychological bulletin*, 140(5):1383, 2014.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Shi, F., Chen, X., Misra, K., Scales, N., Dohan, D., Chi, E. H., Schärli, N., and Zhou, D. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pp. 31210–31227. PMLR, 2023.

Tulving, E. Cue-dependent forgetting: When we forget something we once knew, it does not necessarily mean that the memory trace has been lost; it may only be inaccessible. *American scientist*, 62(1):74–82, 1974.

Tulving, E. and Pearlstone, Z. Availability versus accessibility of information in memory for words. *Journal of verbal learning and verbal behavior*, 5(4):381–391, 1966.

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

Wu, W., Wang, Y., Xiao, G., Peng, H., and Fu, Y. Retrieval head mechanistically explains long-context factuality. In *International Conference on Learning Representations*, volume 2025, pp. 62143–62156, 2025.

Xiao, G., Tian, Y., Chen, B., Han, S., and Lewis, M. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024.

A. Materials and condition templates

All worlds are invented for this paper. Each consists of a 200-word description of a fictional polity, organisation, or research collective, with 4–6 sectors, each sector populated by 3–4 named items with brief property descriptions. Item names are constructed to have no pre-training-corpus matches; properties combine plausible but invented facts (units, dates, agents, processes).

Table 7. The 15 fictional worlds and their sector pairings used as A/B targets in the position-balanced design (Table 7). Each world contributes 10 trials per condition: 5 seeds $\times$ 2 orderings (A before B / B before A in phase 1).

World	Sector A (default)	Sector B (default)
Zalorian Empire Economy	Agricultural	Industrial
Meridian Federation Science	Biological Sciences	Physical Sciences
Thalvian Republic Culture	Musical Traditions	Culinary Traditions
Ashkari Collective Transport	Surface Transit	Aerial Transit
Veridian League Education	Youth Academy	Adult Institute
Kenthara Alliance Medicine	Pharmaceutical	Surgical
Brontium Commonwealth Defense	Naval	Ground
Orvalen Confederation Architecture	Civic	Residential
Nyserine Republic Agriculture	Highland	Lowland
Khaldoryan Federation Energy	Renewable	Conventional
Pelmarc Dominion Communications	Short-Range	Long-Range
Darmun Coalition Sports	Team	Individual
Xeranth Kingdom Textiles	Natural	Synthetic
Vorasian Union Timekeeping	Mechanical	Astronomical
Jelryth Concord Mining	Surface	Deep

Phase-2 condition templates. Table 8 lists the full set of phase-2 condition templates used across the paper:

Table 8. Phase-2 prompt templates. Within-condition wording held fixed across worlds and seeds.

Name	Phase-2 prompt
`control`	(phase 2 skipped)
`experimental`	Elaborate in detail on the ⟨A-sector⟩ of the ⟨world⟩.
`filler`	Elaborate in detail on an unrelated topic from history or science.
`copy`	The following is a description of the ⟨A-sector⟩: ... (model reads, no generation).
`reset`	Consider all the information above.
`summarize`	List all topics mentioned above by name.
`explicit`	List every specific item from both sectors by name.
`detailed`	Briefly describe each specific item with quantities or properties.
`pre-cue`	(prepended before A-elaboration) The items in ⟨B-sector⟩ are: ⟨B-item names⟩.
`interleave`	(instead of free A-elaboration) Describe one item from ⟨A-sector⟩, then one from ⟨B-sector⟩, alternating.

The full prompt templates (phase-1 study, phase-2 elaboration in each of the condition variants, phase-3 test, judge prompt) are released at the repository https://github.com/lszshu/Retrieval-Induced-Forgetting; see `PROMPTS.md` there.

B. Dose-response sweep

Phase-2 elaboration was instructed to produce 114, 200, 355, 500, or 605 words (N=30 trials per dose, normal position, judge-scored on Qwen-2.5-7B). The step function in the main paper text is from this sweep (Table 9):

The threshold appears between 200 w and 355 w; below it, the paradigm produces no measurable RIF; above it, the effect saturates. This is an LLM-specific quantitative prediction absent from the human literature.

