MEMINI: CONTINUAL LEARNING THROUGH ASSOCIATIVE MEMORY IN LLM SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

An LLM system can preserve every incoming document yet lose the ability to use earlier knowledge because new and old evidence compete within finite retrieval and context budgets. We introduce Memini, an external associative memory for continual learning from streams of unlabeled textual observations with all pretrained components held fixed. Drawing on episodic binding and cue-driven recall, Memini preserves each observation as an immutable source episode, while query-independent updates bind recurring concepts to episodes and accumulate source-attributable co-observations across them. At query time, partial cues follow bounded associative routes to recover complete source episodes, either as a standalone readout or in combination with dense retrieval. Reading does not modify persistent memory. On streams constructed from MuSiQue, 2WikiMultiHopQA, and HotpotQA, we evaluate acquisition, retention, and the joint use of evidence received at different updates. Under these schedules, Memini's hybrid readout reduces mean answer-F1 forgetting by 50–72% relative to HippoRAG2 using its released incremental update path. It also improves cross-arrival answer F1 by 1.44–7.71 points over dense retrieval alone. Shuffling concept connections degrades retained answering even with the source archive unchanged, indicating that the organization of stored experience contributes beyond source availability. These results support source-grounded associative memory as a mechanism for system-level continual learning without adapting pretrained models.

1 INTRODUCTION

As new documents arrive after deployment, an LLM system must use their information alongside earlier knowledge. Preserving every document does not guarantee this: new and earlier evidence compete within finite retrieval and context budgets. The system must make new information usable, retain earlier capabilities, and combine evidence received at different times. We study this continual-learning problem by treating incoming documents as a stream of unlabeled textual observations, with all pretrained components held fixed. The system adapts by updating its persistent external memory.

Structured external memories can improve question answering (Rajesh et al., 2026). Yet studies report degraded QA performance as retrieval corpora expand (Gutiérrez et al., 2025) and interference between new and earlier experience during retrieval in frozen-model agents (Hu et al., 2026). Moving adaptation from model weights to external memory therefore does not by itself resolve the stability–plasticity tension. This raises a central question: how should arriving observations reshape persistent memory organization to keep new and earlier knowledge accessible and usable together?

We introduce *Memini*[1], an external associative memory inspired by episodic binding and cue-driven recall (Staresina & Davachi, 2009; Rolls, 2013). It preserves each observation as an immutable source episode and connects episodes through recurring concepts. When concepts occur together in an episode, Memini records that episode as a witness to their association. Witnesses accumulate as new episodes arrive, before future questions are known and without re-ingesting historical sources. At query time, partial cues follow bounded, witness-supported routes through these concepts to recover complete source episodes, including those that do not contain the initial query cue (Figure 1). Memini supports associative recall alone or combines it with dense retrieval, allowing learned connections to complement direct semantic matching.

[1] *Meminī* is Latin for "I remember" or "I hold in mind".

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

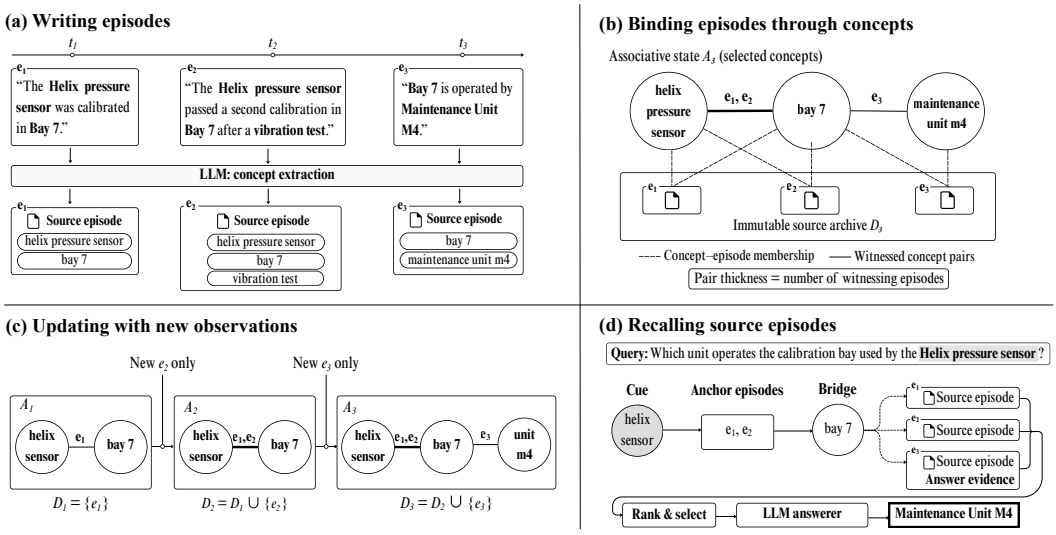


Figure 1: **Memini learns associations that guide episodic recall. (a)** Each observation is preserved as a complete episode and an LLM extracts salient concepts. **(b–c)** Recurring concepts connect episodes; co-observations across distinct episodes add support without reprocessing earlier sources. **(d)** A later question cues Helix. Witnessed routes activate Bay 7, and additional contributing anchors can reinforce this activation. Bay 7 reinstates complete episodes, with e_3 naming the operator.

