EXPERIENCE MEMORY AND EPISODE COMPLETION: A PRE-REGISTERED AUDIT ON TWO LLM AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Experience memory, a bank of distilled past episodes retrieved into an agent's context, is widely reported to improve LLM agents, but outcome-bearing memories may also induce premature commitment. We pre-registered a 33-arm dissection of an ExpeL-style memory bank on three multi-hop QA benchmarks (about 29,700 logged episodes, plus 3,000 in two later blocks): the design was hash-frozen before any test episode, four rounds of numeric go/no-go gates ran without a threshold change, and every deviation is reported (Appendix A). For Qwen3.5-27B, the pre-registered `NoMem–Mem` contrast is equivalent within the registered margin (paired ΔEM $+0.6$pt; TOST 90% CI $[-1.2, +2.4]$ within ± 4pt). None of six content manipulations produces a significant change after Holm correction, and memory does not shorten search: unconditional searches rise, and episodes that reach an answer (a set the treatment itself shifts) use the same number with and without the bank. Memory instead changes how episodes end: the share that exhausts the call budget without answering rises from 14.0% to 25.2%. A post-hoc, single-seed ablation shows that episode-completion mechanics alone (a budget extension and a terminal forced finish) reproduce most of the gain of the evidence-citation instruction arms in three of four decompositions: mechanics recover $+6.2$pt and the instruction adds $+2.0$pt beyond them (90% CI $[-0.2, +4.2]$). Two blocks on a second hardware setup ($4\times$A6000) give, for mechanics and then instruction, $+3.8$ and $+0.4$pt with the bank and $+2.2$ and $+2.6$pt without it. At matched mechanics the memory contrast is $+2.4$pt on both hardware setups, too imprecise to establish either a nonzero effect or equivalence. Hard enforcement does not outperform the instruction, a content-free bookkeeping instruction reaches a similar EM point estimate, and optimizer-evolved instructions (GEPA, MIPROv2) do not significantly outperform the hand-written instruction on the full test set; their minibatch gains do not transfer. Two conclusions reverse on gpt-oss-120b, a mixture-of-experts agent with lower base EM on the matched subset: memory helps ($+8.5$pt, $p=3.7\times10^{-7}$, two seeds) and stored answers carry the value (redaction -5.3pt). Memory-bank effects differed across the two agents studied here and should be reported per agent. For Qwen3.5-27B, completion mechanics plus one instruction, without a bank, trails no memory configuration by more than 3pt (pooled point estimates), at under two-thirds of the tokens per episode of `Mem+Cite`.

1 INTRODUCTION

LLM agents increasingly use *experience memory*: episodes from past tasks are distilled into reusable items (insights, procedures, outcomes) and retrieved into the context of new episodes (Shinn et al., 2023; Zhao et al., 2024). Although prior work reports substantial gains, outcome-bearing memory may also bias an agent toward a previously seen answer and reduce the search it would otherwise perform. We refer to this hypothesized failure mode as *memory-induced shortcutting*.

This study was designed to test that failure mode causally. Before running any test episode, we pre-registered three claims: (C1) outcome-bearing memory items cause premature commitment while answer-redacted, token-matched items do not; (C2) a provenance-free candidate answer alone reproduces the truncation; and (C3) a hard evidential floor removes most memory-induced breaks while the same requirement expressed only as an instruction recovers less than a third. The design (33 arms, minimal-pair payload swaps under pinned retrieval, numeric go/no-go gates, exact

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.

paired statistics with equivalence margins, and treatment-free seed replicates) was frozen by SHA-256 before the main matrix ran (Appendix A). Neither C1 nor C3 held for the primary agent and C2's registered branch metric turned out to be inapplicable; we report the null results and the partial reversal on a second agent.

For Qwen3.5-27B, the pre-registered memory contrast is equivalent within the specified margin. The primary endpoint (no-memory `NoMem` vs. full-bank `Mem`; registered IDs `M0`/`M1`, and Appendix B maps every registered arm name to its registered ID; n=1,000 paired episodes) is $+0.6$pt with a TOST 90% CI of $[-1.2, +2.4]$pt, inside the registered ± 4pt margin. This supports equivalence rather than only a non-significant difference. None of the six content manipulations yields a statistically significant change, and the predicted behavioral signature is not observed: memory does not shorten search (unconditionally 3.31→3.95, because more episodes exhaust the call budget; 2.62 per answered episode in both arms, a set the treatment itself shifts) (§4.1). A mechanics-matched re-test of the same contrast is single-seed and inconclusive (§4.2).

A post-hoc ablation suggests that completion mechanics reproduce most of the gain of the evidence-citation arms (in three of four single-seed decompositions; in the fourth the instruction component is the larger, $+2.6$ vs $+2.2$pt), while the instruction's residual effect is small and imprecisely estimated. This ablation, run after an automated adversarial review, shows that the registered slot contrasts partly reflect the A6/A8 completion mechanics. Enabling the budget-extension and forced-finish mechanics alone yields $+6.2$pt, while memory arms ended answerless in 27.6% of episodes (seed 11). The hand-written evidence instruction adds $+2.0$pt beyond mechanics (90% CI $[-0.2, +4.2]$; an effect up to the equivalence margin cannot be excluded), enforcement does not outperform the instruction alone, and optimizers seeded with the instruction do not improve reliably over it on the full test set; their minibatch gains do not transfer, which is consistent with adaptation to exact-match scoring noise (§4.2–§4.4).

Two conclusions reverse on gpt-oss-120b. Replicating five arms on this agent (two seeds, n=600 paired) shows that memory improves EM by $+8.5$pt and that redacting stored answers reduces EM by 5.3pt; enforcement has a $+1.2$pt higher point estimate than instruction, but the direct enforcement–instruction contrast is not significant on either agent. The mechanics and optimizer arms were not replicated on gpt-oss-120b (§4.5).

Our contributions are: (1) a pre-registered, gate-controlled dissection of agent experience memory, with the gate history, amendments, and deviations published; (2) a minimal-pair arm design that separates memory *presence*, *content*, *capacity*, *relevance*, and *provenance* from the instruction slot and its enforcement; (3) a post-hoc completion-mechanics ablation and a completion-mechanics + instruction-only baseline; and (4) evidence that memory-bank effects can differ across the two agents studied.

2 Related Work

Experience memory for agents. Reflexion (Shinn et al., 2023) and ExpeL (Zhao et al., 2024) distill episodic experience into reusable text; subsequent systems extend the approach to procedural banks and self-editing stores (e.g., Ouyang et al., 2025) (verification table in Appendix H). A memory arm compared against a no-memory baseline differs from it both in the text that fills the context and in how episodes terminate. For Qwen3.5-27B, we find that such a comparison can conflate memory effects with instruction-slot and episode-completion differences.

Utility-aware memory control. Beyond static banks, MemRL learns runtime retrieval utility (Zhang et al., 2026), MemAudit specifies an evaluation protocol for budget-aware write-time selection (Bhargava et al., 2026), and adaptive memory controllers or retention mechanisms manage which items are kept or surfaced (Sun et al., 2026; Dao et al., 2026). These act at retrieval or write time; in our study the bank's residual with the instruction slot filled is already below the null floor (§4.3).

Retrieval design. Retriever choice measurably moves deep-search agents (Chen et al., 2025), and structured task retrieval, whether bridge-conditioned or over proposition graphs, improves multi-hop QA evidence quality or its token cost (Bacellar, 2026; Jenkins et al., 2026). Separately, structured episodic event graphs have been used for long-term agent memory retrieval (SEEM; Lu et al., 2026). The former results concern the *task* retriever, whereas our manipulated variable is the *memory* re-

triever, the setting closer to SEEM; here the registered decoy arm gives a null for memory-retrieval quality at this sample size (items retrieved for other questions leave EM unchanged; `Mem-Decoy`, Table 2) and a post-hoc BM25 swap agrees (`Mem-BM25`, Table 4).

Prompt optimization. DSPy's MIPROv2 (Khattab et al., 2024; Opsahl-Ong et al., 2024) and GEPA-style reflective evolution (Agrawal et al., 2025) treat the instruction as a trainable artifact. We use budget-matched lite variants as *arms*, asking whether optimization, seeded with the hand-written evidential instruction, exceeds it on the full test set; it does not (§4.4). A memory bank and an optimized instruction can encode the same training episodes; our F-A2 family tests whether the bank still adds value once the slot is filled (§4.3).

Pre-registration and equivalence. Our protocol imports pre-registration (Nosek et al., 2018), exact McNemar tests on paired fix/break tables, Holm correction (Holm, 1979), and TOST equivalence testing (Lakens, 2017), plus a treatment-free *null floor*, the largest $|\Delta\mathrm{EM}|$ between re-runs of one condition at different seeds (2.8pt here, larger than our primary effect), below which no effect is claimed.

Metric validity. Exact match is the default metric for multi-hop QA and is sensitive to gold formatting; a registered semantic-fidelity check and a frontier audit (Appendix E) quantify that sensitivity here.

3 Experimental Design

3.1 Environment, data, and agent

Episodes are ReAct loops (Yao et al., 2023): at each turn the agent emits `Search[query]` (BM25 (Robertson & Zaragoza, 2009) over the question's own corpus, top-5 passages) or `Finish[answer]`, at most 8 calls, temperature 0.7 with per-call seeding. Data are the HippoRAG (Gutiérrez et al., 2024) subsets of HotpotQA (Yang et al., 2018), MuSiQue (Trivedi et al., 2022), and 2WikiMultiHopQA (Ho et al., 2020): 1,000 questions each with per-dataset corpora of 9,811/11,656/6,119 passages. Seeded, stratified (50/30/20) splits: S_{bank}=400, S_{opt}=150, S_{dev1}=300 (150 at freeze, extended by amendment A7), S_{dev2}=100, S_{test}=500, with a 1,550-question untouched reserve; all instance IDs are registered. The release holds about 29,700 logged episodes: 25,500 test, 3,000 replication, 400 bank, and 820 development; the two second-hardware blocks add 3,000 (Appendix F). Scoring is SQuAD-normalized EM (primary) and F1 (descriptive). The primary agent is Qwen3.5-27B (Qwen Team, 2026) (bf16; thinking disabled through the chat template, as logged in amendment D1; ordinary calls are capped at 420 output tokens and the A8 terminal call at 800; vLLM 0.19.1 (Kwon et al., 2023)); the replication agent is gpt-oss-120b (OpenAI, 2025) (mxfp4, served at the engine's default reasoning effort; the client sets no reasoning parameter); the distiller/redactor is Qwen3-32B (Qwen Team, 2025), which shares the agent's model family but is a different checkpoint and generation. The memory retriever is a CPU-pinned MiniLM bi-encoder (Reimers & Gurevych, 2019).

