Query-Side Gains and Accumulation Limits in Online In-Weight Memory

Anonymous authors
Paper under double-blind review

Abstract

We study when online LoRA writes into a frozen 9B chat model yield usable memory, measuring readout, selection, and application against frozen controls. Under per-task isolation, InMind reaches 0.424 indirect application versus 0.312 for its frozen question comparator (historical judge protocol, not the new local judge). A pre-registered three-seed rerun meets its criterion under the primary judge; historical second-judge parsing failures leave that replication unresolved. One-shot transfers retain the margin on fresh tasks and PrefEval. A prespecified repair adds a same-inference-budget frozen hypothetical-key front end (0.328 application against 0.432) and an empty-query ablation of the same adapters (0.496 retrieval against 0.584): the application estimate is positive ($+0.104$ at n=125) but imprecise, and neither primary test is significant, raw or Holm-corrected. Historical coherent accumulation retrieves 0.192 versus 0.178 for frozen questions, despite a 0.705 stem-hit recall proxy. A new same-adapter control (20 personas, 146 queries) retrieves 0.212 with the correct query versus 0.082 with an empty one (equal-persona difference 0.147, conversation-cluster sign-flip p=0.0078): question conditioning within the accumulated trained state, but low absolute retrieval and no frozen arm. Paraphrase writes reverse a likelihood index's preference from exact strings to held-out paraphrases, a replicated write-format effect. A fixed-snapshot Mem0 repair isolates a query intervention (0.024 to 0.232, local Qwen judge), with ingestion workloads unmatched. Writes buy query-side readout under isolation; usable memory under accumulation is not established.

1 Introduction

A model that updates its parameters during interaction must make what it learns usable in later answers. We separate four observables: likelihood preference for a written line, generation of the fact, retrieval of the relevant store line, and application in a judged answer. They can come apart; we test their relationship here without claiming a general hierarchy or a single causal bottleneck.

Online means protocol-prescribed updates between interactions, not hard real-time latency or self-chosen supervision. These experiments test isolated fact writes and separate accumulation runs, not one continuously maintained agent.

The substrate is LoRA writing into a frozen instruction-tuned 9B chat model, plus an append-only text log (§2). InMind (Li et al., 2026a) pairs each of 125 personal disclosures ("I have a tree-nut allergy") with a semantically distant query ("walk me through a macaron recipe") requiring unprompted application. Published external-memory systems reach $\leq 14.4\%$ indirect application against an in-context ceiling of 84%; naive RAG reaches 16.0%. The benchmark's always-in-state diagnostic (0.688) prepends a rewritten profile and bypasses query-conditioned retrieval. To our knowledge this is its first parametric-memory arm.

Instrument controls remove the write while retaining the tested readout pipeline; stronger frozen generators and question-conditioned selection require separate controls. Chains have pre-registered falsification lines; audits voided three flattering instruments (App. A). The number of record is the pre-registered multi-seed rerun of the selected recipe, not the exploratory maximum. Mem0 (Chhikara et al., 2025) is rerun, with store and workload differences disclosed (§7).

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.

Contributions. (1) *Application and attribution*: isolated InMind 0.424 vs 0.312, with pre-registered reruns and transfers separated from exploratory selection; the repair's stronger frozen comparator and empty-query ablation leave both primaries nonsignificant, raw and corrected (§3). (2) *Selection under accumulation*: no clear historical written–frozen advantage (0.192 vs 0.178), but same-adapter question conditioning (0.212 correct vs 0.082 absent retrieval; conversation-cluster p=0.0078), at low absolute success and without a frozen arm (§3). (3) *Readout and write format*: a replicated likelihood-key inversion (§5) and exploratory dose, schedule, and content findings with explicit confounds (§4–6). (4) *External-store control*: a fixed-snapshot query intervention, historical configuration comparisons, and an inconclusive small LoCoMo reversal; cross-architecture workloads remain unmatched (§7).

Nearest neighbors. Test-time memory architectures learn a memory module (Behrouz et al., 2025); parametric RAG serves injected knowledge from weights (Su et al., 2025; Tan et al., 2025). Other work routes among offline adapters (Pulici et al., 2026), consolidates into serving weights (Li et al., 2026b), or manages tiered external context (Packer et al., 2023), the family of the Mem0 control. Here the trained adapter generates retrieval text; the frozen base serves the answer. Differentiable Search Index maps queries to identifiers (Tay et al., 2022); indexing through parameters is not a novelty claim. Earlier studies examine streaming re-instatement and log-grounded consolidation (Lin, 2026b;a); this paper examines question-conditioned retrieval and downstream application after parameter writes.

2 Setup and instruments

Protocol. We follow the InMind benchmark's canonical per-task isolation: memory state = a 47-session background corpus (240 user turns indexed) plus one injected fact, a 241-line store per task. Writes are teacher-forced gradient steps on the fact (a `Note` form and a chat-QA form against the task's direct question — benchmark-supplied on InMind and the fresh tasks, self-generated on PrefEval — prefix-masked), $3{\times}10^{-5}$, rank-64 all-linear LoRA on a frozen instruction-tuned 9B chat model (Qwen3.5-9B; Qwen Team, 2026). A step is one gradient update on one line. Under the confirmed recipe the background is written too (41 pre-target and 10 post-target one-step `Note` writes, §3). Learning rate, rank, and k=6 were selected in the exploratory chains; the frozen-recipe transfers price that selection too. Two per-fact doses recur throughout, named by the readout tier they reach: the *certification dose* ($\sim$2 steps, at which the written line's likelihood gain crosses a fixed threshold, $\Delta\ell \geq 0.05$ nats; §4 prices that threshold's specificity) and the *recall dose* ($\sim$6–8 steps, at which the persona-prompted direct question — the model told it may know the user personally — is answered from weights). Serving is always cold: the frozen base generates every user-facing answer from retrieved log lines. Vocabulary: a *chain* is one pre-registered experimental unit; an *arm* one pipeline configuration; the *gold* is a task's injected fact and the store line carrying it (*gold-free* controls remove that write); *chatter* is the background stream, everything but the gold; a line's *register* is its surface voice (the store's lines are first-person conversational); a *front end* produces the retrieval query, and the *statement* front end states one best-guess sentence about the user and embeds it; a *hop question* is the direct memory-check question generated from the indirect one; the write's *pointer* is a keyword elicited from the adapter to cue retrieval; a retrieval gain *cashes* when it survives to the judged answer.

Instruments. Retrieval is scored as gold-in-top-k, k=6 (agentic arms credit gold in any round's top five), over the 241-line store (BGE-small embeddings (Xiao et al., 2024)). Historical end-to-end scores use the benchmark's own application rubric verbatim, judged by a strong external model (Claude Opus 5, accessed 2026-08/09; per-item verdicts archived, so judged numbers are re-scorable without the judge). Each run's answers are judged in independent shards (two to six per run, fixed before judging, arm labels hidden; rubric and judge model fixed within this historical protocol), so no verdict conditions on another shard's answers. Judge noise is anchored at one verdict flip per 90 byte-identical answers (repeatability, not judge validity or human agreement). The 2026-09 repair runs use a separate locked protocol (J2: local Qwen3.8-27B, same rubric verbatim, temperature 0, schema-constrained JSON, incomplete batches rejected rather than scored zero; App. D). J2, the historical Claude protocol, and the earlier zero-coded InMind Qwen re-score are distinct; numbers are compared only within one protocol. We write IA@k for *indirect application* scored end-to-end: the frozen base answers from the top-k retrieved lines and the answer is judged against the rubric ("Application" in tables is IA@6 unless marked; inline pairs written x/y are Retrieval@6/IA@6 unless marked). Two in-context ceilings recur: the benchmark's published 0.84, measured on its backbone

Arm (all serve cold)	Write role	Retrieval@6	Application
Published best external system†	—	—	0.144
Mem0 (vanilla / genre-query configuration)	none	0.168 / 0.424	0.024 / 0.208
Frozen keyword front end (ours)	none	0.152	0.144
Agentic retry loop ($\leq$ 4 searches)	none	0.216	0.176
Per-task write, keyword elicited	pointer	0.216	0.184
Cued agentic (adapter cue)	pointer	0.392	0.272
Cued agentic (*frozen*-cue control)	none	0.360	0.280
Frozen user-voice question embed	none	0.384	0.312
+ cross-encoder rerank of its top-20	none	0.352	0.280
Adapter hop-answer embed	query gen.	0.464	0.360
Real-stream statement embed (exploratory, 1 run)	query gen.	0.576	0.448
Real-stream controls (no writes / chatter-only)	no gold	0.080	0.080
Confirmatory rerun (pre-registered; 3 seeds + one identical-seed repeat):			
seed 42 / 42-repeat	query gen.	0.496 / 0.528	0.392 / 0.400
seed 1234	query gen.	0.560	0.424
seed 777	query gen.	0.600	0.488
number of record (median seed)			**0.424**
One-shot fresh-task transfer (n=100 generated tasks, pre-registered):			
real-stream statement embed (seed 42)	query gen.	0.550	0.430
frozen user-voice question embed	none	0.280	0.230
gold-free control (n=25) / in-context ceiling	no gold / —	0.040 / —	0.040 / 0.870
Coherent accumulation (20 personas, 146 queries; seed-42 recall proxy 0.705):			
written statement / frozen question (seed 42)	query gen. / none	0.192 / 0.178	not judged
hybrid written / forced-guess frozen (seed 43)	query gen. / none	0.274 / 0.240	not judged

Table 1: The canonical program, then its confirmation. Every write arm has a frozen instrument control; n=125 unless marked; every row above the confirmatory block is a single run (with a demonstrated seed spread of $-0.03/+0.06$ on the top pipeline, 0.05-level orderings among the middle rungs should not be over-read); the agentic rows' Retrieval@6 credits gold in any round's top five (§2), a laxer multi-shot criterion. Per-seed exact McNemar against the frozen question front end (0.312; the strongest frozen arm tested at that time; reused without a rerun; greedy decoding does not exclude hardware variation) gives 28:18 (p=0.18), 27:13 (p=0.039), 37:15 (p=0.003) — significant on the pre-registered $2/3$ criterion; a single pooled test (per-task majority vote across the three distinct seeds, post hoc) gives IA 0.432, 31:16, p=0.040. The identical-seed repeat measures one instance of hardware nondeterminism: $93/125$ generated statements differ, yet the aggregate moves 0.032 retrieval / 0.008 application; a later same-seed rerun moved more (Limits). †Quoted from the benchmark (different backbone and judge): directional only. Mem0 stores were independently rebuilt (124/125 memory counts differ); this is not a query-only intervention (§7). Its 0.424 retrieval matching the IA number of record is coincidence.

and judge, and the 0.80 we measure on our own backbone and judge; headline fractions use the published ceiling unless marked. The discriminative readout is $\Delta\ell$: mean per-token log-probability shift of a candidate line, adapter versus pre-write base — a reference-calibrated membership statistic (Carlini et al., 2022; Duan et al., 2024) repurposed as a read head, and never treated as recall: the gap between the index and what can be emitted is the object of study (Song et al., 2026). All arms return few lines, so context dilution is tested only in this setup; prior work reports both distraction and benefit (Jin et al., 2025; Cuconasu et al., 2024). Run-to-run variation from hardware nondeterminism is reported where it is measured (Table 1 caption; Limits).