We evaluate acquisition, retention and cross-arrival recombination, the joint use of evidence received at different updates, on streams constructed from three multi-hop QA benchmarks. Memini combines competitive acquisition with strong answering after subsequent updates. Across all three datasets, its hybrid readout reduces mean answer-F1 forgetting by 50–72% relative to HippoRAG2 and improves cross-arrival answer F1 by 1.44–7.71 points over dense retrieval alone. With source availability fixed, shuffling concept connections degrades retained answering. Together, these findings show how learned associative organization helps keep new and earlier knowledge accessible and useful as memory grows.

2 Problem Formulation

We consider a deployed LLM system that receives an ordered stream of unlabeled textual observations and answers questions using the information received so far. The learner is the system as a whole: all pretrained components remain fixed, while persistent external state changes as observations arrive. The system must remain able to answer questions throughout the stream. In this setting, we assume earlier information remains valid: later observations add evidence without superseding earlier facts. The problem is therefore not resolving contradictions or deciding which conflicting claim is currently true, but continuing to acquire and access evidence as it accumulates.

After update t, the persistent state contains an archive of source observations D_t and an organization A_t through which evidence from those sources is accessed. This state may grow as observations accumulate. Updates are replay-free: the memory writer processes each incoming observation once. Subsequent updates use the new observation and retained memory state without passing earlier sources through the writer again. Updates are also independent of evaluation questions and receive no target answers, gold evidence, task identities, or task-boundary signals. At query time, the system retrieves evidence within finite retrieval and context budgets and answers from it without modifying persistent state. The retrieved evidence remains traceable to its originating observations.

The objective is to recover relevant evidence (*access*) and answer correctly from it (*successful use*) throughout the stream. Meeting this objective requires three capabilities. Information newly supported by the observed stream must become usable (*acquisition*) and remain usable after subsequent updates (*retention*) (Parisi et al., 2019; De Lange et al., 2022). When an answer depends on several observations received at different updates, successful use further requires combining their evidence

(*cross-arrival recombination*). All three should be achieved while limiting *interference*, the loss of earlier capabilities caused by later updates.

3 Related Work

Continual learning methods for LLMs typically update model parameters and must manage both the cost of updates and interference across optimization steps (Wu et al., 2024; Shi et al., 2025). Such updates may also be unavailable to downstream users who access models through inference APIs (Sun et al., 2022). We instead keep pretrained components fixed and adapt persistent external memory, following systems that frame structured memory as non-parametric continual learning (Gutiérrez et al., 2025; Rajesh et al., 2026). We compare these and other external memory systems by what they store, how they retrieve evidence, and what their reported evaluations establish; Table F1 in Appendix F summarizes these distinctions.

Retrieval-augmented generation (RAG) stores passages in a flat index (Lewis et al., 2020). Structured systems add derived organization: RAPTOR builds hierarchical summaries (Sarthi et al., 2024), GraphRAG constructs entity graphs and community summaries (Edge et al., 2025), and LightRAG combines graph and vector indexes with incremental document integration (Guo et al., 2025). HippoRAG2 extracts OpenIE triples and integrates phrases and passages into a shared graph that grows as sources arrive (Gutiérrez et al., 2025), while Panini constructs a separate workspace for each document, containing entities, events, and generated QA pairs (Rajesh et al., 2026). Memini preserves each observation as an immutable source episode and records concept co-observations with episode-specific witnesses, without requiring named relations. Memory updates accumulate support across distinct episodes, with each contribution traceable to its source.

HippoRAG2 runs Personalized PageRank from query-dependent seeds, diffusing activation through the corpus graph to rank source passages (Gutiérrez et al., 2025). Memini also returns source episodes, but uses bounded, witness-supported routes through recurring concepts rather than graph diffusion. Panini decomposes queries into subquestions and retrieves and reranks QA pairs to assemble evidence chains across workspaces. Its answerer receives the selected QA pairs, and the authors report improved multi-hop answering with less answer context (Rajesh et al., 2026). Returning complete sources avoids an additional answer-time dependence on generated QA representations, although extraction and routing errors can still prevent Memini from retrieving required episodes.