3.2 Memory bank and minimal-pair arms

Running `NoMem` on S_{bank} solved 199/400 episodes (EM .498); each solved episode was distilled into an item with `INSIGHT`/`PROCEDURE`/`OUTCOME` lines, the last recording the answer verbatim. The distiller receives the question, the number of searches, and the final answer (trajectories were not logged), so `PROCEDURE` lines are the distiller's reconstruction of a plausible search plan rather than the executed one. Retrieval is dense top-k over the original items. The top-k selection is computed once on the original bank and held fixed across payload variants, so arms differ only in payload text: `Mem` original top-7; `Mem-NoAns` answers redacted (token-parity $\pm 5\%$); `Mem-Swap` answers swapped for equally informative non-answers; `Mem-AnsOnly` answer-only (procedures deleted); `Mem-Decoy` decoy items retrieved for different questions (relevance control; unpinned by design); `Mem-Top3`/`Mem-NoAns-Top3` top-3 (capacity); `Mem-Curated` contradiction-curated bank (190 items); `Guess` a single line presenting the top item's answer as the agent's own unverified guess (provenance control); `Mem+Budget Mem` with a +3-call budget. Full definitions in Appendix B. A post-hoc title-level check finds that 12 of the 199 bank items and 23 of the 150 S_{opt} questions share a gold-document title with an S_{test} question (§6).

3.3 INSTRUCTIONS AND FLOORS

I^* (registered verbatim and hash-pinned; text in Appendix G) asks the agent, before finishing, to output `Evidence: "quote" | "quote"` lines citing passages retrieved *in this episode*; if unable, search further. Soft floor (`+Cite`) supplies I^* as instruction only. Hard floor (`+Enforce`) additionally verifies quotes against actually retrieved text and rejects unsupported finishes (forced re-search, $+3$-call cap, at most three rejections; the A8 terminal answer is not re-checked), so it is a finite-retry policy rather than a guarantee that every final answer is supported. A registered bookkeeping control I_B (`+Tools`: output a `Tools: <list>` line naming the searches you ran) fills the slot with no evidential content. Two execution amendments were adopted during gating, with thresholds unchanged and archived measurements kept: A6 grants instruction-bearing arms one $+3$ extension when the budget would expire with no finish attempted; A8 adds one terminal 800-token call when the budget expires without an answer (logged `forced_final`); its prompt asks for at most one short Evidence line and then `Finish`, so `+Complete` is not entirely instruction-free, and 1.8–4.6% of its episodes still end answerless when the attempt yields no parseable finish (Table 4). A6 and A8 are applied as one package and not separated here. Both were restricted to instruction-bearing arms (and, as logged, the gate-only closed-book arm) so that the already-run bare arms and the frozen primary endpoint stayed untouched; this choice produced the mechanics asymmetry that §4.2 repairs post hoc (Appendix A).

3.4 OPTIMIZER ARMS

GEPA-lite evolves I^* by failure-reflection mutations under a 350-rollout budget on S_{opt} (shared deterministic minibatches; reflection by a frontier model, logged verbatim), run twice: with the bank in context (`GEPA@Mem`) and without (`GEPA@NoMem`). MIPRO-lite proposes six grounded candidates and applies successive halving (`MIPRO@Mem`). Champions were fixed before their S_{test} runs (curves, reflections, and champion texts: Appendices D and G).

3.5 PRE-REGISTERED INFERENCE

EM is gated as the sole primary metric. The single primary test is `NoMem` vs. `Mem` at $\alpha{=}.05$ (exact McNemar on the paired fix/break table); all other comparisons live in three Holm-corrected families fixed at freeze time: F-A4 (break reduction vs. `Mem`), F-A5 (content manipulations vs. `Mem`), and F-A2 (bank value with the instruction slot filled). Two further registered families (an interaction term and dataset slices) are not reported as confirmatory tests (Appendix A). "No difference" claims require TOST: a 10k-resample bootstrap 90% CI of the paired EM difference within ± 4pt, resampling episodes as registered. Because a question's two seeds are not independent, figures and post-hoc analyses use question-cluster bootstrap CIs (both seeds resampled together; primary contrast $[-1.4, +2.7]$, still inside the margin). The null floor is the maximum treatment-free $|\Delta\text{EM}|$ across five same-arm cross-seed replicates. Numeric go/no-go gates preceded the matrix; the complete four-round history, including four gates that initially failed and the execution amendments that resolved them without changing any threshold, is in Appendix A.

4 RESULTS

Sections 4.1 and 4.3–4.5 report the registered analyses: the primary test, the three Holm families, the descriptive exploratory arms outside the registered families, and the replication. Section 4.2 is a post-hoc, single-seed ablation added after an automated adversarial review; it lies outside every registered family and is labeled as such wherever its estimates reappear. Per-episode token cost for every arm is listed in Tables 3, 4, and 16.

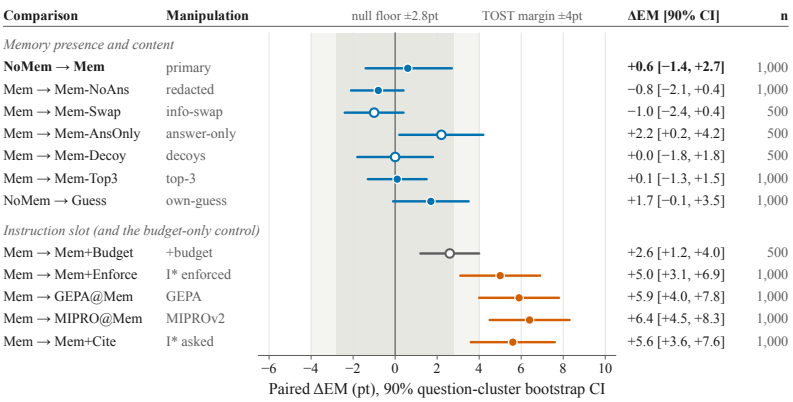


Figure 1: Registered presence, content, provenance, and slot effects (Qwen3.5-27B, S_{test}, paired ΔEM; F-A4 and F-A5 subset plus the provenance control, F-A2 in Table 13). Memory-presence and content manipulations (blue) have point estimates inside the null floor (±2.8pt, dark band) and the TOST margin (±4pt, light band); instruction-bearing arms (orange) clear both; the budget-only control `Mem+Budget` (grey, +2.6pt) clears Holm correction but not the floor. Question-cluster bootstrap 90% CIs (10k resamples). Open markers: seed 11 only (n=500); filled: seeds pooled (n=1,000).

4.1 The primary endpoint is equivalent; the predicted signature is not observed

The primary contrast (Figure 1; Table 2, Appendix C) is +0.6pt (p=.654) with a TOST 90% CI of $[-1.2, +2.4]$pt: formal equivalence at the registered ±4pt margin (also inside ±3pt) and below the 2.8pt null floor. The entire content-manipulation family (F-A5) stays below the floor: redacting answers, swapping them for matched facts, keeping only answers, replacing items with decoys, and halving capacity are all non-significant after Holm correction and sit below the floor (the largest, answers-only, is +2.2pt). We do not observe the predicted behavioral signature: memory does not shorten search. Mean searches rise under memory (3.31→3.95) because every answerless bare-arm episode ends at the 8-call cap (7.6 and 7.9 searches) and memory produces more of them; among episodes that produce an answer both arms average 2.62 searches, a comparison that conditions on an outcome the treatment shifts, which we report as descriptive (Table 6 note). `Guess`, the pure candidate-answer probe, is null on EM (fixes 69/breaks 52, p=.145) and also raises unconditional search (3.53). C2's registered branch metric is inapplicable: R_A=0.48, but both its numerator and denominator are negative (anchor and memory each raise mean calls through a higher answerless share), so there is no truncation to mediate and C2 is reported as moot (Appendix C). Breaks do occur (59/1,000). Echo classification ran only at gate time on the dev set, where 17/17 breaks were non-echo (the wrong answer was not any stored answer); test-set breaks were not classified because model text was not retained, so we report no evidence of the registered echo signature rather than its absence. §4.2 shows that many breaks are answerless episodes, that is, budget exhaustion without an answer rather than a wrong commitment.

4.2 Completion mechanics carry most of the gain (post hoc; one seed per hardware setup)

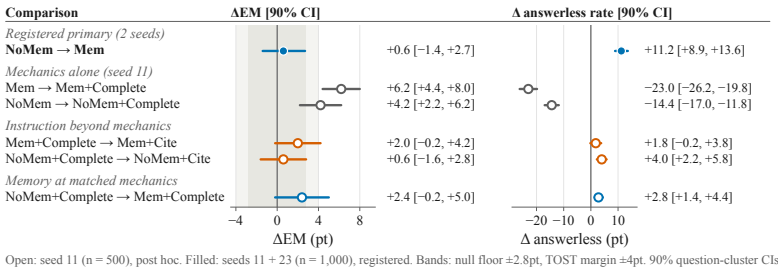


Figure 2: Completion mechanics decomposed (Qwen3.5-27B, S_{test}, paired; left ΔEM with the null floor and TOST bands, right Δ answerless rate). The registered primary contrast (filled, two seeds) leaves EM unchanged and raises the answerless rate by +11pt; the post-hoc mechanics arms (open, seed 11, n=500) recover EM by lowering the answerless rate; instruction beyond mechanics and the bank at matched mechanics leave the EM point estimate inside the margin (CIs cross it) and change the answerless rate by +1.8 to +4.0pt.

The instruction-bearing arms in Table 10 gain +5.0 to +6.4pt over `Mem` on the pooled rows (+7.4pt for the single-seed `GEPA@Mem+Enforce` row). An internal audit of amendments A6/A8 (applied only to instruction-bearing arms; §3) led to a post-hoc ablation in which `+Complete` arms enable the same mechanics with no instruction and no floor (Figure 2; Table 5 in Appendix C; seed 11, n=500). In this single-seed analysis, completion mechanics alone recover +6.2pt on `Mem` (fixes 32, breaks 1) and +4.2pt on `NoMem`.