3 The climb, confirmed

Table 1 compresses the exploratory program (thirteen pre-registered chains, App. A); this section reports its attribution results and the confirmatory rerun that sets the number of record.

The retry budget subsumes the pointer. The adapter's elicited keyword cues an agentic searcher to 0.272, but the same loop cued with the frozen base's keyword scores 0.280; across a budget sweep the cue advantage is an inverted-U, within noise from budget 4 on (Fig. 4): answer-side churn over the shared hits cancels its persistent retrieval edge end-to-end (Lin et al., 2025).

Register match, not machinery. Embedding a generated first-person memory-check question ("Did I tell you about…?") retrieves at 0.384 with no training anywhere — register match with the log's first-

person lines, the HyDE/HyPE mechanism (Gao et al., 2023; Vake et al., 2025) at query time. Under isolation it beats every trained front end we built except the adapter's *answer* to the same question ($0.464/0.360$, $+0.048$, paired $24{:}18$, one-sided sign test $p{=}0.22$ — underpowered), and embedding the frozen base's own answer collapses into refusals (0.080): it was the strongest frozen comparator tested at that point. The repair's explicit hypothetical key later retrieves slightly above a regenerated question reference (0.408 vs 0.384; App. D), and other frozen alternatives remain untested (Limits).

Single-fact readout under a background write stream. Writing 41 background turns before the fact and ten after it (a partial background stream; schedule effects confounded, §4) lifts the simplest pipeline — state your best guess about the user, embed the sentence, serve cold — to $0.576/0.448$ in the exploratory run. Two gold-free controls (no writes; chatter-only writes) collapse to 0.080 with byte-identical refusals — a limitation of the unwritten statement form, not of frozen retrieval, with obedience and ignorance confounded by the persona's refusal instruction (App. A, step 10). That the write supplies content shows in the confirmatory runs, where statements containing the fact's words (0.58–0.65 of tasks, the harness's own stem rule) retrieve at 0.75–0.83 against 0.10–0.18. Because that exploratory number was the maximum over many evaluated arms, we froze the pipeline and reran it fresh under a pre-registered protocol (three seeds plus an identical-seed repeat, lines fixed before data; the seed varies torch-level randomness only, decoding stays greedy): the **number of record is** 0.424 (Wilson 95% CI $[0.34, 0.51]$) — 0.50 of the benchmark's published in-context ceiling (0.53 of the 0.80 we measure on our own backbone and judge) — with the exploratory 0.448 inside the seed spread $[0.392, 0.488]$ (which does not rule out selection bias) and the pre-registered retrieval-side falsification line untriggered (spread 0.496–0.600 against the frozen front end's 0.384). Of the headline, 0.312 is available training-free; the write buys the last $+0.112$ (median seed; 95% CI $[0.02, 0.21]$), at the cost accounted in §7. The historical second-judge check is incomplete and leaves the two-of-three criterion unresolved (Limits; App. D).

The margin survives tasks nobody saw. Confirmation still conditions on thirteen chains of design iteration over the same 125 tasks, which seed resampling does not touch; a pre-registered one-shot transfer prices that. We generated 100 new tasks to the benchmark's recipe under a frozen, hashed generation prompt, with a pipeline-blind overlap gate and a ceiling measured before any arm ran (App. A, step 14), then ran the frozen recipes once, one seed per arm, no tuning. The write pipeline lands at 0.550/0.430 and the frozen question front end at 0.280/0.230 — paired $33{:}13$, one-sided $p{=}0.0023$ (margin $+0.20$, 95% CI $[0.07, 0.33]$). The gold-free control collapses to 0.040; against the fresh in-context ceiling of 0.870 the pipeline sits at 0.494, reproducing the headline fraction to within a point. The pipeline transfers at its development level (0.430 vs 0.424) while the frozen comparator transfers *below* its own (0.230 vs 0.312), so part of the widened margin is comparator weakness on the new tasks — which were generated by the assisting AI that also judges (correlation disclosed; Limits). A second, independently authored transfer removes the generator side of that correlation: on PrefEval (Zhao et al., 2025) — 125 tasks drawn by fixed seed from that benchmark's human-curated preference-following pool and scored by its own four-template rubric verbatim — the same frozen recipes land at 0.904 vs 0.560 (write pipeline vs frozen question front end), paired $48{:}5$, one-sided $p{=}3.5 \times 10^{-10}$ (margin $+0.34$, 95% CI $[0.25, 0.44]$). Its anchors: in-context ceiling 0.976 ($n{=}125$), gold-free floor 0.320 ($n{=}25$, accidental compliance at 0.000 retrieval). Explicit preference application is an easier register than implicit association — 0.93 of ceiling, not 0.50 — so absolute levels compare only within a benchmark; what transfers is the margin, on tasks this project never authored (provenance caveat: Limits), over the frozen question front end only.

The edge concentrates at small k. The write's retrieval edge concentrates at small k (rank curves, post hoc on the exploratory artifacts: written statement 0.520 vs frozen question 0.248 at $k{=}1$, $2.1\times$, narrowing to $1.3\times$ by $k{=}20$; the written front end at $k{=}1$ out-retrieves the frozen one at $k{=}20$). The confirmatory protocol carried a pre-registered precision line: IA@1 within 0.06 of IA@6, falsified if it fell more than 0.12 *below*. The trigger never fired; the observed gaps run $+0.016$ to $+0.064$, and post-hoc paired tests put IA@1 above IA@6 in all four runs, significantly in two (per-run counts: App. A, step 11). The $k{=}1$ gain is *precision-conditional*, not generic: in the pre-registered tuned-frozen-stack arm (§7) serving one line *hurts* (paired $2{:}15$, $p{=}0.0023$) — the opposite sign. On the fresh transfer tasks and on PrefEval the two operating points show no detected difference for the write pipeline (IA@1 0.420 vs 0.430, $7{:}8$; 0.896 vs 0.904, $4{:}5$); those discordant counts are too small for an equivalence claim, and the k-by-pipeline interaction was never tested directly. On the development tasks the write buys precision, not recall; the number of record stays IA@6, and $k{=}1$ is a candidate operating point with neither non-inferiority nor safety established.

Primary contrast ($n = 125$)	A	Control	Difference [95% CI]	Holm p
E2: A vs frozen HYP1, IA@6	0.432	0.328	0.104 [0.000, 0.208]	0.123
E3: correct vs absent, retrieval@6	0.584	0.496	0.088 [0.008, 0.176]	0.123

Table 2: Prespecified repair on previously-seen development tasks, with fresh seed-42 adapters and complete J2 judging of both reader arms. A: the confirmed write pipeline's own query front end; HYP1: frozen hypothetical key; absent: same adapter, question slot emptied. CIs are marginal paired-bootstrap intervals, not inversions of the tests; Table 1's historical Claude scores use a different judge.

Two prespecified repairs, neither conclusive. Two gaps above — no explicit frozen hypothetical-key comparator, and no query ablation separating query-conditioned selection from target salience — were addressed in one locked diagnostic on the same 125 tasks, with fresh seed-42 adapters and matched key budgets (Table 2; App. D). Neither prespecified endpoint reaches significance, before or after Holm correction: IA@6 0.432 versus 0.328 for the frozen hypothetical front end (+0.104, 28:15, exact two-sided p=0.066, Holm 0.123), and gold-in-top-6 0.584 versus 0.496 for the absent query (+0.088, 20:9, p=0.061, Holm 0.123). Inference follows the prespecified exact tests; the bootstrap intervals are marginal, and neither null establishes equivalence. Secondary retrieval contrasts favor A over HYP1 (+0.176, p=0.005) and correct over permuted (+0.168, p=0.0005) but cannot replace the primaries; the absent query's 0.496 is consistent with target salience without proving it, and permuted donors are not a reviewed irrelevant-query control. A post-feedback anchor reads the frozen question reference at 0.320 under J2, level with HYP1 (App. D).

Accumulation shows no clear advantage. The results above isolate one target fact per adapter. On 20 coherent LoCoMo personas, with facts written in dialogue order and intervening turns included, statement retrieval is $28/146$=0.192 versus $26/146$=0.178 for frozen questions (paired 16:14, descriptive task-level p=0.86). The direct-recall stem-hit proxy is 0.705; a cross-fact stem heuristic matches other facts of the same persona on 0.47 of queries. Both use five-character stem matches, not semantic correctness judgments; no end-to-end IA was judged (Table 1). These tasks cluster within personas and source conversations, so the pooled test is not a persona-level confirmation. A question-plus-statement hybrid reaches 0.253 retrieval (post hoc, App. A, step 20), and a separately trained seed-43 follow-up of matched-generation hybrids (0.274 written versus 0.240 frozen forced guess, 22:17, pre-registered one-sided p=0.26) fails its co-primary criterion; it is a new run, not a same-adapter ablation of the seed-42 result (App. A, step 21).

Accumulated-state question conditioning. A separately locked seed-43 run trains one adapter for each of the same 20 personas, then probes all 146 tasks with correct, absent, and within-persona deranged queries at the identical final state. The prespecified equal-persona correct−absent difference is 0.147 (two-sided sign-flip p=0.0078 over 10 source-conversation clusters); task-weighted retrieval is $31/146$=0.212 versus $12/146$=0.082, a different estimand. Deranged queries retrieve $9/146$=0.062 (descriptive). No observation is excluded. This supports question conditioning within accumulated trained states, under the test's independence and sign-symmetry assumptions, rather than a randomized-treatment causal claim. Absolute retrieval remains low; this run has neither a frozen comparison nor end-to-end IA judging. The previously-seen cohort, fixed write recipe, manipulation checks and exploratory few-cluster interval are detailed in App. D.