In their corpus-expansion studies, HippoRAG2 and Panini evaluate fixed question cohorts whose supporting evidence is available initially. HippoRAG2 maintains higher F1 than dense retrieval, although both degrade at similar rates on MuSiQue (Gutiérrez et al., 2025). Panini adds only distractors and reports smaller declines than BM25 and direct embedding retrieval (Rajesh et al., 2026). Related interference also arises in sequential interactive tasks. Hu et al. (2026) attribute forgetting in frozen agents to newer memories dominating retrieval despite earlier experience remaining stored. Together, these findings highlight the challenge of maintaining earlier capabilities as external memory grows. Our evaluation complements the corpus-expansion studies above by separately measuring acquisition when supporting observations arrive and retention of those same questions after further updates, through both retrieval and answering, alongside cross-arrival recombination.

Models of human episodic memory and associative learning motivate a functional blueprint for supporting these capabilities through rapid encoding of distinct experiences (McClelland et al., 1995; Kumaran et al., 2016), strengthening associations among co-occurring elements (Hebb, 1949), and recovering complete memories from partial cues while keeping experiences distinct (Rolls, 2013). Memini adapts these principles to persistent external memory by preserving individual episodes while allowing experience to shape the associations through which their evidence is later recovered.

4 Memini: Experience-Driven Associative Memory

Memini stores observations as source episodes in D_t and represents the organization A_t through shared concepts, concept–episode memberships, and witnessed concept associations (Figure 1). All persistent changes occur during writing, while reading computes temporary query-dependent activation over the resulting memory.

4.1 Memory Formation from Experience

For each arriving episode e, an LLM writer identifies salient local concepts, assigns them coarse types, and groups coreferent mentions: textual spans referring to the same concept. A deterministic checker accepts a proposed mention only if it occurs in the declared title or body; concept identification and coreference grouping remain semantic decisions of the LLM.

The integrator receives only the concepts with accepted mentions and maps those whose checked names are equal after normalization to the same global concept g. Identity is lexical and independent of the assigned types. Within episode e, it consolidates this evidence for g into one concept–episode membership. Let $n_m(g, e)$ count the mentions after consolidation and $I_{\text{title}}(g, e) \in \{0, 1\}$ indicate whether any occurs in the source title. With the title bonus $\beta_{\text{title}} = 0.5$, the membership weight is

$$w_{\text{mem}}(g, e) = \log_2(2 + n_m(g, e)) \; (1 + \beta_{\text{title}} \, I_{\text{title}}(g, e)) \, . \tag{1}$$

The logarithmic term gives repeated mentions diminishing influence; a title mention signals prominence, multiplying the weight by 1.5. For titleless episodes, $I_{\text{title}}(g, e) = 0$.

From the same accepted evidence, Memini records all unordered pairs of distinct local concepts. After global mapping, these records are consolidated into at most one *pair observation* per episode and unordered pair of distinct global concepts. Whereas memberships link individual concepts to episodes, pair observations link two concepts through a witnessing episode, supporting recall from a cue concept to a co-observed concept. Each record retains its witnessing episode, shared-sentence evidence, and whether either concept has a title mention; sharing a body sentence is not required for co-observation. Later route scoring uses these fields and the episode's distinct global-concept count. Repeated mentions can strengthen a membership without adding another witness for the same pair.

4.2 Accumulation of Associative Support

Support for a concept pair consists of persistent co-observation evidence contributed by distinct episodes. For distinct global concepts a and b, let $\mathcal{W}_t(a, b)$ denote their witness set and $n_t(a, b)$ their accumulated *exposure*:

$$\begin{aligned} \mathcal{W}_t(a, b) &= \{e \in D_t \mid \{a, b\} \text{ is recorded as a pair observation in } e\} \, , \\ n_t(a, b) &= |\mathcal{W}_t(a, b)| \, . \end{aligned} \tag{2}$$

The set preserves source attribution, while its size counts distinct witnessing episodes. In Figure 1, when e_2 arrives, the Helix–Bay 7 witness set changes from $\{e_1\}$ to $\{e_1, e_2\}$, while e_1's stored text and local evidence remain unchanged. During recall, Memini reads this evidence without modifying it and computes temporary, query-dependent activation that guides source-episode retrieval.