`Mem+Budget` (unconditional +3 calls, no forced finish) recovers +2.6pt (90% CI [+1.2, +4.0] vs seed-11 `Mem`; Figure 1), a different intervention from A6's conditional extension. The hand-written instruction adds +2.0pt beyond mechanics with the bank (p=.18; 90% CI [−0.2, +4.2]) and +0.6pt without it (90% CI [−1.6, +2.8]); both intervals include zero. In the bare registered arms (Table 5 note), episodes end answerless in 16.2% (`NoMem`) and 27.6% (`Mem`) of seed-11 episodes (14.0% and 25.2% pooled over both seeds). Memory raises the answerless rate by +11.4pt at seed 11 (+11.2pt pooled), opposite to the registered shortcutting prediction, and `+Complete` reduces answerless termination substantially. Figure 2 shows both quantities: each mechanics step lowers the answerless rate by 14 to 23 points as it raises EM, whereas instruction beyond mechanics and the bank at matched mechanics change EM within noise and raise the answerless rate by 1.8 to 4.0 points. Of `Mem`'s 138 answerless seed-11 episodes, `+Complete` turns 27 into correct answers (16 by the forced finish, 7 within the extended budget, 4 with neither mechanism engaged) and 88 into wrong ones; its forced-finish answers are correct 21% of the time against 64% for self-finished episodes (Table 6). By seed, the instruction arm's gain over `Mem` is +8.2pt (seed 11) and +3.0pt (seed 23); Table 7 accounts for where each gain lands, without implying a causal share. The enforcement, optimizer, and `+Tools` arms also carry A6/A8 and were not decomposed. Two single-seed blocks run later on different hardware (4×A6000, same code, prompts, bank, and questions; seed 23, n=500 each; Table 9; setup in Appendix F) repeat the decomposition with and without the bank: `Mem`→`Mem+Complete` +3.8pt [+2.4, +5.4] and `Mem+Complete`→`Mem+Cite` +0.4pt [−1.8, +2.6]; `NoMem`→`NoMem+Complete` +2.2pt [+0.8, +3.6] and `NoMem+Complete`→`NoMem+Cite` +2.6pt [+0.2, +5.2], so without the bank the instruction component is the larger of the two in this block. On that hardware the matched-mechanics memory contrast `NoMem+Complete`→`Mem+Complete` is +2.4pt [+0.0, +4.8] and the bare contrast `NoMem`→`Mem` is +0.8pt [−1.8, +3.4] (paired question bootstrap 90% CIs); these blocks are reported beside the primary 4×A100 block, never pooled with it. The instruction-versus-enforcement, optimizer, and second-agent bare-arm contrasts (§4.3–§4.5) do not depend on this asymmetry. In the A100 block, the memory effect under matched mechanics is +2.4pt (`NoMem+Complete` vs `Mem+Complete`; 90% CI [−0.2, +5.0]); that interval and the A6000 one both include zero and the ±4pt margin, so this comparison is underpowered to establish either an effect or equivalence, and the registered null floor, computed on other arms, is not applied here.

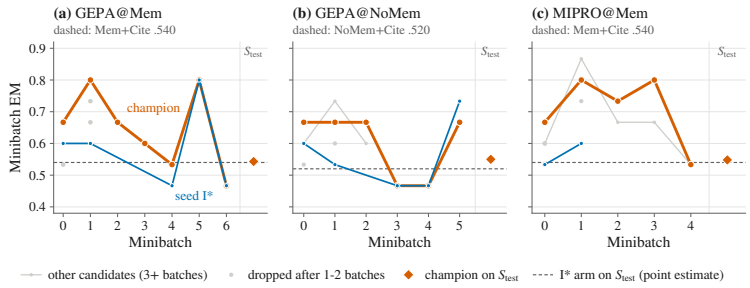


Figure 3: Optimizer training curves (per-candidate minibatch EM on shared S_{opt} batches; x is the shared batch index, not evaluation order, and every new candidate starts at batch 0). Gray = non-surviving candidates (lines: three or more batches; dots: one or two). Champions (orange) beat the I^* seed (blue) on average over the batches both ran, though not on every batch (the seed wins the last batch in b). On the full S_{test} (diamonds), (a) and (c) land near the hand-written instruction's S_{test} point estimate (dashed; `Mem+Cite` .540 in a and c, `NoMem+Cite` .520 in b) and (b) sits 3pt above it (descriptive, §4.4).

4.3 Mechanics plus instruction pass the registered tests

All instruction-bearing arms also use the completion mechanics of §4.2, so the contrasts in Table 10 (Appendix C) compare mechanics plus instruction against a bare `Mem`; `Mem+Budget` and `Mem-Curated` carry neither. Each arm that fills the instruction slot survives Holm correction, whereas bank curation does not. The pre-registered C3 ordering is not observed: C3 predicted that hard enforcement would remove $\geq$50% of breaks while the instruction alone would recover $<$1/3; with mechanics in both arms, the instruction recovers 61% and enforcement 44% of `Mem`'s 59 breaks. The ordering is reversed in point estimate (direct contrast -0.6pt, 90% CI $[-1.6, +0.4]$; no equivalence test registered). These gains also cannot be attributed solely to repairing memory-induced breaks: `NoMem+Cite` gains $+4.2$pt over `NoMem` with no memory at all (pooled; .478 to .520), and at seed 11 that gain ($+4.8$pt, .464 to .512) decomposes into $+4.2$pt from the completion mechanics alone and $+0.6$pt from the instruction (Table 5). The `+Tools` arms, whose registered bookkeeping instruction has no evidential semantics, reach .522 to .548 (seed 11, n=500) with 7.8 to 9.9 searches per episode and 62% and 84% of episodes ending in the A8 forced finish (Table 3); at seed 11 they match the I^* arms (`Mem+Cite` .550, `NoMem+Cite` .512) while searching far more; on this single-seed comparison, the `+Tools` result is consistent with a non-evidential, slot-filling explanation of the residual gain, so we do not attribute the $+2.0$pt residual beyond mechanics specifically to evidential content. The registered coupling family F-A2 gives the complementary comparison: once the instruction slot is filled, adding the bank contributes $+2.0$pt (`NoMem+Cite` to `Mem+Cite`; p=.075, Holm ns; question-cluster 90% CI $[+0.2, +3.8]$; Table 13, CI in Table 8), below the 2.8pt null floor. A residual bank effect of about 2pt is not excluded.

4.4 Optimizers do not beat the hand-written instruction on the full test set

GEPA-lite's champion, an I^* mutation (append: copy the answer exactly as written in the quoted passage), beat I^* by $+13$pt on its first two minibatches ($+7$pt over the five batches both candidates ran; Figure 3); MIPRO-lite's champion behaved similarly. On the full S_{test}, the champions remain within the observed variation around the hand-written instruction: `GEPA@Mem` .543 vs `Mem+Cite` .540 (p=.84), `MIPRO@Mem` .548 (p=.49), and `GEPA@Mem+Enforce` vs `Mem+Enforce` p=1.0 (seed 11, n=500). The registered semantic-fidelity check is consistent with this reading: a frontier audit of EM=0 episodes with F1$\geq$.5 (15 of the 57 such episodes in `Mem`, seed 11) found 15/15 semantically correct answers penalized by missing aliases or gold formatting ("Detroit" vs "Detroit, Michigan"; "22,500" vs "22,500 acres"). Surface-form clauses adapting to batch-specific formatting is one explanation; selection optimism over 15-question minibatches is another, and our data do not separate the two. `GEPA@NoMem` has the highest point estimate in the study (EM .550), but its validation round did not confirm the gain (champion .533 vs seed .556 on the shared batches), so we report it descriptively, as pre-registered for arms outside every inferential family.

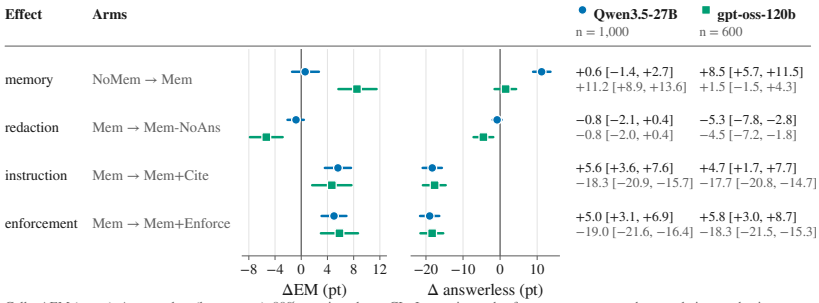


Figure 4: The same four interventions on two agents (left: paired ΔEM; right: paired Δ answerless rate; question-cluster 90% CIs in the cells). Qwen3.5-27B: blue, seeds 11+23, n=1,000; gpt-oss-120b: green, n=600. The memory contrast is equivalent within the registered $\pm$4pt margin on Qwen and $+8.5$pt on gpt-oss (p=3.7×10^{-7}); redaction produces no detectable change on Qwen and -5.3pt on gpt-oss (p=.001); memory raises the answerless rate on Qwen ($+11.2$pt) but not detectably on gpt-oss, while instruction and enforcement lower it by 17 to 19pt on both. Both instruction arms carry A6/A8; the direct enforcement–instruction difference (-0.6 vs $+1.2$pt) is not significant on either agent and lies outside the registered families.

4.5 Agent-relativity: the memory and redaction conclusions reverse on the lower-baseline agent

On gpt-oss-120b (Figure 4; base EM .312 vs Qwen's .465 on the same 300 questions, seeds 11+23 (seed 23 added after the freeze as a logged deviation, Appendix A), n=600, computed from the Qwen logs on the replication question IDs), memory adds $+8.5$pt, and redacting stored answers removes most of this gain. This is the answer-dependence posited by C1, but in the beneficial direction: the gain depends on the stored answers and search does not fall, rising from 2.21 to 2.44 (1.73 to 1.91 among answered episodes). Memory does not increase the answerless rate detectably on this agent ($+1.5$pt, 90% CI $[-1.5, +4.3]$), so the termination effect also differs across agents. Enforcement has a $+1.2$pt higher point estimate than instruction (43 fixes/36 breaks, p=.50, 90% CI $[-1.2, +3.5]$); the direct contrast is not significant. Both instruction arms carry A6/A8 on this agent, and no `+Complete` arm was run, so their gains are not separated from completion mechanics. We read the memory, redaction, and enforcement contrasts of §4.1 and §4.3 as specific to the agents studied rather than as general properties of memory; the mechanics and optimizer results (§4.2, §4.4) were not replicated. Within Qwen3.5-27B, MuSiQue alone shows the same signature (memory $+2.0$pt, redaction -3.3pt; no test; Table 15).