Table 9. Qwen-2.5-7B dose-response, judge-scored.

Phase-2 dose (words)	RIF (pp)	p
114	$+3.6$	ns
200	$+12.9$	<0.01
355	$+27.1$	$<10^{-4}$
500	$+27.8$	$<10^{-4}$
605	$+28.4$	$<10^{-4}$

C. Attention-mask knockout: full procedure

The §6.3 knockout uses the `attention_mask` parameter of `model.generate()` to zero out the A-elaboration response positions during phase-3 generation. Token-span localisation:

1. Run phase 2 normally, store the model's response text.
2. Construct the phase-3 prompt by concatenating the original context, the phase-2 prompt, the phase-2 response, and the phase-3 question.
3. Tokenise the phase-3 prompt with `return_offsets_mapping=True`.
4. Locate the phase-2 response's character span in the prompt using a first-80-character locator on the response text.
5. Use the offset mapping to convert the character span to a token range $[s, e)$.
6. Construct the attention mask as $\mathbf{1}_n$ with positions $[s, e)$ set to 0.
7. Pass the mask to `model.generate(input_ids=..., attention_mask=...)`.

The masked positions contribute nothing to softmax normalisation or to attended values at every layer at decode time. The `attention_mask` intervention tests whether removing A's attendable presence at decode rescues recall.

Per-condition matched seeds. To minimise sampling noise, the knockout pairs each `experimental` trial with a same-seed `masked_exp` trial. The phase-2 response is generated once per seed, then the phase-3 generation is repeated twice with and without the attention mask, deterministically given the same prompt prefix and seed.

Sanity check on the mask. Response-content comparison: the masked-condition phase-3 response diverges from the unmasked-condition response after 81 shared characters on average; both conditions tackle the same B-sector (when prompted). The mask *does* have a generation-level effect; it is just not the predicted recovery effect.

D. Compute and reproducibility

Table 10 summarises the wallclock and GPU-hours per experiment block.

Table 10. Compute used for all reported results. All inference; no training. Academic HPC cluster, bf16 on A100 (40/80 GB) and H200 (141 GB) nodes.

Experiment block	Wall (h)	GPU-hours
Phase A existence + 4 properties (Qwen-7B, N=150+)	~ 30	30 A100
Phase B/C cure-vs-prevention (Qwen-7B, 4-cure, prev)	~ 25	25 A100
Qwen-72B scaleup (2×H200, 8h walltime)	~ 8	16 H200
Qwen-72B 4-cure / pre-cue / interleave (2×H200)	~ 24	48 H200
Cure-attention v2 (15 trials, eager attention)	~ 0.4	0.4 H200
Knockout v4 (30×15, 30 m)	~ 0.2	0.2 H200
LLM-judge rescoring	~ 4	4 A100
Total		$\sim$ 125 A100+H200 hours

Environment. Python 3.13.5, `torch` 2.11.0+cu130, `transformers` 5.5.0, `numpy` 2.4.4, `scipy` 1.17.1, `matplotlib` 3.10.8. Single `uv.lock` (1359 lines, 67 direct+transitive packages) pinned to a project-local `.venv` (4.9 GB). A fresh `uv sync --frozen` reproduces the environment in ~ 10 seconds (login-node CPU, after first download), and an end-to-end sanity script (`jobs/misc/run_repro_smoke.sh`) verifies import / project-module loading / figure regeneration in 28 seconds total.

Aggregate data release. The cross-model summary (`results/cross_model_summary.json`) aggregates the main results into a single JSON: scaleup, 4-cure mitigation, pre-cue timing, and interleave. All paper figures and tables derive from this single source. Materials, prompts, and runner scripts are released alongside the code.

Model checkpoints. All models are released by their authors at `Qwen/Qwen2.5-7B-Instruct` and `Qwen/Qwen2.5-72B-Instruct`.

Sampling. Temperature $T = 0.7$, top-$p = 0.95$, max_new_tokens varies by phase (600 for phase 2 elaboration, 300 for phase 3 test). All experiments seeded; per-experiment seeds recorded in per-trial JSON output.