4 The certification–retrieval dose wall

The remaining sections characterize the write itself. Exposure-dependent recognition-before-recall ordering is known for fine-tuned memories (Orhan, 2023); here we evaluate the likelihood signal as a retriever ranked against a full store under sequential writes. At the certification dose the $\Delta\ell$ threshold is met on every task — by construction, it is the stopping rule — and nothing else works. Under the stream the threshold has no specificity of its own: the gold's $\Delta\ell$ (median 0.41 nats) sits below the best *never-written* same-register line (median 1.88) on $125/125$ tasks, and still on $68/125$ at the recall dose. A chatter-only null (n=25) prices the threshold: after the 41 pre-gold chatter writes the *unwritten* gold already passes 0.05 on $15/25$, yet the two-step write lifts the same line by +0.34 nats (median; $25/25$ above its unwritten value, two-sided $p=6\times10^{-8}$; AUC 0.976), so a calibrated certificate would separate written from unwritten gold in-stream (post hoc) while never-written lines still outrank it: register spillover lifts every same-register line above the two-step needle

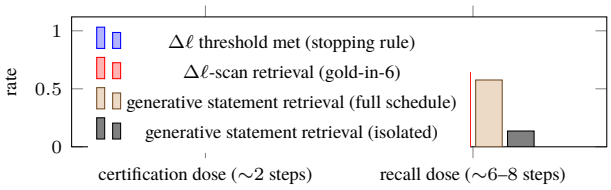

Figure 1: The certification–retrieval comparison across unmatched write schedules (n=125, single runs; the generative bar's 0.576 is the exploratory value of Table 1, confirmed replications 0.496–0.600). At ∼2 steps the threshold is met by construction, yet $\Delta\ell$ as a *retriever* ranks the just-written fact near the bottom of the store (236/241 median) and the generative route sits at the frozen floor (isolation factorial, 0.000 echo-gated). The recall-dose generative bar comes from a schedule with ten post-gold writes, the scan bars from one without (§4). The isolated-schedule generative bar is plotted alongside: 0.000 to 0.576 changes both dose and schedule, 0.000 to 0.136 changes dose alone.

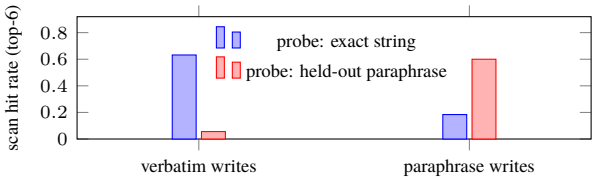

Figure 2: The write-format key inverts: under verbatim writes the likelihood scan retrieves the exact string and not its paraphrase; paraphrase writes flip both bars (§5).

(App. B: the same separation under isolated writes). The generative route is dead in every cell of a 2×2 probe factorial (App. A, steps 5–6). At the recall dose both routes open (generative statement 0.576; verbatim scan 0.640, median rank 2; Fig. 1). These dose-axis numbers are single runs — exploratory; the confirmatory protocol never reran the certification-dose arm. There is no cheap discriminative shortcut: the two stopping rules are separated by a ∼3× dose wall, and below it the index is not weak but *actively wrong*.

Form-sensitive readout and an unresolved schedule effect. Between the doses sits an access pathology: a fact written to the recall criterion can answer its trained question while off-form prompts ("state what you remember") fall into the learned-default refusal (Kang et al., 2025), and re-voicing the probe *as the user would ask it* unlocks genuine recall from 3.2% to 32.8% (not the reversal curse (Berglund et al., 2024): argument order never changes, only the speech-act register). A post-hoc comparison across three write schedules finds better readout in the schedule with subsequent writes: probed immediately after its own writes a fact is locked (retrieval 0.136 in isolation), while a full schedule adding 41 pre-gold and 10 post-gold writes reaches 0.576 retrieval with refusals falling 0.71 → 0.12 (App. C). The comparison is confounded: it also changes pre-gold exposure, the effective additional dose is unmeasured, the schedules differ in total steps and in the last-written form, and an instructed-refusal reading (Qi et al., 2024; Zhang et al., 2025a) is not excluded, so whether content, dose, or last-written form releases the lock is untested. Prior work improves extraction through fact augmentation (Allen-Zhu & Li, 2024), joint storage/query-format training (Zhang et al., 2025b), and meaningful prompts after rote memorization (Wu et al., 2026); mixed data instead protects retention (Chen et al., 2026). Our schedule association motivates a matched-dose, matched-final-form test and does not establish that later unrelated writes cause recovery.

5 The write-format key

What is the index keyed on? Under verbatim writes, the recall-dose scan retrieves the exact string (0.632; replicated at 0.640 by the independent dose-wall run of §4 and at 0.648 by a rerun; the paraphrase-write regime below replicated once, 0.176/0.624) and nothing else: a held-out paraphrase scores 0.056 (Fig. 2). Writing a fresh paraphrase each round *inverts* the key: the held-out paraphrase rises to 0.600 while the original string — still trained, via the chat-QA form — falls to

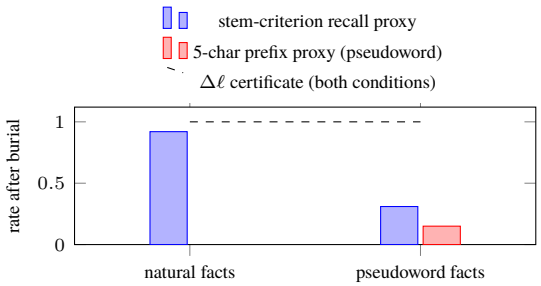

Figure 3: Natural and pseudoword burial outcomes, same write harness (n=125 per condition; primary bars share the stem criterion; burial registers differ). Forty one-step writes leave natural-fact recall at 0.92–0.95 (chatter burial; lower of two write regimes shown) but pseudoword recall under same-register burial at 0.31 (0.15 with the stricter 5-character prefix, 0.10 full value), while every likelihood certificate survives in both conditions (dashed line) with pseudoword margins *rising* from 3.8 to 5.2 nats through the burial (§6).

0.184. A family-level control separates content from style: under paraphrase writes, the held-out paraphrase of the *written* fact gains $\Delta\ell = 1.77$ (mean) while style-matched paraphrases of five *other, never-written* facts — same generator, same register — gain only 0.09; the content lift (1.68) is positive on 125/125 tasks. The key follows the surface family the writes exercised, not the exact string; whether it sits on meaning or on lexical overlap with the (unlogged) trained paraphrases is untested. A reversal probe on the same adapters scores 0.08–0.16 against forward recall of 123/125, but 21 answers (verbatim regime) restating the fact are scored as misses and 67 are instructed refusals, so it prices the scorer as much as the argument-order axis (Berglund et al., 2024). Against fact augmentation as an extraction enabler (Allen-Zhu & Li, 2024), this bidirectional inversion measures a tradeoff between surface families in a likelihood index. The result is limited to that index; paraphrase writes were not tested through statement generation, embedding retrieval, or judged application.

6 Content-dependent failure under unmatched burial conditions

Forty subsequent one-step chatter writes leave high recall after an eight-step *natural-language* fact (0.92–0.95 stem-criterion recall at n=125 across both write regimes, certificates 1.00 with rising margins); an n=8 smoke test with same-register fact burial points the same way. Rerunning the harness with unguessable pseudoword facts ("My transit code is roptavvor"; burial now same-register pseudoword writes) collapses recall to 0.15 (5-character prefix; 0.31 stem, the like-for-like criterion; 0.10 full value) while every certificate survives with *rising* margins and the scan still returns the fact in the top six (Fig. 3). Retention under sequential writes is known to depend on presentation breadth and provenance (O'Neill, 2026; Chen et al., 2026); *lexical/entity familiarity* is a candidate explanation this comparison does not isolate. Burial register (the natural-under-same-register cell is the n=8 smoke), criterion, and tokenization also differ between the cells, so the content attribution is a direction until the natural-value-in-template cell is run (cf. Gekhman et al., 2024).

Of 125 burial tasks, 68 strict-recall failures still emit a pseudoword-shaped string, all novel *chimeras* but degenerately recombined, while the certificate margin rises — visible to a likelihood read head, invisible to accuracy alone (syllable decomposition, sampling baselines, and the templated-line caveat: App. C).

7 Design considerations, limits, and cost

The program suggests six bounded design considerations, not validated deployment rules (Limits); (6) closes the extraction-memory paragraph below. (1) *Serve cold, write hot*: every end-to-end gain we measured lived on the query side, and as a reader of identical retrieved lines the adapter loses to the frozen base under the confirmed protocol (IA@6 0.448 vs 0.512, paired 3:11, two-sided p=0.057; IA@1 1:12, p=0.003; one seed, uncorrected) and in the accumulated-store regime alike (App. A, step 18); self-recitation consolidation lands below doing nothing (Table 3). (2) *The index supplies fact-*

bearing query text under isolation: its observed value is eliciting words about the written fact, and the empty-query ablation (0.496 against 0.584, §3) leaves query-independent target salience unseparated from query-conditioned selection; the keyword pointer prices at zero once an agentic loop may re-query, so the gain matters most where budgets bind ($2.1\times$ at $k{=}1$, post hoc). (3) *Certification is not retrieval*: a calibrated likelihood lift at two steps certifies presence, not findability; an uncalibrated one certifies neither (§4), so budget the recall dose if the weights must be searched (reached *across* turns by one-step writes plus miss-gated replay, Table 3). (4) *The write format chooses the key*: paraphrase writes re-key the likelihood index past the exact string, at the verbatim key's cost, untested end-to-end (§5). (5) *Test write schedules at matched dose and final form*: the observed schedule association is not yet a causal rule (§4).

The frozen stack, tuned. The margin behind considerations (1)–(2) is priced against a frozen alternative given the obvious upgrades under pre-registration: a cross-encoder reranker (bge-reranker-v2-m3 (Chen et al., 2024)) over the frozen question's top-20 lowers retrieval to $0.216/0.352$ at $k{=}1/k{=}6$ (from $0.248/0.384$) and trails the untuned stack end-to-end (0.280 vs 0.312 at $k{=}6$; paired $8{:}12$, $p{=}0.50$); a stronger embedder (BGE-large) retrieves at $0.232/0.312$ on the same two operating points. Neither upgrade helped; coherent hybrid controls are in App. A, step 21, and the frozen alternatives still untested under per-task isolation are listed in Limits.