5 Discussion

Interpreting the Qwen3.5-27B result. In this environment, Qwen3.5-27B already searches (2.6 searches per answered episode) and is not detectably anchored by candidate answers supplied directly (`Guess` is null on EM). Deleting the bank's procedures (`Mem-AnsOnly`) leaves EM within the null floor, which is consistent with procedures that describe behavior the agent already performs; redacting its answers (`Mem-NoAns`) also stays within the floor, and answer-bearing variants yield only small point-estimate gains ($+1.7$ to $+2.2$pt for `Guess` and `Mem-AnsOnly`) without reducing search. The main behavioral change associated with memory is a higher rate of answerless termination: more episodes exhaust the budget without answering, while search per answered episode is unchanged. Completion mechanics reduce this termination behavior; content manipulations do not.

Implications for memory evaluation. For Qwen3.5-27B, memory gains measured against an empty instruction slot and default episode mechanics cannot be attributed to the bank alone. Our mechanics + instruction baseline without a bank (`NoMem+Cite`, EM .520; no storage or retrieval infrastructure, and 36% fewer tokens per episode than `Mem+Cite`) trails no memory configuration by more than 3pt on point estimates. For the one registered pair, `NoMem+Cite` to `Mem+Cite`, the question-cluster 90% CI is $[+0.2, +3.8]$ (Table 13; CI in Table 8); the highest point estimate

in the study also has no bank (GEPA@NoMem, .550, descriptive only). We suggest reporting (i) a completion-mechanics baseline with a filled instruction slot (we did not detect a distinct benefit of evidential content in the available control), (ii) an answer-redacted bank, and (iii) a cross-seed null floor.

Token cost. On Qwen3.5-27B, the bank raises mean tokens per episode from 5.4k (NoMem) to 9.6k (Mem) and, with the instruction, from 10.4k (NoMem+Cite) to 16.3k (Mem+Cite); completion mechanics alone use 7.8k tokens (NoMem+Complete). On gpt-oss-120b, the same bank increases cost from 5.8k to 9.7k tokens per episode and is associated with a $+8.5$pt gain (Tables 3, 4, and 16). Memory retrieval costs approximately 1.6–1.8$\times$ the tokens in both settings; the measured accuracy benefit differs across agents.

Metric sensitivity. Exact match penalizes semantically correct answers and offers a surface-form optimization target; paired designs stay internally valid, but optimizer gains and absolute EM levels need semantic audits before they count as improvement (§4.4).

6 LIMITATIONS

The study covers one environment family (multi-hop QA over fixed corpora) and two agents; a capability sweep would strengthen the cross-agent conclusion. The bank is modest (199 items); the capacity arms vary only the retrieved count (top-7 vs top-3), not bank size. The bank distills solved episodes only (an ExpeL-inspired recipe) from the question, search count, and final answer rather than the executed trajectory. The registered gold-document decontamination rule requires the overlap count to be reported: a post-hoc title-level check finds 12 of 199 bank items and 23 of 150 S_{opt} questions sharing a gold-document title with a test question; these overlaps touch 51 of 500 test and 34 of 300 replication questions; excluding them moves Qwen contrasts by less than 0.7pt (primary $+0.1$pt, 90% CI $[-2.0, +2.2]$) and gpt-oss contrasts by at most 0.9pt away from zero (Table 8); other contamination is not ruled out. The redacted bank retains an audited 4–8% semantic leak (Appendix E), which biases the redaction contrasts on both agents toward zero. The null floor is the maximum over the five registered replicates; other same-arm seed pairs move more (Mem s11/s23 3.2pt, Guess 3.8pt), close to the $\pm$4pt margin, so floor statements are read with that spread in mind. The mid-gating amendments A6/A8 applied completion mechanics only to instruction-bearing arms; the post-hoc ablation of §4.2 quantifies this asymmetry, but the mechanics-matched memory comparison has one seed per hardware setup (A100 $+2.4$ $[-0.2, +5.0]$; A6000 $+2.4$pt $[+0.0, +4.8]$) and neither interval settles effect or equivalence, so the primary equivalence claim rests on the registered NoMem vs. Mem contrast only. On the second agent the instruction arms carry the same mechanics and no mechanics-only arm was run, so their gains are not separated from completion mechanics. Per-episode logs record final answers, call counts, citations, and scores but not retrieved passages or model text, so failure modes are classified only at the outcome level.

7 CONCLUSION

Under pre-registered controls, the NoMem–Mem contrast for Qwen3.5-27B is equivalent within the registered $\pm$4pt margin; in post-hoc single-seed blocks, episode-completion mechanics alone reproduce most of the evidence-citation arms' gain (three of four decompositions; in the fourth the instruction component is larger), and a residual memory effect of about $+2.4$pt at matched mechanics is neither established nor ruled out on either hardware setup. A no-bank mechanics + instruction baseline trails no memory configuration by more than 3pt at under two-thirds of the token cost of Mem+Cite. On gpt-oss-120b, memory helps and redaction hurts, so the measured effects differ across the two agents studied. Memory evaluations should include matched instruction and completion-mechanics controls. We will release the gate history, amendments, and artifacts so others can run the same audit.

AI USE STATEMENT

Large language models are the object of this study (the agents, the memory distiller and redactor, the optimizer arms, and the frontier audit annotators; §3, Appendices D and E). In the research process we used generative AI (Anthropic Claude models, through a coding assistant) for tasks that the ICLR

policy lists as requiring disclosure: *hypotheses and methodology design* (the registered claims and arm design were drafted with LLM assistance, partly through an ideation pipeline derived from The AI Scientist-v2 (Yamada et al., 2025), then revised by the authors over three design-review rounds before the freeze); *implementing methods and data analysis* (the harness, analysis scripts, and table and figure generators; the optimizer-reflection step of the GEPA-lite and MIPRO-lite arms, deviation D6, with all prompts released); and *interpreting results*, where automated adversarial reviews and analysis drafts proposed readings that the authors checked against the logs before adopting or rejecting them. We additionally used it for tasks the policy lists as recommended disclosures: *retrieval and discovery* (part of the related work, each entry then checked against its source) and *drafting and editing* the manuscript, including automated review passes whose suggestions the authors accepted or rejected individually. Generative AI was not used to generate or clean benchmark data, to translate, or to develop theory or proofs. Verification: the registered contrasts were recomputed from the per-episode logs by a re-analysis script separate from the experiment code, every arXiv citation was verified against its source (Appendix H), and the harness passed smoke runs and four gate rounds before the matrix ran. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics statement

Public benchmark data only; no human subjects; the per-episode logs contain no personal data; compute (allocation time, estimated) $\sim$6 GPU-days on 4$\times$A100 plus about one GPU-day on 4$\times$RTX A6000 for the second-environment blocks.

Reproducibility statement

Code, prompts (verbatim, hash-pinned), splits (instance IDs), per-episode JSONL logs, the gate history, amendments, and optimizer reflections will be released upon publication; the serving stack is pinned (vLLM 0.19.1). The logs record final answers, call and search counts, citations, scores, token counts, and configuration hashes; retrieved passages and model text were not stored. Audit records are aggregate summaries with the sampling seed; per-item verdicts were not retained. Tables 3, 11–14, 16, 17, and the C2 fragment are generated by `gen_tables.py` from the logs and then renamed and reformatted by the release scripts with values unchanged; the token column of Tables 3, 4, and 16 is transcribed from the released re-analysis summary; Tables 2, 10, 1, 4, 5, and 15 are transcribed from those outputs and from the released re-analysis script; Table 6 is generated by `gen_transitions.py` from the logs, Tables 7 and 8 by `scripts/cite_transitions.py` and `scripts/r04_sensitivity.py`, and Table 9 by `scripts/analyze_rerun2.py`. Per-episode seeds are logged, with a documented caveat that concurrent batched inference is not bit-reproducible (Appendix F).

References

Lakshya A. Agrawal et al. GEPA: Reflective prompt evolution can outperform reinforcement learning. *arXiv preprint arXiv:2507.19457*, 2025.

Andre Bacellar. BridgeRAG: Training-free bridge-conditioned retrieval for multi-hop question answering. *arXiv preprint arXiv:2604.03384*, 2026.

Nishant Bhargava et al. MemAudit: An exact package-oracle evaluation protocol for budgeted long-term LLM memory writing. *arXiv preprint arXiv:2605.02199*, 2026.

Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, Andrew Liu, Joshua Green, Kshama Patel, Ruoxi Meng, Mingyi Su, Sahel Sharifymoghaddam, Yanxi Li, Haoran Hong, Xinyu Shi, Xuye Liu, Nandan Thakur, Crystina Zhang, Luyu Gao, Wenhu Chen, and Jimmy Lin. BrowseComp-Plus: A more fair and transparent evaluation benchmark of deep-research agent. *arXiv preprint arXiv:2508.06600*, 2025.

Quang Dao, Purvi Kathalkar, and Kenneth Eaton. Weighted memory tree: Remembering what matters for long-horizon LLM agents. *arXiv preprint arXiv:2608.20631*, 2026.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *International Conference on Computational Linguistics (COLING)*, 2020.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 1979.

Stockton Jenkins, Ramya Korlakai Vinayak, and Junjie Hu. GRASP: Graph agentic search over propositions for multi-hop question answering. *arXiv preprint arXiv:2605.16598*, 2026.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. DSPy: Compiling declarative language model calls into self-improving pipelines. In *International Conference on Learning Representations (ICLR)*, 2024.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *ACM Symposium on Operating Systems Principles (SOSP)*, 2023.

Daniël Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4), 2017.

Zhengxuan Lu, Dongfang Li, Yukun Shi, Beilun Wang, Longyue Wang, and Baotian Hu. Structured episodic event memory. *arXiv preprint arXiv:2601.06411*, 2026.

Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11), 2018.

OpenAI. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Krista Opsahl-Ong, Michael J. Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. Optimizing instructions and demonstrations for multi-stage language model programs. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Siru Ouyang, Jun Yan, et al. ReasoningBank: Scaling agent self-evolving with reasoning memory. *arXiv preprint arXiv:2509.25140*, 2025.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

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

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4), 2009.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Xiaolong Sun, Qichao Wang, Hangyu Li, and Liang Chen. Verifiable memory: Learning unified memory management with local and global verifiers for large language model agents. *arXiv preprint arXiv:2608.03137*, 2026.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics (TACL)*, 10, 2022.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The AI Scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Shengtao Zhang, Jiaqian Wang, Ruiwen Zhou, Junwei Liao, Yuchen Feng, Zhuo Li, Yujie Zheng, Weinan Zhang, Ying Wen, Zhiyu Li, Feiyu Xiong, Yutao Qi, Bo Tang, and Muning Wen. MemRL: Self-evolving agents via runtime reinforcement learning on episodic memory. *arXiv preprint arXiv:2601.03192*, 2026.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. ExpeL: LLM agents are experiential learners. In *AAAI Conference on Artificial Intelligence*, 2024.

A Pre-registration protocol and complete gate history

Freeze. DESIGN v2 (33 arms, claims C1–C3, families, thresholds) frozen 2026-08-24 after three adversarial review rounds (R1: 7 validity + 8 statistics blocking findings; R2: 6+4; R3: 3), all adopted; SHA-256 of design, thresholds, instructions, splits, and code recorded in `prereg/FREEZE.sha256`. The design, threshold, instruction, and split entries are unchanged since 2026-08-24; the code entries were re-issued at each logged execution-level fix (dates in `DEVIATIONS.md`), and the final `agent.py` entry postdates the matrix because it adds the post-hoc `mech` flag. `AMENDMENTS.md` was appended after the freeze and its manifest entry is stale. Post-freeze edits require logged deviations. Fourteen entries were logged: nine execution-level entries before the analysis, including a client-argument bug; engine-version correction 0.13.0→0.19.1; embedder GPU→CPU; error-record retry semantics; a `Mem+Budget`/`Mem-Curated` (registered M6/M8) dispatch omission; an incomplete replication-model download and its substitution; and the frontier-assistant substitution for optimizer reflection (deviation D6); and five entries dated 2026-09-02 to 09-04 that register the post-hoc arms (`+Complete`, `Mem-BM25`), the replication's second seed, and their results before they entered the paper. Two further departures are recorded here rather than in the deviation log: gate G-6 was registered as a 150-item human check and run as an automatic substring check on all 199 items plus a 25-item frontier semantic audit, and the seed-23 half of `Mem-Curated+Enforce` was dropped in the registered drop order.

Registered but not reported. The interaction family, the pre-specified slice family, the citation audit separating stated from executed evidence under claim C3, and the registered minimum detectable effects for subgroup claims are not reported in this paper.

Gate history (thresholds never modified).

Gate	threshold	R1	R2	R3	R4 (final)
G-1 bank effect (dev)	fixes$\geq$6 & breaks$\geq$6	11/8 pass	24/17 pass	pass	pass
G-2 `+Cite` compliance	$\geq$65/100	64 fail	68 pass	81 pass	80 pass
G-3 non-echo breaks	$\geq$10 & $\geq$20%	8 fail	17 (100%) pass	pass	pass
G-4 item truthfulness	$\geq$70%	100% pass	pass	pass	pass
G-6 redaction leak (auto; see text)	$\leq$10%	13.1% fail	0% pass	pass	pass
G-8 `+Enforce` unparseable	$<$5%	27% fail	19% fail	8% fail	**3%** pass

The four round-1 failures were resolved by: an iterative redaction repair (G-6; 26 leaky items, final mechanical neutral-filler fallback, then 0% mechanical leak and a 25-item frontier semantic audit at $\sim$4–8%); amendment A7 extending S_{dev1} 150→300 by deterministic continuation of the registered shuffle (G-1/G-3 re-measured at n=300); amendment A6 (no-finish budget extension) and A8 (terminal forced-finish, + a max-token/parse hardening after a replay demonstrated truncation races) for G-2/G-8. Diagnostics that motivated each amendment, and the archived pre-amendment measurements (`*.pre_A6`, `*.pre_A8`, `*.pre_A8b`), are in the release. G-5 (closed-book rate

≤60%): 18.6%, pass. G-7 (retriever relevance): audited at gate time, pass; the released gate log has no dated entry for this audit.

Null floor. Five treatment-free replicates: `NoMem` s11/s37 2.2pt; `Mem` 0.2pt; `Mem+Cite` 2.2pt; `Mem+Enforce` 0.4pt; `NoMem` s11/s23 2.8pt ⇒ floor **2.8pt**.

B COMPLETE ARM TABLE AND CONFIGURATIONS

Table 1 reproduces the registered arm table. Pinned selection: the top-k selection is computed once on original-bank embeddings and shared by `Mem`/`Mem-NoAns`/`Mem-Swap`/`Mem-AnsOnly`/`Mem-Top3`/`Mem-NoAns-Top3` (top-1 by `Guess`/`Guess-NoHedge`); `Mem-Decoy` is unpinned by design, since its manipulated factor is selection relevance. Per-arm JSON configs and episode mechanics (including A6/A8) ship in the release.

Table 1: The registered arm table (DESIGN v2 §4, frozen 2026-08-24; all arms on S_{test}, row 33 on its 300-question subset). Seeds $\{11, 23\}$ mark 2-seed decisive arms; seed 37 rows are treatment-free null replicates. Seeds and episode counts are as run: row 16 was registered at seeds 11,23 (1,000 episodes) and its seed-23 half was dropped in the registered drop order; row 33 was registered at seed 11 (1,500) and gained a second seed after the freeze; the registered exploratory frontier probe, outside the 33 slots, is unrun. Rows are the 33 numbered slots of the registered design; rows 29–33 bundle replicates and the replication. Post-hoc arms (Table 4) have log IDs `M0_MECH`, `M1_MECH`, `M1B`. *Registered* = the arm ID in the pre-registration and in the released log filenames. Name grammar: `+Cite` = I^* as instruction (soft floor); `+Enforce` = I^* with quote verification (hard floor); `GEPA@`/`MIPRO@` = optimizer-evolved instruction (soft floor); `+Tools` = I_B (soft floor); no suffix = empty instruction slot. All instruction-bearing arms (`+Cite`, registered Fs; `+Enforce`, Fh; `GEPA@`; `MIPRO@`; `+Tools`) run with the completion mechanics A6/A8; `+Complete` (post-hoc, Table 5) enables those mechanics alone; `+Tools` carries the bookkeeping instruction I_B (`Tools: <list>`).

#	Arm	Registered	Memory	Seeds	Episodes
1	`NoMem`	`M0`	none	11,23	1,000
2	`Mem`	`M1`	top-7	11,23	1,000
3	`Mem-NoAns`	`M2`	redacted	11,23	1,000
4	`Guess`	`ANCHOR`	own-guess	11,23	1,000
5	`NoMem+Cite`	`M0_Fs`	none	11,23	1,000
6	`Mem+Cite`	`M1_Fs`	top-7	11,23	1,000
7	`Mem-NoAns+Cite`	`M2_Fs`	redacted	11,23	1,000
8	`Guess+Cite`	`ANCHOR_Fs`	own-guess	11,23	1,000
9	`NoMem+Enforce`	`M0_Fh`	none	11,23	1,000
10	`Mem+Enforce`	`M1_Fh`	top-7	11,23	1,000
11	`GEPA@Mem`	`G1_M1`	top-7	11,23	1,000
12	`GEPA@NoMem`	`G0_M0`	none	11,23	1,000
13	`Mem-Top3`	`M5`	top-3	11,23	1,000
14	`Mem-NoAns-Top3`	`M5r`	redacted top-3	11,23	1,000
15	`Mem-Curated`	`M8`	curated	11,23	1,000
16	`Mem-Curated+Enforce`	`M8_Fh`	curated	11	500
17	`MIPRO@Mem`	`MIPRO_M1`	top-7	11,23	1,000
18	`Mem-NoAns+Enforce`	`M2_Fh`	redacted	11	500
19	`Guess+Enforce`	`ANCHOR_Fh`	own-guess	11	500
20	`Mem-Swap`	`M2i`	info-matched	11	500
21	`Guess-NoHedge`	`ANCHOR_NH`	own-guess, no hedge	11	500
22	`Mem-AnsOnly`	`M3`	answer-only	11	500
23	`Mem-Decoy`	`M4`	decoy	11	500
24	`Mem+Budget`	`M6`	top-7, +3 calls	11	500
25	`GEPA@Mem+Enforce`	`G1_Fh`	top-7	11	500
26	`NoMem+Tools`	`IB_M0`	none	11	500
27	`Mem+Tools`	`IB_M1`	top-7	11	500
28	`NoMem-ClosedBook`	`M0_CB`	none (closed book)	11	500
29–32	null replicates		`NoMem`, `Mem`, `Mem+Cite`, `Mem+Enforce` at seed 37		2,000
33	replication		`NoMem`, `Mem`, `Mem-NoAns`, `Mem+Cite`, `Mem+Enforce` on gpt-oss-120b, $S_{test}[300]$, seeds 11,23 (registered: seed 11)		3,000

C FULL RESULTS

Tables 3, 11–14, 16, and the C2 fragment are generated by `gen_tables.py` directly from the per-episode JSONL logs, then renamed and reformatted by the release scripts with values unchanged; Table 4 and the token columns are transcribed from the released re-analysis summary; Table 2 (registered presence and content family) and Table 5 (post-hoc mechanics ablation) give fixes, breaks, and exact McNemar p for the contrasts drawn in Figures 1 and 2; Table 6 is generated by `gen_transitions.py` from the logs; Table 7 (instruction gains by seed), Table 8 (title-overlap sensitivity), and Table 9 (the second-environment block) by the scripts named in the reproducibility statement. Table 3: complete per-arm results (registered arms). Table 4: the post-hoc arms. Tables 11–13: the three registered Holm families; question-cluster 90% CIs for every contrast are released with the logs. Table 14: the treatment-free null floor. Table 15: per-dataset contrasts. Table 16: the replication agent. Exploratory arms outside the registered families (`GEPA@NoMem`, `Mem-NoAns+Enforce`, `Guess+Enforce`, `Mem-Curated+Enforce`) are descriptive-only as registered; `Guess-NoHedge` serves only the C2 guard; `NoMem-ClosedBook` serves only gate G-5.

C2 branch metric. Registered branch metric, on the full paired set (n=1,000):

$$R_A = \frac{\overline{\text{calls(NoMem)} - \text{calls(Guess)}}}{\overline{\text{calls(NoMem)} - \text{calls(Mem)}}} = \frac{-0.198}{-0.412} = 0.48.$$

Both differences are *negative* (memory and anchor each *increase* calls), so the registered branch (thresholds on truncation $\geq$0.60/$<$0.30) is **inapplicable**: there is no truncation to mediate, and C2 is reported as moot rather than decided.

Table 2: Memory presence and content effects, Qwen3.5-27B, S_{test} (paired; 2 seeds pooled n=1,000 unless noted). The registered null floor (max treatment-free cross-seed $|\Delta|$) is **2.8pt**.