Cost. Workloads are unmatched across architectures. Per task, the repair run spends about 28 s in gradient steps and 6 s in gate generations; Mem0 ingestion takes $\sim$$1650$ s per task under shared service load, with 0.14 s per search (detail: App. D). These are descriptive timings, not a speed comparison.

Extraction memory: configuration versus query intervention. The original Mem0 comparison reaches 0.024 versus 0.208 application, but its arms independently rebuilt their stores (124/125 memory counts differ): a configuration contrast, not a query-only effect (App. A, step 12). The extracted store also trails the raw log under the same query and embedder (0.208 vs 0.312, $2{:}15$, $p{=}0.0023$); content loss and register mismatch remain possible, unisolated explanations.

A pre-specified repair holds the original vanilla snapshot fixed for all 125 tasks (content and vector hashes unchanged before and after both queries), regenerates both answer arms with the same cold 9B reader, and grades them blind under J2 (App. D). Genre-matched querying raises IA@6 from 0.024 to 0.232 ($26{:}0$, exact two-sided $p{=}2.98\times10^{-8}$, paired-bootstrap 95% CI $[0.136, 0.280]$). This supports the query intervention on that fixed snapshot; the new-judge values are not interchangeable with the original Claude scores. Consideration (6) remains: *compare raw and extracted stores within each query regime*; neither architecture is established as generally preferable.

A small-sample reversal. A pre-registered LoCoMo falsification test (Maharana et al., 2024) fired its descriptive $+0.05$ trigger for the extracted store (0.475 vs 0.400, $n{=}40$, $p{=}0.58$; 0.800 vs 0.500, $n{=}10$, $p{=}0.25$) without establishing a general explicit-recall advantage (detail: App. D).

Limits. The deepest validity threat was task-set reuse: the pipeline was designed over thirteen chains evaluated on the same 125 tasks, and the confirmatory rerun assesses seed variation after selection; it does not remove adaptive overfitting to the task set. The pre-registered fresh-task transfer (§3) prices it, with its own disclosed residue: the 100 new tasks were generated by the assisting AI to the benchmark's recipe and judged by a sibling model (their ceiling, 0.870, runs slightly above the benchmark's own); disclosure is not independence. The PrefEval transfer (§3) closes the generator side, but the judge applying that benchmark's rubric is still a Claude-family model (PrefEval's tasks were themselves Claude-assisted before curation). A second-family judge pass prices that residual: Qwen re-applying the same four templates to all 650 answers gives recorded scores 0.912 vs 0.520 (paired $52{:}3$; item-level agreement 0.86–0.96). Raw judge responses were not archived, so missing-tag parsing cannot be audited (App. D); it also shares the *serving* backbone's family, so the two judges are differently correlated rather than uncorrelated. The InMind Qwen re-score incorrectly assigned zero to 28 parsing failures among 875 responses. Across all binary completions of the missing labels, all three margins remain positive but one to three seeds pass the two-sided test; this is not a complete replication (App. D). Human validation remains undone. The main results use one 9B substrate and one embedder: a one-shot second-backbone run (Mistral-7B, App. A step 17) reproduces the direction of the margin and the precision effect, not the IA@6 margin's significance ($12{:}5$, one-sided $p{=}0.07$). The frozen alternative's two obvious tunings were tested and hurt (§7), and an explicit hypothetical-key front end was added by the repair (§3) on the development tasks only

(the transfer margins are over the frozen question front end); multi-query fusion and ingestion-side question indexing remain untested under per-task isolation, so the margin is over tested frozen configurations. The isolated InMind comparisons follow the benchmark's per-task isolation (one target fact per adapter, with background in the store). The target receives direct-QA supervision and roughly eight updates against one per background turn. The repair's absent- and permuted-query arms probe that confound on the isolated adapters without resolving it; no reviewed irrelevant-query arm was run, and frozen indexing of the same direct-QA supervision has not been compared. The corresponding coherent same-adapter test detects question conditioning, but has no frozen-arm comparison and only ten source-conversation clusters. Two existing one-shot accumulation streams bound that question: all 125 mutually inconsistent facts as one persona collapse *both* front ends (0.08 written vs 0.12 frozen retrieval@6) while the direct-recall proxy holds 0.53 — a selection failure despite residual recall (step 19); on the 20 coherent personas the written edge is a direction only (§3). The query-side gain is demonstrated per fact, without isolating selection from salience. Published baselines in Table 1 are quoted, not rerun (Mem0 excepted); cross-setup deltas are directional, never leaderboard positions. Run-to-run variation exceeds the identical-seed pair's $\sim$0.01–0.03: a later seed-42 rerun (App. A, step 18) landed at IA@6 0.512, above the confirmed spread, so single-run deltas of that size elsewhere (e.g. the budget-sweep endpoints) are directions, and the number of record stays the pre-registered median. The $2/3$ criterion is conditional on one frozen comparator run: its unchanged per-item verdict vector enters all three tests. The seeds reuse the same 125 tasks and are correlated. Exact McNemar treats the 125 pairs as independent though they share one corpus and judge, and no family-wise correction was applied to those historical seed tests; we report the specified seed criterion and individual transfer tests without claiming family-wise control across the program. Conventions: InMind McNemar p's are two-sided exact binomial on the discordant pairs; the transfer tests are one-sided, as pre-registered; other historical p values are descriptive post-hoc diagnostics, two-sided unless marked; paired-difference CIs are Wald intervals unless otherwise specified. The confirmatory design was underpowered for its own median effect: with $\sim$35% discordant pairs at n=125 the minimal detectable paired difference at 80% power is about 0.15, against observed $+0.08/+0.11/+0.18$. The pooled post-hoc margin is $+0.12$, 95% CI $[0.01, 0.23]$ — the transfers provide additional task-level evidence, but no claim is made that they survive arbitrary multiplicity corrections. The content comparison uses unmatched burial conditions; and the 0.50 headline is a cross-setup ratio (our number over the benchmark's published ceiling) — it does not transfer across content classes, judges, or backbones.

8 Conclusion

Under per-task isolation the write buys fact-bearing retrieval text: InMind 0.424 versus 0.312 for its frozen question comparator, with two one-shot transfers retaining the margin. Two boundaries qualify it. A prespecified repair — a same-inference-budget frozen hypothetical front end, and the same adapters queried with an empty slot — leaves both primary tests nonsignificant before and after correction; query-independent target salience remains unresolved under isolation. Under accumulation, historical runs show no clear written-versus-frozen advantage; the new same-adapter control supports question conditioning, at low absolute retrieval and with no frozen arm. The replicated write-format inversion and exploratory readout failures characterize access to written information, with schedule and content mechanisms unresolved. Under isolation the update supplies query-side text about what was written; making that selectable and applicable as it accumulates is the open problem. These results do not demonstrate a superior deployed memory system.

Ethics statement

This work writes personal disclosures into model weights — the memory regime most likely to hold sensitive user data. The experiments use external benchmark data and the synthetic tasks described above; no deployment user logs were collected for this study. Two properties of the studied architecture bear on deployment ethics, in opposite directions: per-user adapters are deletable as units (removal of the cartridge and log is data deletion, unlike facts absorbed into shared weights), but the likelihood signal cuts both ways — calibrated against the stream's baseline (§4), personal content is *detectable* in weights long before it is extractable, a privacy audit tool and an inference-attack surface at once. The obvious misuse is covert profiling — writing *inferred* rather than disclosed facts

without telling the user. Deployed systems of this kind need consent for writing, disclosure of what is written, per-user isolation of adapters, and deletion guarantees covering both store and log.

Reproducibility statement

The anonymized artifact bundle in the supplementary material contains the full pre-registration ledger (per-chain falsification lines, amendment history, and the voided-instrument audits), the harness scripts with each arm's environment settings and seeds, historical raw per-task result files, per-item judge verdicts under the benchmark's rubric, the 100 generated fresh tasks with the hashed generation prompt, the fixed-seed draw lists for the PrefEval and LoCoMo slices, and a README mapping each table and figure to the script and config that produces it. Backbone (Qwen3.5-9B), embedder (BGE-small), and judge protocol are named in §2; judged numbers are re-scorable from the archived per-item verdicts, and re-judgeable only up to judge availability. Gradient-based runs use a single consumer GPU (AMD Radeon RX 7900 XTX, 24 GB, ROCm; PyTorch 2.5.1, PEFT 0.19.1); separate reader and judge service runtimes are recorded in the repair artifacts. External assets and licenses: Mem0 (version 2.0.19), InMind (public repository, license undeclared at the time of writing), PrefEval and LoCoMo (both CC BY-NC 4.0, used for non-commercial research evaluation only); no benchmark text is redistributed with the bundle — per-item files carry task ids, retrieval ranks, judge verdicts, and model outputs only.

AI assistance disclosure

Large language models (Claude, Anthropic; Codex, OpenAI for the present revision) were used under the authors' direction, and the authors bear responsibility for all claims. The present revision's experimental protocols were drafted and locked by Codex before execution under the author's authorization. Following the ICLR 2027 author policy we disclose each use. *Synthetic data:* the 100 fresh InMind-style tasks (App. A, step 14) and the 146 persona tasks of step 20 were generated by an LLM from frozen, hashed prompts and gated by scripts (the persona tasks additionally audited by an independent LLM pass); the 125 development tasks, PrefEval and LoCoMo are external. *Hypotheses and methodology:* chain designs, falsification lines, and amendments were drafted by the LLM and approved or redirected by the author(s); several of the paper's corrections (voided instruments, the criterion mislabel of Fig. 3, the non-specificity of the certificate) originate in LLM audit passes commissioned by the author(s). *Implementation:* the harness scripts, judge sharding, and analysis code were written with LLM assistance and executed on author-controlled hardware; Codex operated the repair experiments under author authorization. *Automated pre-submission feedback:* the Paper Assistant Tool offered by ICLR 2027 reviewed the revised manuscript once; its comments led to wording, unit, and cross-reference clarifications and to the frozen question reference's application anchor (App. D); no claim was added on its advice. *Evaluation:* end-to-end answers were judged by an LLM applying the benchmarks' own rubrics verbatim (§2); a second-family LLM judge and the archived per-item verdicts are provided as checks; no human validation slice exists yet (Limits). *Literature:* related work was located by LLM-assisted web search and verified against the cited sources in the assisted audit; final author verification remains required. *Interpretation and writing:* results were interpreted jointly; the manuscript was drafted and edited by the LLM and revised by the author(s); reviewer-style audits of the draft were also LLM-run. The supplement supports arithmetic recomputation from archived outputs and judge labels; this does not independently validate the labels. Full experimental re-enactment also requires external datasets and private intermediate artifacts not redistributed here. Earlier numerical audit passes did not detect the separately rebuilt Mem0 stores. The present revision corrects that experimental-definition error and distinguishes arithmetic verification from checks of experimental controls; human validation of the revised claims remains required.