Comparison	Manipulates	fixes	breaks	ΔEM	p	verdict
`NoMem` vs `Mem` (**primary**)	memory presence	65	59	+0.6pt	.654	**TOST-equivalent**
`Mem` vs `Mem-NoAns`	answers redacted	22	30	−0.8pt	.332	ns
`Mem` vs `Mem-Swap`(n=500)	answer→other fact	7	12	−1.0pt	.359	ns
`Mem` vs `Mem-AnsOnly`(n=500)	answers only	24	13	+2.2pt	.099	ns
`Mem` vs `Mem-Decoy`(n=500)	decoy items	15	15	0.0pt	1.0	ns
`Mem` vs `Mem-Top3`	top-7→top-3	34	33	+0.1pt	1.0	ns
`Mem` vs `Mem-NoAns-Top3`	redacted top-3	29	37	−0.8pt	.389	ns
`NoMem` vs `Guess` (provenance)	candidate answer only	69	52	+1.7pt	.145	ns

Table 3: Complete per-arm results on S_{test} (Qwen3.5-27B). *tokens* = prompt plus completion tokens summed over the episode's calls, mean per episode; *forced* = A8 terminal forced-finish rate. Blank EM_{s23} = arm run at seed 11 only.

Arm	n	EM	EM_{s11}	EM_{s23}	F1	searches	calls	tokens	forced
`NoMem` none	1,000	.478	.464	.492	.585	3.31	4.33	5,357	0.0%
`Mem` bank top-7	1,000	.484	.468	.500	.571	3.95	4.75	9,633	0.0%
`Mem-NoAns` redacted	1,000	.476	.474	.478	.568	3.94	4.74	9,659	0.0%
`Mem-Top3` bank top-3	1,000	.485	.484	.486	.578	3.92	4.74	7,662	0.0%
`Mem-NoAns-Top3` redacted top-3	1,000	.476	.466	.486	.568	3.93	4.75	7,774	0.0%
`Guess` own-guess	1,000	.495	.476	.514	.596	3.53	4.53	6,071	0.0%
`NoMem+Cite`	1,000	.520	.512	.528	.637	4.19	5.54	10,376	13.3%
`Mem+Cite`	1,000	.540	.550	.530	.643	4.66	5.88	16,343	17.0%
`Mem-NoAns+Cite`	1,000	.523	.516	.530	.635	4.75	5.95	16,692	16.5%
`Guess+Cite`	1,000	.524	.532	.516	.635	4.18	5.54	11,338	13.3%
`NoMem+Enforce`	1,000	.523	.520	.526	.639	4.32	6.29	11,966	15.9%
`Mem+Enforce`	1,000	.534	.544	.524	.642	4.76	6.34	17,647	19.2%
`Mem-Curated` curated	1,000	.478	.468	.488	.569	3.94	4.74	9,607	0.0%
`GEPA@Mem` GEPA	1,000	.543	.548	.538	.646	4.73	6.34	17,842	14.7%
`GEPA@NoMem` GEPA	1,000	.550	.558	.542	.652	4.23	5.91	11,448	12.0%
`MIPRO@Mem`	1,000	.548	.546	.550	.647	4.86	6.00	16,394	16.0%
`Mem-Swap` info-swap	500	.458	.458		.550	3.93	4.71	9,819	0.0%
`Mem-AnsOnly` answer-only	500	.490	.490		.578	3.77	4.61	7,068	0.0%
`Mem-Decoy` decoys	500	.468	.468		.553	4.03	4.78	9,745	0.0%
`Mem+Budget` bank, +3 calls	500	.494	.494		.585	4.69	5.57	14,000	0.0%
`Guess-NoHedge` no-hedge	500	.490	.490		.592	3.36	4.48	5,877	0.0%
`Mem-NoAns+Enforce`	500	.534	.534		.647	4.87	6.47	18,172	17.2%
`Guess+Enforce`	500	.526	.526		.641	4.22	6.07	12,281	14.4%
`Mem-Curated+Enforce`	500	.530	.530		.640	4.76	6.44	18,062	18.4%
`NoMem+Tools`	500	.522	.522		.637	7.78	9.01	25,868	62.0%
`Mem+Tools`	500	.548	.548		.652	9.89	11.02	42,069	84.4%
`NoMem-ClosedBook` closed-book	500	.186	.186		.288	0.00	1.33	601	25.8%
`GEPA@Mem+Enforce`	500	.542	.542		.645	4.91	7.58	21,349	20.4%

Table 4: Post-hoc arms (outside the registered design; S_{test}, seed 11, n=500). *forced* = A8 terminal forced-finish rate; *answerless* = episodes with no final answer; *tokens* as in Table 3. Paired contrasts use question-cluster 90% CIs.

Arm	EM	F1	searches	calls	tokens	forced	answerless
`NoMem+Complete` (`M0_MECH`)	.506	.631	3.57	4.84	7,790	8.0%	1.8%
`Mem+Complete` (`M1_MECH`)	.530	.639	4.63	5.73	15,169	18.4%	4.6%
`Mem-BM25` (`M1B`; BM25 memory retriever)	.466	.555	4.07	4.84	9,929	0.0%	29.2%

Paired contrasts: `NoMem`→`NoMem+Complete` $+4.2$ $[+2.2, +6.2]$; `Mem`→`Mem+Complete` $+6.2$ $[+4.4, +8.0]$ (Table 5); `Mem`→`Mem-BM25` -0.2 $[-1.6, +1.2]$, fixes 9 / breaks 10, p=1.0.

Table 5: Post-hoc mechanics ablation (added after an automated adversarial review; outside the registered families; S_{test}, seed 11 only, n=500). `+Complete` = A6/A8 completion mechanics enabled with the instruction slot empty and no floor (the A8 terminal prompt itself asks for one Evidence line; §3); the last row is the mechanics-matched memory test. 90% CI: question-cluster bootstrap, 10k resamples; the $\pm$4pt margin is a reference, not a registered test.

Comparison	n	fixes	breaks	ΔEM(pt)	p	90% CI
`Mem` → `Mem+Complete`	500	32	1	$+6.2$	7.9×10^{-9}	$[+4.4, +8.0]$
`NoMem` → `NoMem+Complete`	500	30	9	$+4.2$	.0011	$[+2.2, +6.2]$
`Mem+Complete` → `Mem+Cite`	500	28	18	$+2.0$	.18	$[-0.2, +4.2]$
`NoMem+Complete` → `NoMem+Cite`	500	26	23	$+0.6$	.78	$[-1.6, +2.8]$
`NoMem+Complete` → `Mem+Complete`	500	36	24	$+2.4$	.155	$[-0.2, +5.0]$

Answerless-episode rates, bare arms, seed 11: `NoMem` 16.2%, `Mem` 27.6% (pooled seeds 11+23: 14.0%, 25.2%).
Forced-finish fired in 18.4% of `Mem+Complete` episodes.

Table 6: Paired outcome transitions for the completion-mechanics contrasts (Qwen3.5-27B, S_{test}). Each episode is classed as correct (EM 1), wrong (an answer with EM 0), or answerless (no final answer; every such bare-arm episode ended at the 8-call cap). Rows: the episode's outcome under the base arm; columns: its outcome under the treatment arm on the same question and seed. The last three columns describe the treatment arm's terminal forced-finish answers: how many episodes ended that way, their EM, and the EM of episodes that finished on their own. Searches by outcome (seeds 11+23): answered episodes average 2.62 in both `NoMem` and `Mem`; answerless episodes 7.6 and 7.9.

			under the base arm		under the treatment arm			forced finish (treatment)		
Contrast	seeds	n	outcome	n	correct	wrong	answerless	n	EM	EM, unforced
`NoMem` → `Mem`	11+23	1,000	correct	478	419	21	38	0	—	0.647
			wrong	382	55	230	97			
			answerless	140	10	13	117			
`NoMem` → `NoMem+Complete`	11	500	correct	232	223	9	0	40	0.150	0.548
			wrong	187	12	174	1			
			answerless	81	18	55	8			
`Mem` → `Mem+Complete`	11	500	correct	234	233	1	0	92	0.207	0.639
			wrong	128	5	123	0			
			answerless	138	27	88	23			
`Mem+Complete` → `Mem+Cite`	11	500	correct	265	247	14	4	88	0.273	0.661
			wrong	212	24	168	20			
			answerless	23	4	11	8			

Table 7: Where the instruction arms' gains come from, by seed (Qwen3.5-27B, S_{test}, n=500 per seed). Outcomes are classed as in Table 6. Fixes are split by the base outcome they came from; breaks are correct base episodes that became wrong or answerless. An accounting decomposition of the paired net gain, not a causal share.

Contrast	seed	EM base→treat	answerless base→treat	fixes: wrong	fixes: answerless	breaks	net	forced n	forced EM	unforced EM
`Mem` → `Mem+Cite`	11	0.468→0.550	27.6%→6.4%	22	32	13	+41	88	0.273	0.661
`Mem` → `Mem+Cite`	23	0.500→0.530	22.8%→7.4%	12	23	20	+15	82	0.256	0.640
`NoMem` → `NoMem+Cite`	11	0.464→0.512	16.2%→5.8%	23	19	18	+24	63	0.238	0.591
`NoMem` → `NoMem+Cite`	23	0.492→0.528	11.8%→3.8%	27	10	19	+18	70	0.271	0.596

Table 8: Title-overlap sensitivity (registered rule R04). A test question is excluded if its gold-document title is shared by a bank item (21 questions) or an S_{opt} question (39); the union removes 51 of 500 S_{test} questions (46 musique, 3 2wikimultihopqa, 2 hotpotqa) and 34 of the 300 replication questions. Retrieval and answers are the logged ones; only the question set changes. Question-cluster 90% CIs.

Contrast	seeds	n_q	fixes/breaks	ΔEM	90% CI	set
`NoMem` → `Mem`	11+23	500	65/59	+0.6	[−1.4, +2.7]	all
`NoMem` → `Mem`	11+23	449	55/54	+0.1	[−2.0, +2.2]	overlap removed
`NoMem+Cite` → `Mem+Cite`	11+23	500	67/47	+2.0	[+0.2, +3.8]	all
`NoMem+Cite` → `Mem+Cite`	11+23	449	61/41	+2.2	[+0.3, +4.1]	overlap removed
`Mem` → `Mem+Complete`	11	500	32/1	+6.2	[+4.4, +8.0]	all
`Mem` → `Mem+Complete`	11	449	30/1	+6.5	[+4.5, +8.5]	overlap removed
`NoMem` → `NoMem+Complete`	11	500	30/9	+4.2	[+2.2, +6.2]	all
`NoMem` → `NoMem+Complete`	11	449	29/9	+4.5	[+2.2, +6.7]	overlap removed
`Mem+Complete` → `Mem+Cite`	11	500	28/18	+2.0	[−0.2, +4.2]	all
`Mem+Complete` → `Mem+Cite`	11	449	27/17	+2.2	[−0.2, +4.7]	overlap removed
`NoMem+Complete` → `Mem+Complete`	11	500	36/24	+2.4	[−0.2, +5.0]	all
`NoMem+Complete` → `Mem+Complete`	11	449	31/23	+1.8	[−0.9, +4.5]	overlap removed
`NoMem` → `Mem` (gpt-oss)	11+23	300	76/25	+8.5	[+5.7, +11.5]	all
`NoMem` → `Mem` (gpt-oss)	11+23	266	69/21	+9.0	[+6.0, +12.2]	overlap removed
`Mem` → `Mem-NoAns` (gpt-oss)	11+23	300	29/61	−5.3	[−7.8, −2.8]	all
`Mem` → `Mem-NoAns` (gpt-oss)	11+23	266	24/57	−6.2	[−8.8, −3.6]	overlap removed

Table 9: Completion decomposition re-run on a second hardware setup (A6000, seed 23: 4×A6000, two TP=2 replicas, same code, prompts, bank, and questions; n=500 questions per arm). Block 1 (bank) and block 2 (no bank) were each fixed before running (`PLAN.md` in the release); they share environment, seed, and questions, so the cross-block rows are valid paired contrasts, whereas nothing here is pooled with or subtracted from the A100 block. Paired question bootstrap 90% CIs; exact McNemar p is descriptive.

Arm	block	EM	F1	answerless	forced	searches	calls	tokens
`NoMem`	2	0.482	0.592	12.6%	0.0%	3.33	4.34	5,318
`NoMem+Complete`	2	0.504	0.624	1.0%	8.8%	3.55	4.72	7,159
`NoMem+Cite`	2	0.530	0.636	5.4%	12.2%	4.09	5.39	9,763
`Mem`	1	0.490	0.584	21.8%	0.0%	3.91	4.73	9,530
`Mem+Complete`	1	0.528	0.628	2.4%	17.8%	4.51	5.56	14,124
`Mem+Cite`	1	0.532	0.633	8.0%	18.6%	4.74	5.92	16,402

Contrast	block	fixes	breaks	ΔEM	90% CI	p	Δanswerless	90% CI
`Mem` → `Mem+Complete`	block 1	21	2	+3.8	[+2.4, +5.4]	6.6e-05	−19.4	[−22.4, −16.6]
`Mem+Complete` → `Mem+Cite`	block 1	22	20	+0.4	[−1.8, +2.6]	.878	+5.6	[+3.6, +7.8]
`Mem` → `Mem+Cite`	block 1	36	15	+4.2	[+1.8, +6.6]	.005	−13.8	[−16.6, −11.0]
`NoMem` → `NoMem+Complete`	block 2	15	4	+2.2	[+0.8, +3.6]	.019	−11.6	[−13.8, −9.4]
`NoMem+Complete` → `NoMem+Cite`	block 2	35	22	+2.6	[+0.2, +5.2]	.111	+4.4	[+3.0, +5.8]
`NoMem` → `NoMem+Cite`	block 2	44	20	+4.8	[+2.2, +7.4]	.004	−7.2	[−9.4, −5.0]
`NoMem` → `Mem`	cross	34	30	+0.8	[−1.8, +3.4]	.708	+9.2	[+6.4, +12.0]
`NoMem+Complete` → `Mem+Complete`	cross	32	20	+2.4	[+0.0, +4.8]	.126	+1.4	[+0.0, +2.8]
`NoMem+Cite` → `Mem+Cite`	cross	28	27	+0.2	[−2.2, +2.6]	1.0	+2.6	[+0.2, +5.0]

Table 10: Break-reduction family F-A4 (vs `Mem`), Holm-corrected within family; seeds 11+23 pooled, n=1,000 unless noted. The instruction-bearing arms carry the completion mechanics A6/A8 (§4.2); `Mem+Budget` and `Mem-Curated` carry neither. "recov." = `Mem`'s breaks recovered (59 pooled; 33 at seed 11 for the n=500 rows). Rows with n=500 are seed 11 only and are compared with seed-11 `Mem`.

Arm	EM	searches	fixes/breaks vs `Mem`	recov.	p	Holm
`Mem` (bank only)	.484	3.95	reference			
`Mem+Cite` (I^* asked)	.540	4.66	89/33	36/59	4.0×10^{-7}	sig.
`GEPA@Mem` (GEPA)	.543	4.73	85/26	33/59	1.7×10^{-8}	sig.
`MIPRO@Mem`	.548	4.86	88/24	30/59	9.0×10^{-10}	sig.
`Mem+Enforce` (I^* enforced)	.534	4.76	78/28	26/59	1.3×10^{-6}	sig.
`GEPA@Mem+Enforce`(n=500)	.542	4.91	51/14	17/33	4.5×10^{-6}	sig.
`Mem+Budget`(n=500) (budget only)	.494	4.69	16/3	7/33	.0044	sig.
`Mem-Curated` (curation)	.478	3.94	9/15	5/59	.307	ns

Table 11: Family F-A4 (break-reduction vs `Mem`), Holm-corrected within family.

Comparison	n	fixes	breaks	ΔEM(pt)	p	Holm
`Mem` bank top-7 → `Mem+Cite`	1,000	89	33	+5.6	4.0×10^{-7}	sig.
`Mem` bank top-7 → `GEPA@Mem` GEPA	1,000	85	26	+5.9	1.7×10^{-8}	sig.
`Mem` bank top-7 → `MIPRO@Mem`	1,000	88	24	+6.4	9.0×10^{-10}	sig.
`Mem` bank top-7 → `Mem-Curated` curated	1,000	9	15	−0.6	.307	ns
`Mem` bank top-7 → `Mem+Enforce`	1,000	78	28	+5.0	1.3×10^{-6}	sig.
`Mem` bank top-7 → `GEPA@Mem+Enforce`	500	51	14	+7.4	4.5×10^{-6}	sig.
`Mem` bank top-7 → `Mem+Budget` bank, +3 calls	500	16	3	+2.6	.0044	sig.

Table 12: Family F-A5 (content manipulations vs `Mem`), Holm-corrected within family.

Comparison	n	fixes	breaks	ΔEM(pt)	p	Holm
`Mem` bank top-7 → `Mem-NoAns` redacted	1,000	22	30	−0.8	.332	ns
`Mem` bank top-7 → `Mem-Swap` info-swap	500	7	12	−1.0	.359	ns
`Mem` bank top-7 → `Mem-AnsOnly` answer-only	500	24	13	+2.2	.099	ns
`Mem` bank top-7 → `Mem-Decoy` decoys	500	15	15	+0.0	1.0	ns
`Mem` bank top-7 → `Mem-Top3` bank top-3	1,000	34	33	+0.1	1.0	ns
`Mem` bank top-7 → `Mem-NoAns-Top3` redacted top-3	1,000	29	37	−0.8	.389	ns

Table 13: Family F-A2 (does memory still add value once the instruction slot is filled?), Holm-corrected.

Comparison	n	fixes	breaks	ΔEM(pt)	p	Holm
`NoMem+Cite` → `Mem+Cite`	1,000	67	47	+2.0	.075	ns
`Mem-NoAns+Cite` → `Mem+Cite`	1,000	53	36	+1.7	.089	ns
`Guess+Cite` → `Mem+Cite`	1,000	56	40	+1.6	.125	ns
`NoMem+Tools` → `Mem+Tools`	500	34	21	+2.6	.105	ns

Table 14: Treatment-free null floor: identical condition, different seed. The registered floor is the maximum row.

Arm	replicate	EM_a	EM_b	$\lvert\Delta\rvert$(pt)
`NoMem` none	s11 vs s37	.464	.486	2.2
`Mem` bank top-7	s11 vs s37	.468	.466	0.2
`Mem+Cite`	s11 vs s37	.550	.528	2.2
`Mem+Enforce`	s11 vs s37	.544	.540	0.4
`NoMem` none	s11 vs s23	.464	.492	2.8

Table 15: Per-dataset paired ΔEM (pt), Qwen3.5-27B; seeds 11+23 pooled, n = paired episodes; no per-corpus test. The primary null and the slot effect are similar across corpora; the answer-content signature appears only on MuSiQue.

Comparison	HotpotQA (n=500)	MuSiQue (n=300)	2Wiki (n=200)
`NoMem`→`Mem`	-0.2	$+2.0$	$+0.5$
`Mem`→`Mem+Cite`	$+5.4$	$+4.7$	$+7.5$
`Mem`→`Mem-NoAns` (redaction)	0.0	-3.3	$+1.0$

Table 16: Replication agent gpt-oss-120b on the registered subset (seeds 11+23 pooled, n=600; EM/F1-only claims). *tokens* as in Table 3. Searches among answered episodes: `NoMem` 1.73, `Mem` 1.91.

Arm	n	EM	F1	searches	tokens
`NoMem` none	600	.312	.476	2.21	5,806
`Mem` bank top-7	600	.397	.538	2.44	9,655
`Mem-NoAns` redacted	600	.343	.506	2.31	9,067
`Mem+Cite`	600	.443	.599	3.62	16,721
`Mem+Enforce`	600	.455	.599	4.47	23,769
Paired contrasts (name, n, fixes, breaks, ΔEM(pt), p):					
`NoMem` none → `Mem` bank top-7	600	76	25	+8.5	3.7×10^{-7}
`Mem` bank top-7 → `Mem-NoAns` redacted	600	29	61	-5.3	.00097
`Mem` bank top-7 → `Mem+Cite`	600	66	38	+4.7	.0078
`Mem` bank top-7 → `Mem+Enforce`	600	67	32	+5.8	.00056

D OPTIMIZER CURVES, REFLECTIONS, AND CHAMPION INSTRUCTIONS

Budgets. `GEPA@Mem` 330/350 rollouts; `GEPA@NoMem` 270/350; `MIPRO@Mem` 255/350; shared deterministic minibatches (seed 20260831). **Curves (per-candidate minibatch EM).** `GEPA@Mem`: I^* .587 (n=75) → champion `cand_001` .648 (n=105); the five gen-2 stacking mutations were evaluated on two minibatches each (means .600–.700, n=30) and none was promoted, so the champion is a single clause. `GEPA@NoMem`: champion .600 vs seed .560; validation flat (.533/.556), reported descriptively. `MIPRO@Mem`: grounded proposal `cand_006` .707 (n=75), head-to-head .689 vs .622. **Champion texts** are inlined verbatim in Appendix G. All reflection prompts and outputs (frontier assistant, D6) are published in `runs/opt/*/reflect_*.md`. Table 17 lists every candidate's full trajectory.