References

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.1, knowledge storage and extraction. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2309.14316.

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, 2025. arXiv:2501.00663.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: LLMs trained on "A is B" fail to learn "B is A". In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.12288.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914, 2022. doi: 10.1109/SP46214.2022.9833649. arXiv:2112.03570.

Howard Chen, Jiayi Geng, Adithya Bhaskar, Dan Friedman, and Danqi Chen. Continual memorization of factoids in language models. *Transactions on Machine Learning Research*, 2026. arXiv:2411.07175.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335, 2024. doi: 10.18653/v1/2024.findings-acl.137. arXiv:2402.03216.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025. Preprint; no venue assigned as of 2026-09-05.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. The power of noise: Redefining retrieval for RAG systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 719–729, 2024. doi: 10.1145/3626772.3657834. arXiv:2401.14887.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models? In *Conference on Language Modeling (COLM)*, 2024. arXiv:2402.07841.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers*, pp. 1762–1777, 2023. doi: 10.18653/v1/2023.acl-long.99.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. Does fine-tuning LLMs on new knowledge encourage hallucinations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7765–7784, 2024. doi: 10.18653/v1/2024.emnlp-main.444. arXiv:2405.05904.

Bowen Jin, Jinsung Yoon, Jiawei Han, and Sercan O. Arik. Long-context LLMs meet RAG: Overcoming challenges for long inputs in RAG. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.05983.

Katie Kang, Eric Wallace, Claire Tomlin, Aviral Kumar, and Sergey Levine. Unfamiliar finetuning examples control how language models hallucinate. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), Volume 1: Long Papers*, pp. 3600–3612, 2025. doi: 10.18653/v1/2025.naacl-long.183. arXiv:2403.05612.

Ruizhe Li, Mingxuan Du, Benfeng Xu, and Zhendong Mao. Keep it InMind: Benchmarking the implicit-association blind spot in agent memory, 2026a. arXiv:2607.24368.

Wenzhi Li, Dong Nie, Rui Lan, Tongtong Lyu, Peiyao Wang, Lingzi Hong, Weihang Pan, Binbin Lin, Boyuan Pan, and Yao Hu. Dual-layer agentic memory with fast write routing and slow consolidation. *arXiv preprint arXiv:2608.22215*, 2026b. Preprint; no venue assigned as of 2026-09-05.

Huifeng Lin, Gang Su, Jintao Liang, You Wu, Rui Zhao, and Ziyue Li. Fishing for answers: Exploring one-shot vs. iterative retrieval strategies for retrieval augmented generation. *arXiv preprint arXiv:2509.04820*, 2025. Preprint (under review per the arXiv record); no venue assigned as of 2026-09-05.

Xuhao Lin. The lifelong loop: Error-gated consolidation and the cost of nights without a log, 2026a. URL https://doi.org/10.5281/zenodo.21310038. Preprint, 2026-07-11.

Xuhao Lin. Persistence is not accumulation: Erasure, suppression, and re-instatement in online in-weight memory for language models, 2026b. URL https://doi.org/10.5281/zenodo.21232648. Preprint, 2026-07-07.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers*, pp. 13851–13870, 2024. doi: 10.18653/v1/2024.acl-long.747. arXiv:2402.17753.

Charles O'Neill. Can a language model learn facts continually in its weights? *arXiv preprint arXiv:2607.11020*, 2026. Preprint; no venue assigned as of 2026-09-05.

A. Emin Orhan. Recognition, recall, and retention of few-shot memories in large language models. *arXiv preprint arXiv:2303.17557*, 2023. Preprint; no venue assigned as of 2026-09-05.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Martino M. L. Pulici, Cuong Xuan Chu, Evgeny Kharlamov, and Volker Tresp. A storage-retrieval gap in parametric knowledge graph memory. *arXiv preprint arXiv:2608.25489*, 2026. arXiv:2608.25489; accepted at SKGi 2026 per the arXiv record.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.03693.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Xiangchen Song, Zhenhao Chen, Lingjing Kong, Shaoan Xie, Xinshuai Dong, Guangyi Chen, and Kun Zhang. Beyond perplexity: A behavioral evaluation framework for deployment-memory claims in LLM test-time training. *arXiv preprint arXiv:2607.00368*, 2026. Preprint; no venue assigned as of 2026-09-05.

Weihang Su, Yichen Tang, Qingyao Ai, Junxi Yan, Changyue Wang, Hongning Wang, Ziyi Ye, Yujia Zhou, and Yiqun Liu. Parametric retrieval augmented generation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 1240–1250, 2025. doi: 10.1145/3726302.3729957. arXiv:2501.15915.

Yuqiao Tan, Shizhu He, Huanxuan Liao, Jun Zhao, and Kang Liu. Dynamic parametric retrieval augmented generation for test-time knowledge enhancement. *arXiv preprint arXiv:2503.23895*, 2025. Preprint; no venue assigned as of 2026-09-05.

Yi Tay, Vinh Q. Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, Tal Schuster, William W. Cohen, and Donald Metzler. Transformer memory as a differentiable search index. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2202.06991.

Domen Vake, Jernej Vičič, and Aleksandar Tošić. Bridging the question-answer gap in retrieval-augmented generation: Hypothetical prompt embeddings. *IEEE Access*, 13:129952–129961, 2025. doi: 10.1109/ACCESS.2025.3589499. arXiv:2607.29402.

Qinyuan Wu, Soumi Das, Mahsa Amani, Bishwamittra Ghosh, Mohammad Aflah Khan, Krishna P. Gummadi, and Muhammad Bilal Zafar. Rote learning considered useful: Generalizing over memorized data in LLMs. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2507.21914.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general Chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 641–649, 2024. doi: 10.1145/3626772.3657878. arXiv:2309.07597.

Mengqi Zhang, Xiaotian Ye, Qiang Liu, Pengjie Ren, Shu Wu, and Zhumin Chen. Uncovering overfitting in large language model editing. In *International Conference on Learning Representations (ICLR)*, 2025a. arXiv:2410.07819.

Ying Zhang, Benjamin Heinzerling, Dongyuan Li, and Kentaro Inui. Understanding fact recall in language models: Why two-stage training encourages memorization but mixed training teaches knowledge. *arXiv preprint arXiv:2505.16178*, 2025b. Preprint (v2, May 2026); no venue assigned as of 2026-09-05.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. Do LLMs recognize your preferences? Evaluating personalized preference following in LLMs. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2502.09597.

A The canonical program, chronologically

Table 1 presents the program as a ladder; this appendix presents it as it happened — because the order in which the numbers arrived, and the instruments that had to be voided along the way, are part of the evidence. The narrative steps below group the thirteen exploratory chains thematically rather than one-to-one — a step may span several chains (a sweep and its control, a probe and its factorial, are separate ledger entries) — and steps 11–20 are the ten later pre-registered additions; the step-to-chain mapping is explicit in the ledger. Every chain was pre-registered with falsification lines before its run; probe changes after a falsified target are declared as amendments; diagnostics computed after unblinding are labeled post hoc. The full ledger, exact configs, raw result files, and per-item judge verdicts ship in the anonymized artifact bundle.

1. The frozen baselines, and a calibration tie. Rerunning the training-free arms under the benchmark's canonical per-task isolation flipped the rankings we had measured in a preliminary accumulated-store regime: question-side embedding collapses (0.016 gold-in-4, 0.000 application) against the off-genre 241-line haystack, an answer-draft front end multiplies retrieval $3.5\times$ but drowns absolutely (the gold ranks $110/241$ against a typical advice-style draft), and the survivor is the simplest front end — elicit one keyword from the question with the frozen base, embed it: $0.152/0.144$, matching the published best system's value (14.4%) at its relative-to-ceiling number (0.171). The tie is a coincidence of level under a different backbone and judge, and we use it only as the cross-rig calibration point. The encoder's retrieval-instruction prefix leaves that collapsed question-side probe's scores unchanged, consistent with a semantic rather than a prefix-formatting cause.

2–4. The pointer is priced at zero. An agentic retry loop ($\leq$ 4 searches, frozen everywhere) reaches 0.176, with 20 of 27 hits coming from re-queries. A per-task write with keyword elicitation reaches 0.184, and yields the program's first uncontaminated dose–elicitation gradient (base keyword $0.152 \rightarrow$ certification-tier $0.160 \rightarrow$ recall-tier 0.216 gold-in-6). Cueing the agent with the adapter's keyword reaches 0.272 — but the pre-registered control, the same loop cued with the frozen base's keyword, reaches 0.280: the seeds agree on $92/125$ tasks (90 answers byte-identical; one judge flip across those ninety prices judge noise at $\sim 1/90$), and the budget sweep of Fig. 4 shows the cue advantage is an inverted-U dead from budget 4 on. The write's keyword-level pointer is subsumed by retries.

5–6. A voided probe localizes the form lock. The two-hop probe — frozen base converts the indirect question to a direct memory-check question, adapter answers, the answer is embedded — *failed as first built* (0.184 retrieval), and the failure was an instrument artifact: an echo audit found 19 of its 23 "hits" were lexical echoes of the probe question; genuine at-probe recall was 3.2%. We voided the probe and declared an amendment rather than tuning it silently. The autopsy localized the lock: the generated question addressed the user as "you," off the trained form. Re-voicing the hop question as the user would ask it unlocked genuine recall to 32.8% and retrieval to 0.464. A 2×2 factorial (who writes the question $\times$ who answers it) attributes the effect to the answering pathway, and the certification-tier probe is dead in every cell — the first sighting of the dose wall of §4.

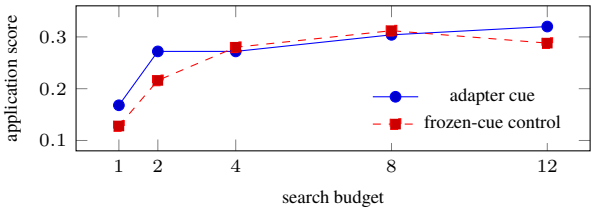

Figure 4: The cued-agentic advantage is an inverted-U in retry budget: $+0.056$ at budget 2, within noise from budget 4 on (mean $+0.006$; the $+0.032$ at budget 12 is single-arm noise); n=125 per point (§3).