Table 17: Complete optimizer trajectories: per-candidate minibatch EM (shared deterministic batches of 15).

Run	Candidate	minibatch EM	mean	n	
GEPA@Mem	cand_000	b0:.60, b1:.60, b4:.47, b5:.80, b6:.47	.587	75	seed I^*
GEPA@Mem	cand_001	b0:.67, b1:.80, b2:.67, b3:.60, b4:.53, b5:.80, b6:.47	.648	105	**champion**
GEPA@Mem	cand_002	b0:.53, b1:.73	.633	30	
GEPA@Mem	cand_003	b0:.67, b1:.73	.700	30	
GEPA@Mem	cand_004	b0:.53, b1:.67	.600	30	
GEPA@Mem	cand_005	b0:.60, b1:.60	.600	30	
GEPA@Mem	cand_006	b0:.60, b1:.67	.633	30	
GEPA@NoMem	cand_000	b0:.60, b1:.53, b3:.47, b4:.47, b5:.73	.560	75	seed I^*
GEPA@NoMem	cand_001	b0:.67, b1:.67, b2:.67, b3:.47, b4:.47, b5:.67	.600	90	**champion**
GEPA@NoMem	cand_002	b0:.60, b1:.73, b2:.60	.644	45	
GEPA@NoMem	cand_003	b0:.53, b1:.60	.567	30	
GEPA@NoMem	cand_004	b0:.60	.600	15	
GEPA@NoMem	cand_005	b0:.67	.667	15	
MIPRO@Mem	cand_000	b0:.53, b1:.60	.567	30	seed I^*
MIPRO@Mem	cand_001	b0:.60, b1:.87, b2:.67, b3:.67, b4:.53	.667	75	
MIPRO@Mem	cand_002	b0:.60	.600	15	
MIPRO@Mem	cand_003	b0:.60	.600	15	
MIPRO@Mem	cand_004	b0:.60, b1:.73	.667	30	
MIPRO@Mem	cand_005	b0:.53	.533	15	
MIPRO@Mem	cand_006	b0:.67, b1:.80, b2:.73, b3:.80, b4:.53	.707	75	**champion**

E FRONTIER AUDITS

Redaction semantic audit (25 seeded items): 23 clean; 1 semantic leak (word-order variant defeating the substring check); 1 borderline (unique-birthdate identifier). Estimated semantic leak 4–8% vs the 10% gate. **EM scorer audit** (15 seeded EM=0, F1≥.5 episodes): 15/15 semantically correct (missing-alias and gold-format artifacts); implications discussed in §4.4.

F INFRASTRUCTURE AND DETERMINISM

4×A100-80GB; vLLM 0.19.1; agent replicas TP=1 on 3 GPUs, annotator/replication on the fourth; CPU-pinned MiniLM memory retriever. Determinism caveat: with per-call seeds fixed, a failed episode succeeded on solo replay, because concurrent batching is not bit-reproducible; we therefore treat seeds as noise control rather than as replay guarantees, and publish per-episode outcome logs instead (retrieved passages and model text were not stored). **Second environment.** The two blocks in Table 9 ran on 4×RTX A6000 48GB (driver 535.86) with two tensor-parallel-2 vLLM 0.19.1 replicas (max-num-seqs 32, max-model-len 16384), the same bf16 weights, and byte-identical code, prompts, bank, and question set; the pilot, the throughput rule that fixed the block size before any outcome was read, the second block's pre-registration, and the analysis are in `PLAN.md` of that release directory. Because the A100 arms were not re-run there, no contrast crosses the two environments; contrasts between the two A6000 blocks share environment, seed, and questions.

G VERBATIM PROMPTS

All prompts below are extracted mechanically from the hash-pinned sources (`prereg/instructions.json`, `code/agent.py`, `code/build_bank.py`, champion candidate files).

System prompt (SYS_BASE; {instruction}/{memory} are the manipulated slots)

```
You are a research assistant answering questions using a search tool over
    a fixed document collection.

On each turn, do ONE of:
  Search[your query]   — retrieve relevant passages
  Finish[final answer] — give the final answer (a short phrase)

Think briefly before acting. Use multiple searches when the question has
   multiple parts.{instruction}{memory}
```

I^* — registered evidence instruction (hash-pinned)

```
Before giving a final answer, output a line 'Evidence: "<exact quote>" |
   "<exact quote>"' quoting passages you retrieved in this episode that
   support the answer. If you cannot quote supporting passages, search
   for more evidence first.
```

I_B — registered bookkeeping control instruction

```
Before giving a final answer, output a line 'Tools: <list>' stating which
    searches you ran this episode and in what order.
```

Hard-floor rejection message (`+Enforce`)

```
Your answer was not accepted: before finishing you must output a line '
   Evidence: "<exact quote 1>" | "<exact quote 2>"' quoting passages
   retrieved IN THIS EPISODE that support the answer. If your current
   evidence is insufficient, Search again first.
```

A8 terminal forced-finish message

```
You have no search budget left. You MUST now answer. Respond with at most
    one short Evidence line, then Finish[your best answer] (a short
   phrase). Do NOT Search. Keep the response brief.
```

`GEPA@Mem` champion (cand_001)

```
Before giving a final answer, output a line 'Evidence: "<exact quote>" |
   "<exact quote>"' quoting passages you retrieved in this episode that
   support the answer. If you cannot quote supporting passages, search
   for more evidence first. When you answer, copy the answer EXACTLY as
   it is written in the passage you are quoting — same spelling, same
   singular/plural, same digits-or-words — and give ONLY that short
   phrase, with no added qualifiers like country names or extra
   descriptors.
```

`GEPA@NoMem` champion (cand_001)

```
Before giving a final answer, output a line 'Evidence: "<exact quote>" |
   "<exact quote>"' quoting passages you retrieved in this episode that
   support the answer. If you cannot quote supporting passages, search
   for more evidence first. When you answer, copy the answer EXACTLY as
   it is written in the passage you are quoting — same spelling, same
```

```
    singular/plural, same digits-or-words — and give ONLY that short
    phrase, with no added units or qualifiers.
```

`MIPRO@Mem` champion (cand_006)

```
Answer in the exact surface form found in the source text (spelling,
    singular/plural, digits vs words). For multi-hop questions resolve
    the bridge entity first. Quote your supporting line as Evidence:
    "..." before Finish.
```

Bank distillation prompt (Qwen3-32B)

```
You will see a question-answering episode that ended CORRECTLY. Distill
    it into a reusable memory item.
Format exactly:
INSIGHT: <one transferable lesson about how to solve questions like this>
PROCEDURE: <which searches, in what order, and how to disambiguate>
OUTCOME: The answer was '<final answer>'.

Episode:
Question: {q}
Searches: {s}
Final answer: {a}
```

Answer-redaction prompt (`Mem-NoAns` construction)

```
Rewrite this memory item with EVERY answer, result value, or completion
    statement removed, keeping the procedure intact. Replace removed
    content with neutral filler of similar length so total length stays
    within 5%. Do NOT reveal what the answer was, even indirectly.

Item:
{item}

Rewritten item:
```

H CITATION VERIFICATION APPENDIX

As part of our verification procedure, every external identifier was checked against its source (retrieval dates 2026-09-02 for the original entries, 2026-09-17 for the five entries added in revision, and a full re-check of all thirteen arXiv identifiers on 2026-09-25, which corrected one entry); Table 18. All arXiv identifiers listed in Table 18 were re-checked on 2026-09-25.

Table 18: Citation verification log: externally checked identifiers, retrieval dates 2026-09-02 (original entries) and 2026-09-17 (entries added in revision, marked †); ‡ = title, authors and venue corrected on 2026-09-25 after a re-check of every arXiv identifier (the earlier entry carried a different title and an unverified venue).

Key	ID	Source	Exact title (verified)
agrawal2025gepa	arXiv:2507.19457	arxiv.org	GEPA: Reflective Prompt Evolution Can Outperform Reinforcement Learning
ouyang2025reasoningbank	arXiv:2509.25140	arxiv.org	ReasoningBank: Scaling Agent Self-Evolving with Reasoning Memory
openai2025gptoss	arXiv:2508.10925	arxiv.org	gpt-oss-120b & gpt-oss-20b Model Card
qwen3report	arXiv:2505.09388	arxiv.org	Qwen3 Technical Report
qwen2026blog	vendor blog	qwen.ai	Qwen3.5: Towards Native Multimodal Agents (no arXiv TR for the dense 27B line as of retrieval)
memrl2026	arXiv:2601.03192	arxiv.org	MemRL: Self-Evolving Agents via Runtime Reinforcement Learning on Episodic Memory
bhargava2026memaudit	arXiv:2605.02199	arxiv.org	MemAudit: An Exact Package-Oracle Evaluation Protocol for Budgeted Long-Term LLM Memory Writing
browsecompplus2026‡	arXiv:2508.06600	arxiv.org	BrowseComp-Plus: A More Fair and Transparent Evaluation Benchmark of Deep-Research Agent
bridgerag2026	arXiv:2604.03384	arxiv.org	BridgeRAG: Training-Free Bridge-Conditioned Retrieval for Multi-Hop Question Answering
sun2026vermem†	arXiv:2608.03137	arxiv.org	Verifiable Memory: Learning Unified Memory Management with Local and Global Verifiers for Large Language Model Agents
dao2026wmt†	arXiv:2608.20631	arxiv.org	Weighted Memory Tree: Remembering What Matters for Long-Horizon LLM Agents
jenkins2026grasp†	arXiv:2605.16598	arxiv.org	GRASP: Graph Agentic Search over Propositions for Multi-hop Question Answering
lu2026seem†	arXiv:2601.06411	arxiv.org	Structured Episodic Event Memory
yamada2025aiscientistv2†	arXiv:2504.08066	arxiv.org	The AI Scientist-v2: Workshop-Level Automated Scientific Discovery via Agentic Tree Search