7–8. The frozen champion, and a margin we refuse to celebrate. Embedding the frozen base's user-voiced question alone scores $0.384/0.312$ — the frozen top (shared with a budget-8 frozen-cue agentic arm), from register match with the log's first-person lines, no training anywhere. The adapter's hop *answer* scores $0.464/0.360$: $+0.048$ over its own instrument control, consistent across the program but underpowered at n=125 (paired 24:18, one-sided sign test p=0.22) — reported as a direction, never a finding.

9. Realism, and two caveats banked at design time. Writing a 41-turn subset of the task's background stream into the adapter before the fact, and ten further turns after it, revived the statement probe (dead in isolation: 0.136 retrieval, 0.71 refusals) to 0.576, and the simplest pipeline — state your best guess about the user, embed the sentence, serve cold — reached 0.448 application (exploratory; step 11 sets the number of record). Two caveats were banked in the ledger the day the arm was designed: part of the gain could be the index learning the *register* of the very store it retrieves from, and no frozen control existed yet under the real-stream protocol, so the step was first recorded as attributed to neither instrument nor memory.

10. The gold-free control pair. A pre-registered control pair ran the identical statement pipeline with the gold write removed — once with no writes at all, once with the 51 chatter writes but no gold. Both collapse to $0.080/0.080$, the statement a byte-identical "I don't remember" refusal on $125/125$ tasks in each: the write changes this readout, but refusal instructions confound the frozen control. Query-independent salience has not been excluded. The scope is deliberately narrow: frozen *retrieval* did not collapse (frozen question $0.384/0.312$); the controls show failure of this statement prompt without the write. Post-hoc diagnostics on the same artifacts: $+0.136$ over the best frozen front end (32:15 discordant, exact McNemar p=0.019; 35:18, p=0.027 against the tied agentic comparator), and the rank curves of §3 ($2.1\times$ at k=1).

11. The confirmatory rerun sets the number of record. Because 0.448 was the maximum over many evaluated arms, we froze the pipeline and pre-registered a fresh rerun: three seeds plus an identical-seed repeat, all lines fixed before data — falsify if the median seed fails to beat the frozen front end's 0.384 retrieval-side, confirm if the per-seed exact McNemar against 0.312 passes on at least $2/3$ seeds. Result: IA@6 $0.392/0.424/0.488$ (McNemar 28:18 p=0.18; 27:13 p=0.039; 37:15 p=0.003) — confirmed, number of record 0.424, with the exploratory 0.448 inside the seed spread (without ruling out selection bias). The identical-seed pair measures one instance of hardware nondeterminism ($93/125$ statements differ; aggregates move 0.032/0.008), and IA@1 beats IA@6 in all four runs — 8:1 (p=0.039), 10:2 (p=0.039), 10:3 (p=0.092), 9:7 (p=0.80), uncorrected, with the direction consistent in all four and significant in two — the precision claim cashed end-to-end.

12. The on-rig baseline. Finally we reran Mem0 (default extraction pipeline, our backbone and embedder, independently built stores) on the same task inputs. Three comparison lines and their trigger thresholds were pre-registered; the paired test statistics were computed at unblinding. Line 1 (calibration): the vanilla score landed below the pre-registered replication band and below every published external-system score on this benchmark (0.024; the published best across the six quoted systems is 0.144, measured with a stronger backbone) — reported as an out-of-band result, with backbone strength and query-path configuration sensitivity not separable on our rig. Line 2 (planned instrument transfer, trigger $\geq +0.10$): the genre-query configuration scores 0.208 versus 0.024 ($8.7\times$, 25:2, $p<10^{-5}$). Execution instead rebuilt stores independently; 124/125 task memory counts differ. Thus the trigger describes a configuration contrast and cannot identify query-only transfer. Line 3:

the instrumented extraction configuration trails the raw log (0.208 vs 0.312, 2:15, p=0.0023), without separating content loss from register effects. The identical-store repair is now complete (E1, App. D), with newly generated answers and uniform J2 judging; it does not retroactively validate this historical contrast.

13. The frozen stack gets its budget — and spends it badly. The steelman objection holds that the frozen side never got the tuning the write side got. A pre-registered arm gave it the two obvious upgrades: a cross-encoder reranker over the frozen question's top-20 candidates, and a stronger embedder. Both made retrieval *worse* (0.216/0.352 reranked and 0.232/0.312 BGE-large, vs 0.248/0.384 untuned, all retr@1/retr@6; the pre-rerank sanity anchor reproduced the untuned numbers exactly), neither pre-registered trigger fired (retrieval@1 $0.216 \ll 0.480$, the confirmed written front end's retrieval@1 floor; IA@1 0.176 and IA@6 0.280 against the written pipeline's IA@6 floor 0.392), and the reranked stack showed the *inverse* k=1 effect (2:15, p=0.0023) that upgraded the precision claim to a precision-conditional mechanism *hypothesis* (§3).

14. One hundred tasks nobody saw. The last chain answered the last methodological attack — adaptive overfitting to the development tasks. One hundred new tasks were generated to the benchmark's recipe (prompt frozen and hashed before the first call; the 125 existing tasks supplied only as an exclusion list; two paraphrase-guard regenerations, both ledgered), had their in-context ceiling measured pipeline-blind before any arm ran (not used to exclude tasks), and were then hit once by the frozen recipes: write pipeline 0.550/0.430, frozen question 0.280/0.230 (33:13, one-sided p=0.0023), gold-free control 0.040 (n=25), ceiling 0.870. The development-set headline fraction (0.50 of ceiling) reproduces at 0.494; the k=1 advantage does not reproduce (7:8, no detected difference) and stays hypothesis-grade. Every number in this step was one shot: no task was rerun, no recipe adjusted.

15. A rule falsified, on schedule. The program's last pre-registration took rule 6 — index the raw log — to a second, independently authored benchmark (LoCoMo: ten long conversations ingested whole by the same Mem0 pipeline, a fixed-seed 50-question slice (20 multi-hop + 20 temporal + 10 single-hop), raw-log vs extraction-store arms under identical embedder, query, serve, and a blinded judge, plus a no-store floor). The falsification line ($B - A \geq +0.05$ on multi-hop+temporal) fired: extraction edges the raw log there (0.475 vs 0.400; n=40, paired 8:5, two-sided p=0.58) and has a higher single-hop point estimate (0.800 vs 0.500; n=10, 3:0, p=0.25; floor 0.100). The descriptive trigger restricts the raw-log recommendation, but this small slice does not establish a general advantage for extraction in explicit recall. Of this program's falsification lines, this is the one that fired — and reporting it is the point of having them.

16. An independently authored benchmark, its own rubric. The final transfer took the frozen recipes to PrefEval: 125 tasks drawn by fixed seed from its human-curated explicit-preference pool, the pipeline generating its own memory-check question from the topic name alone (the benchmark supplies none). The smoke run caught an echo artifact in the first gate design — an LLM-generated check question carrying the preference's content words fired the recall gate at 2–4 steps, and the underdosed golds ranked 117–240 while every properly dosed gold ranked 1: the dose wall's signature (§4), reproduced uninvited on external data — so the gate was rebuilt leakage-free before the pre-registered run. One shot: retrieval 0.912 vs 0.424 gold-in-6; the benchmark's own four-template preference-following accuracy 0.904 vs 0.560 (paired 48:5, one-sided p=3.5×10^{-10}), in-context ceiling 0.976, gold-free floor 0.320 (n=25) at 0.000 retrieval (accidental compliance, not memory); k=1 is 0.896 versus 0.904 at k=6 (4:5), with no detected difference and no equivalence conclusion. The frozen front end's 0.424 retrieval is numerically coincident with the InMind number of record. Two smoke passes preceded the run (the failing v1 gate and the validating v2), both ledgered; no smoke number is banked anywhere. A pre-registered second-family judge pass (Qwen, temperature 0, the same four templates verbatim, blind to the first judge's verdicts) re-scored all 650 answers: 0.912 (coincident with the retrieval figure) vs 0.520, paired 52:3, one-sided p=7.7×10^{-13}, item-level agreement 0.86–0.96 — the margin does not live in the judge.

17. A second backbone, one shot. A pre-registered single-seed rerun of the frozen recipes on Mistral-7B-Instruct-v0.3 (the ungated non-Qwen instruct model available on the rig; recipe verbatim, echo guard as in step 16) asked whether the margin is a fact about test-time LoRA writes or about Qwen3.5-9B. The recall gate fired on only $29/125$ tasks (Mistral's persona answers refuse rather than guess), so the write ran at the fixed maximum dose on the rest. Retrieval@6 0.280 vs 0.112; blind Opus shards then gave IA@6 0.112 vs 0.056 against that backbone's own in-context ceiling of

0.808 — paired 12:5, one-sided as pre-registered, p=0.07, $+0.056$ with 95% CI $[-0.01, +0.12]$: the pre-registered line reads *direction only*, and no cross-backbone effect size is compared. Post hoc, the precision structure of §3 reappears: IA@1 0.184 beats IA@6 (paired 12:3, p=0.035) for the write pipeline while one line *hurts* the frozen front end (0:5), and IA@1 against the frozen front end's IA@6 is 22:6 (p=0.004).

18. The adapter as reader, under the confirmed protocol. Rule 1 rested on a preliminary accumulated-store arm (frozen reader 0.216 vs adapter reader 0.136, paired 12:2, $p\approx0.013$); a pre-registered rerun of the confirmed pipeline (seed 42) generated every answer twice from the same retrieved lines, frozen base and adapter. The adapter loses: IA@6 0.448 vs 0.512 (paired 3:11, two-sided p=0.057), IA@1 0.448 vs 0.536 (1:12, p=0.003); on the tasks where the gold was retrieved it reads worse (0.75 vs 0.85 at k=6), and only one of 125 hot answers degenerates (repetition-loop check), so the loss is reading, not collapse. The cold arm is a fourth seed-42 run and landed above the confirmed spread (0.512; retrieval 0.600), disclosed in Limits and not banked.

19. All 125 facts in one adapter, one persona. A pre-registered single-seed stream wrote every development fact into one adapter (fixed 8 gold steps each, two background turns between facts; store = background plus every fact written so far) and probed every written fact at five checkpoints. The written front end's retrieval@6 is 0.08 at the end against the frozen question's 0.12 (paired 6:11, two-sided p=0.33), and between 0.05 and 0.14 at every checkpoint: the pre-registered line reads *sharpener dies under accumulation*. The failure is selection: the statement names a different written fact on $95/125$ tasks and its own on 17, while direct-question recall holds 0.53 and the certificate 0.71. Direct recall by age falls from 0.64 to 0.40 between 50 and 75 later facts ($0.64/0.64/0.40/0.48/0.48$ by quintile, newest first) — the natural-fact counterpart of the 3–8-fact half-lives of the pseudoword streams (App. B, Table 3, first row) — and the next ten unwritten facts' median $\Delta\ell$ falls to -1.45 as the adapter accumulates, so written and unwritten separate more cleanly than in step 18's stream. The 125 facts are 125 personas merged into one user with mutually inconsistent disclosures (46 in health alone), so the "best guess about this user" prompt is ill-posed there and the wrong-fact statements are same-domain three times above chance ($58/95$): an upper bound on selection failure, not a deployment estimate. A coherent many-fact persona is the next cell; no IA was judged (retrieval already decides the line).

20. Coherent many-fact personas. The super-user of step 19 is incoherent, so a pre-registered stream used 20 real personas from LoCoMo (one adapter each): their evidence-linked facts were filtered to stable actionable states by one assisting-AI pass, given InMind-recipe indirect queries by a second under a coherent-persona clause (no other fact of the persona may constrain the query; 168 of 456 skipped for that reason), gated for zero content-word overlap, and audited by an independent third pass that kept 146 of 286 (ambiguity, irrelevance, leaks, four contradictions); every dropped fact still entered the stream as a one-step turn write and a store line. Facts were written in dialogue order at eight steps with the persona's intervening turns between them; the store is the persona's own turns so far. Pooled over the 146 facts: written statement retrieval@6 0.192, frozen hop question 0.178 (paired 16:14, two-sided p=0.86; the pre-registered line reads *direction*), the direct-recall stem proxy 0.705, and heuristic cross-fact stem matches on 0.47 of tasks. A hybrid front end that embeds the frozen question and the statement together, recorded as a descriptive arm, reaches 0.253 (19:8 over the question, p=0.052; 12:3 over the statement, p=0.035; post hoc). On real conversational logs every front end sits far below its InMind-isolation level; the queries are assisting-AI generated (correlation as in step 14) and per-persona counts are 3–15, so no per-persona claim is made. Pooled task-level tests do not account for clustering by persona or source conversation.

21. Matched-generation hybrid controls under coherent accumulation. An existing follow-up (chain 64, seed 43; 20 personas, 146 final-checkpoint queries) retrained the stream and compared question-plus-statement hybrids. The frozen statement initially refused throughout; two pre-run smoke amendments removed the system prompt and prefilled the assistant with "You" to elicit a forced guess. Each hybrid uses two generations capped at 48 tokens. Written hybrid (HA) retrieval@6 is 0.274 versus 0.240 for frozen forced-guess hybrid (HF2): 22:17, one-sided p=0.26, $+0.034$, paired Wald 95% CI $[-0.049, 0.118]$. The co-primary comparison does not pass its pre-registered criterion; superiority remains unestablished. Bare written statement is 0.137 versus frozen question 0.178; the direct-recall stem proxy is 0.678. This is a separate seed from step 20, not an ablation on its adapters. Fact-level tests are descriptive without persona/conversation clustering. These refusal-avoidance amendments differ from explicitly labeling generated content as hypothetical; that control has since been run under per-task isolation (App. D), not under accumulation. A post-hoc

conversation-cluster sensitivity analysis gives an equal-persona difference of 0.050 (rather than the task-weighted 0.034), sign-flip $p{=}0.211$, and cluster-bootstrap 95% interval $[-0.015, 0.117]$. It uses ten source conversations and assumes independent clusters and sign symmetry under the null; its small-cluster interval is exploratory.

Voided instruments. Three instruments returned flattering artifacts and were voided by our own audits before any number was kept: (a) the echo-credited hop probe above; (b) a first likelihood-scan retrieval probe that returned a perfect score — voided because with only the fact ever written the scan string is byte-identical to the training string and the probe measures training-set membership, so a perfect score is the null (the real-stream scan of §4, where the needle ranks $236/241$, replaced it); and (c) a recall criterion whose substring check credited query-echoed and refusal-vocabulary words (the assistant persona answers direct personal questions with a privacy refusal), leaving roughly one clean case in 24 — rebuilt with a persona prompt that legitimizes memory answers, an honest "I don't remember" exit, fabrication controls, and stem matching. Each void is recorded in the ledger with the audit that triggered it.

B The wider program

Earlier reports describe streaming re-instatement and recognition (Lin, 2026b), and log-grounded consolidation versus self-recitation (Lin, 2026a). Table 3 combines those prior results with subsequent matched-budget editing, read-head, and 9B maintenance audits archived in the anonymized bundle. It compresses this five-act background program to the verdicts the main text leans on; these are not new InMind results. Every row was measured on synthetic sequential streams of templated attribute–value facts with unguessable pseudoword values (the register of §6), read by likelihood, recognition, and direct-recall probes with a general-reasoning firewall; none uses InMind. Only the last row is measured on this paper's 9B substrate; the others are 2B/1.7B, used for existence and direction, never effect size.

Claim	Key numbers	Evidence class
Capacity is re-instatement, not persistence: first-miss 3–8 writes under every mechanism while end-state retention varies $12\times$	mechanism ladder: naked $3.7 \rightarrow$ EWC+miss-replay 44.2 (of 48 facts retained)	3–5 seeds
Recognition survives recall death, cross-mechanism	matched-budget editing arm: recall 0–$1/48$ vs recognized 45–$47/48$	cross-mechanism, 2 substrates
Consolidating from the model's own outputs poisons	self-recitation retains 4.3 vs doing-nothing 25.0 of 144 facts; error-gated selection wins at matched budget with 2.1–$2.2\times$ recency bias	replicated, 2 substrates
Candidate-conditioned familiarity is a reliable read head; candidate-free readout is not	$\Delta\ell$ AUC 0.93–0.99 written-vs-never-written; pre-write base at chance (0.49–0.54); self-argmax gates fail	3 seeds; 3 independent gate failures
The maintained regime (one-step writes + miss-gated replay) holds at scale	384 facts at 0.815–0.849 recall, $O(1)$/turn, flat-to-inverted age curve, capability firewall $+0.00$ (a general-reasoning benchmark unchanged under the full write stream; zero delta on each seed, not a rounded mean)	3 seeds, 9B

Table 3: Verdict table for the wider program (the long-form audit in the anonymized bundle, where per-seed numbers and the mechanism vocabulary — miss-gated replay, self-recitation, EWC (elastic weight consolidation) — live). No inferential statistics are attached; the InMind sections never borrow significance from this table.

The five-act audit summarized in Table 3 — capacity as re-instatement, recognition surviving recall death, log-grounded consolidation versus self-recitation poisoning, the candidate-conditioned read head, and the maintained 384-fact regime — is released in full (per-seed raw results, per-chain pre-registrations, amendment history, and the voided-instrument audits) with the anonymized artifact bundle in the supplementary material; the InMind sections of this paper never borrow significance from it.

C Deferred readout-mechanism detail

Write-schedule readout rates (§4). Probed immediately after its own writes, the isolated probe has a statement-instance proxy of 0.136, retrieval 0.136, and 0.712 refusals (`ptrc_b_full.json`, `PT_RC`).

The full schedule (41 pre-gold, 10 post-gold writes) reaches 0.576 retrieval; its per-task flags give $75/125 = 0.600$ statement instances and $15/125 = 0.120$ refusals (`rs_full.json`). The schedules are unmatched in total steps, effective dose, and last-written form, and an instructed-refusal reading is not excluded (§4), so the comparison is a post-hoc association, not a scheduling rule.

Chimera decomposition of burial failures (§6). Scripted analysis in the bundle; it characterizes one output pattern. Of 125 burial tasks, 68 strict-recall failures still emit a pseudoword-shaped string; all are *chimeras* — novel strings equal to neither the true value nor any competitor — but the recombination is degenerate: 31 of 68 are one syllable repeated, and four syllables account for 91% of the emitted syllables. Own-syllable retention (the share of a chimera's syllables that belong to the target value) is 0.29, against an expectation of 0.14 under a uniform draw from the 20-syllable pseudoword lexicon and 0.18 under a draw matched to the syllable frequencies of the 40 burial values ($1.6\times$). In these 68 failures the emission sequence concentrates on the most-written syllables, carrying marginally above-chance target content, while the certificate margin rises — read on the whole templated line, which the burial writes share; a value-token-only read is what the "strengthening" half still needs.

D Repair experiments and protocol amendments

E1: fixed-snapshot Mem0 query intervention. Before new inference, the original vanilla store was selected for its baseline role (35,519 extracted memories across 125 task users). Both query arms used verified copies of that snapshot, with no add/update/delete operations. Canonical payload-plus-vector hashes were unchanged for every task and the complete store; the source SQLite byte hash was unchanged. The benchmark question and the existing frozen genre-matched question each retrieved up to six memories with the same BGE-small embedder, Mem0 2.0.19, threshold 0.1, and no reranker. Both answers were newly generated with cold Qwen3.5-9B, bf16, temperature zero, no thinking, and 300 new tokens. Arm order was randomized with seed 20260905. This repairs query isolation, not the unmatched ingestion workload.

The primary endpoint was paired IA@6 under the canonical application rubric. The first blind Qwen3.8-27B judging attempt completed 244/250 items; six exhausted three identical retries with prose instead of JSON, five reaching the token limit. No incomplete aggregate was computed and no failure was scored as zero. An explicit format amendment (J2) rejudged all 250 items uniformly with constrained JSON containing binary score and a reason of at most 240 characters; rubric, model, temperature, thinking setting and 300-token budget were unchanged. Two nonstudy synthetic items checked format support. J2 completed 250/250 and supplies every E1 score reported here; the J1 protocol and its per-item successes and errors are archived in the bundle, not mixed into the scores. Scores can depend on formatting constraints, so J2 is a distinct amended judge protocol, not an unnoticed replacement for J1 or historical Claude grading. Genre-query and vanilla pass 29 and 3 of 125 tasks, respectively, with 26 wins, zero losses, three joint passes and 96 joint failures. The paired-bootstrap interval reported in §7 uses 10,000 draws with seed 20260905. Human validation has not been performed.

E2/E3: a same-inference-budget frozen hypothetical comparator and a query ablation. The protocol was locked before any GPU output: two primary endpoints, A−HYP1 IA@6 and correct−absent gold-in-top-6, both exact two-sided paired tests, Holm-corrected across exactly those two, with paired-bootstrap 95% intervals ($10{,}000$ draws, seed 20260905); every other contrast is descriptive. A later amendment changed checkpoint-integrity checking only (exact PEFT key/shape/value comparison on disk, an initialization-reset reload comparison, and a live-state hash bound across sessions); data, prompts, seed, training steps, decoding, readouts, endpoints, and the multiplicity family were unchanged. The full sample was fixed at the 125 development tasks before the run; a two-task smoke was engineering-only and unscored.

Each task gets one shared seed-42 write under the confirmed recipe (41 pre-gold background `Note` steps, gold `Note`+chat-QA rounds until the original recall gate or 12 rounds, ten post-gold `Notes`; rank-64 all-linear LoRA, 3×10^{-5}, Qwen3.5-9B bf16, one visible GPU). Five readouts share that adapter with no write between probes: the original indirect query (A), a permuted query, an absent query, the frozen hypothetical key (HYP1), and a frozen 32-token question reference (Q_EMB, retrieval-only, regenerated to this run's specification rather than reused from the historical files). A and HYP1 each get one 48-token key, one embedding, one top-6 search, and one 300-token cold

read from the frozen base — matched in generation count and maximum generated length, not in prompt tokens or total training cost. HYP1's system prompt instructs the frozen base to invent one plausible first-person disclosure that would change the answer, explicitly as a search hypothesis with no access to the user's history; the invented assertion is used only as a search key, and only retrieved real log lines reach the reader. The permuted arm substitutes another task's indirect query under a fixed seeded derangement over all 125 ids with no self-donors; donors were not reviewed for relevance and are not an unrelated-query control. The absent arm empties the question slot of the same `PERSONA+RECALL_Q` wrapper, leaving every other instruction in place; the resulting wrapper is awkward, and this ablates query content rather than instantiating a generic "recall any fact" probe. Neither the permuted nor the absent arm falls back to the original query. Verified on the stored readouts: in all five arms, every generated retrieval key was non-empty on all 125 tasks, so the harness's empty-key fallback was never exercised.

All 125 tasks completed with no missing ids. Answers were graded blind under J2. E2: IA@6 $54/125{=}0.432$ (A) versus $41/125{=}0.328$ (HYP1); 26 joint passes, 56 joint failures, 28:15 discordant; difference 0.104, exact two-sided $p{=}0.065994$, Holm $p{=}0.1228567$, bootstrap 95% CI $[0.000, 0.208]$. E3: gold-in-top-6 0.584 (correct) versus 0.496 (absent); 53 joint passes, 43 joint failures, 20:9 discordant; difference 0.088, exact two-sided $p{=}0.061428$, Holm $p{=}0.1228567$, bootstrap 95% CI $[0.008, 0.176]$. Neither primary endpoint meets the corrected 0.05 threshold. The two intervals are marginal and were computed by a different method than the prespecified tests; they are not simultaneous, and they do not override the exact tests. Neither non-rejection is evidence of equivalence. Descriptive contrasts on the same artifacts: retrieval A 0.584 versus HYP1 0.408 (40:18, $p{=}0.005355$, CI $[0.056, 0.288]$), retrieval correct 0.584 versus permuted 0.416 (28:7, $p{=}0.000508$, CI $[0.080, 0.256]$), and the frozen question reference at 0.384. These are descriptive; the run's success criterion was the two primary endpoints, and the descriptive retrieval contrasts are not substituted for them.

Post-feedback anchor: the frozen question reference under J2. Automated pre-submission feedback asked whether HYP1 is in fact a stronger end-to-end comparator than the original frozen question front end under matched judging. A pre-registered addendum read Q_EMB's stored top-6 retrievals with the same frozen reader (greedy, 300 tokens; the A arm's stored prompts were reproduced byte-for-byte by the same construction) and judged the 125 answers blind under the J2 protocol on a recreated server of the same build, model file and arguments. Q_EMB reaches IA@6 $40/125{=}0.320$: level with HYP1 (0.328; 20:19, exact two-sided $p{=}1.0$, bootstrap 95% CI $[-0.088, 0.104]$) and below A (0.432; 29:15, $p{=}0.049$, CI $[0.008, 0.216]$). The pre-registered reading is that HYP1 was not weaker than the historical comparator, so the E2 null is not an artifact of a weak control; the anchor is outside the Holm family and changes no verdict. That the frozen reference scores 0.320 under J2 and 0.312 under the historical judge is a consistency check between judges, not a replication.

Scope. The tasks are previously-seen development data, so this is a repair diagnostic and supports no generalization claim. J2 is a distinct judge protocol, so the A arm's 0.432 is not a re-measurement of the Claude-judged 0.424 and the two must not be differenced. A versus HYP1 changes the front-end recipe and its prompt together, so it prices a system, not the effect of the written parameters alone. The target still receives direct-QA supervision and roughly eight updates against one per background turn, so the absent arm's 0.496 is consistent with query-independent salience without proving it. The coherent control below is a separate prespecified experiment; the historical accumulation runs in App. A, steps 20–21, do not replace it. Local verification covered the input manifest hash, every arm's prompts, row and checkpoint markers, and the reported arithmetic; adapter weights stayed on the execution host, where serialization roundtrips were checked.

E3C: coherent-accumulation question-conditioning control. All 20 personas, 146 tasks and 10 source conversations completed under seed 43, with no omissions, aborted attempts, or score-based exclusions. One adapter per persona receives that speaker's turns through the last target disclosure. Each non-gold turn gets one `Note` update; each gold turn gets four fixed `Note`+chat-QA rounds (eight updates), with four recorded recall gates and no early stopping. The final adapter is probed with correct, absent and fixed within-persona deranged queries, each capped at 48 tokens, with unchanged weights and store across probes. The donor belongs to the same speaker and conversation; topical relatedness was not reviewed, so derangement is not an unrelated-query control. All keys were non-empty.

The single prespecified primary is the equal-persona mean correct−absent Hit@6 difference, 0.147193. Enumerating all 1024 source-conversation sign flips gives 8 equally or more extreme assignments, two-sided p=0.0078125. This assumes independent conversation clusters and sign-symmetric contributions under the null; it is not a randomized-treatment causal test. The 20,000-draw cluster bootstrap (seed 20260905) gives $[0.075540, 0.227729]$, an exploratory few-cluster interval, not a coverage guarantee. This single-primary family was locked separately from E23; its success does not replace E23's two nonsignificant primaries. Task-weighted rates are correct $31/146$=0.212329, absent $12/146$=0.082192, and deranged $9/146$=0.061644; their correct−absent difference (0.130137) is distinct from the equal-persona primary. Descriptively, deranged queries retrieve the donor's gold on $30/146$=0.205479, versus the original target on $9/146$.

Manipulation checks retain all tasks: direct recall's five-character stem-hit proxy is $95/146$=0.650685; mean likelihood lift is 2.691576, with none nonpositive. These are readout checks, not semantic recall validation or exclusion criteria. The run records 6167 gradient steps, 584 gate generations and 438 arm generations. Every checkpoint and live-state binding was checked on the execution host; local verification independently recomputed marker hashes, all probe prompts and rankings, and statistics. Adapter binaries remain on that host. The cohort is previously seen, only one seed was run, and fixed-round training differs from historical runs. No frozen arm or end-to-end IA was measured. The result supports question conditioning within this accumulated trained state; it does not establish that writing beats a frozen retrieval recipe or yields reliable long-term memory.

Historical second-judge missingness. The earlier InMind Qwen grading script coded unparseable responses as zero: 28 of 875 items (seven arms of 125). The IA@6 seed arms have 7, 5, and 2 missing labels; their shared frozen comparator has 3. The other 11 failures are in IA@1 arms. Its stored p values are one-sided; under the paper's two-sided convention the zero-coded results are $0.302/0.066/0.013$, not a clean two-of-three replication. We retain that historical file and enumerate every binary completion of failed labels, keeping parsed labels fixed and each frozen-comparator completion shared across seeds. IA@6 difference bounds are $[0.040, 0.120]$, $[0.080, 0.144]$, and $[0.120, 0.160]$; the number of positively significant seeds ranges from one to three. These are missing-label bounds, not confidence intervals or corrected verdicts. They establish direction conditional on the parsed scores, but leave the replication criterion unresolved. The repair runs instead reject incomplete judging batches.

The historical PrefEval Qwen script maps an absent answer tag to "No" and stores only derived labels, not the raw judge responses. Its parser-failure frequency is therefore unknown from the archived file. The reported 650-item scores and agreement describe those recorded labels; they do not establish complete valid parsing or human agreement. This limitation is distinct from the 28 known InMind failures.

LoCoMo reversal detail. A pre-registered falsification test took the raw-log recommendation to LoCoMo (Maharana et al., 2024), a second, independently authored benchmark (ten conversations ingested whole by the same Mem0 pipeline, a fixed-seed 50-question slice, identical embedder, query, and serve, a blinded judge, and a no-store floor; App. A, step 15). Its descriptive $+0.05$ trigger fired: extraction's point estimate is 0.475 versus 0.400 on multi-hop+temporal (n=40, paired 8:5, two-sided p=0.58) and 0.800 versus 0.500 on single-hop (n=10, 3:0, p=0.25; floor 0.100). The small slice reverses the point estimate without establishing a general explicit-recall advantage; a triggered descriptive threshold is not inferential evidence of superiority.

Cost detail. The shared 9B backbone and GPU do not imply matched workloads. LoRA uses 41 pre-target and 10 post-target background steps plus about eight target steps; the repair run logs $7{,}367$ gradient steps in 3505 s and 496 recall-gate generations in 786 s across the 125 tasks (about 28 s of gradient steps and 6 s of gate generations per task), excluding loading and other overhead. Its query front end adds one generation (capped at 48 tokens) and one embedding. Mem0 ingests each task's 241 lines through LLM extraction into a per-task store, $\sim$285 memories in $\sim$1650 s per task, with 0.14 s per search, measured with its two arms sharing service load. These timings are descriptive, not an equal-stream speed comparison. Full wall-clock accounting, gate calls, token counts, adapter storage and loading, and matched ingestion/query workloads are needed for deployment pricing.