4.3 Query-Conditioned Associative Recall

Recall follows a bounded route from cue concept a through a witnessing anchor episode e and co-observed bridge concept b to a target episode u containing b. In Figure 1, Helix is the cue, calibration episode e_1 or e_2 is the anchor, Bay 7 is the bridge, and operator episode e_3 is the target. Stored witnesses provide candidate anchors; the question determines which contribute.

Cue entry. An LLM extracts concepts explicitly named in question q. Memini matches these cues to existing global concepts by name or checked mention surface and normalizes positive match masses to obtain cue mass $c_q(g)$, using the matching and ambiguity rules in Appendix C.3.

Route scoring. Let $\mathcal{E}(g)$ contain all episodes with a membership for concept g. The normalized membership share is

$$m(g, e) = \frac{w_{\text{mem}}(g, e)}{\sum_{e' \in \mathcal{E}(g)} w_{\text{mem}}(g, e')} . \tag{3}$$

These shares distribute activation across a concept's episodes, with $m(g, e) = 0$ where no membership exists. Let s_e count distinct global concepts in e. The sentence factor $\rho(e, a, b)$ is 1 for a shared body sentence and 0.25 otherwise; the title factor $\tau(e, a, b)$ is 1.5 if either concept has an accepted title mention and 1 otherwise. Local pair quality is

$$q_{\text{pair}}(e, a, b) = \min\left(1, \frac{\rho(e, a, b) \, \tau(e, a, b)}{\log_2(2 + s_e)}\right) . \tag{4}$$

This favors sentence and title evidence while discounting episodes with many concepts. With frozen embeddings $\mathbf{z}_q$ and $\mathbf{z}_y$ of the complete question and episode or concept name y, semantic focus is $g_q(y) = \max(0, \cos(\mathbf{z}_q, \mathbf{z}_y))$. For witnessing anchor $e \in \mathcal{W}_t(a, b)$, the route contribution is

$$r_q(a, e, b) = c_q(a)\, m(a, e)\, g_q(e)\, q_{\text{pair}}(e, a, b)\, g_q(b). \tag{5}$$

Semantic similarity thus operates within associative recall, scoring routes constrained by stored associations rather than ranking the archive directly. Stored pairs are unordered; a and b denote their query-specific cue and bridge roles.

Bridge activation. Memini retains the strongest cue contribution for each anchor–bridge pair (e, b). Let $k_q(b)$ count distinct anchors with positive retained contributions and $r_q^{\max}(b)$ be their maximum. For an active bridge, its activation is

$$\alpha_q(b) = r_q^{\max}(b) + \big(1 - r_q^{\max}(b)\big)\, \big(1 - \exp\big[-r_q^{\max}(b)\, (k_q(b) - 1)\big]\big)\,. \tag{6}$$

With one contributing anchor, activation equals the strongest route value; additional anchors provide bounded reinforcement through their count, with diminishing increments toward one. Bridges without positive routes have zero activation. Unlike stored pair exposure $n_t(a, b)$, $k_q(b)$ counts anchors contributing to this question across cues, each counted once per bridge.

Episode selection. Each target episode u receives its strongest bridge contribution:

$$S_q(u) = \max_b \left[\alpha_q(b)\, m(b, u)\, g_q(u)\right]. \tag{7}$$

An episode therefore gains no additive boost merely because several bridges reach it. Unreached episodes score zero; positive scores are normalized to sum to one and ranked, preserving their order. Parallel routes can recover several episodes despite the fixed traversal depth, including episodes that do not contain the original cue.

Memini's associative memory supports two readouts, associative and hybrid (Figure 2). The associative readout selects the five highest-ranked positive-scoring episodes, or all if fewer exist. The hybrid combines this ranking with an independent dense ranking of the same archive. Dense retrieval matches episodes to the complete question without requiring witnessed routes, but may miss indirectly relevant evidence; associative recall can expose it through stored connections when usable routes exist. Under the evaluated five-episode budget, the hybrid retains the dense top four and adds the highest-ranked positive associative episode not already selected, using the next dense episode as fallback. Both readouts supply the question and complete source episodes, with provenance, to the same frozen LLM answerer.

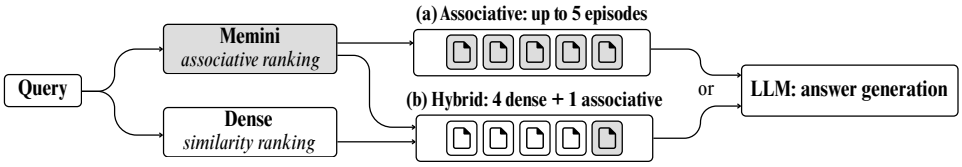


Figure 2: **Two readouts under the evaluated five-episode budget. (a)** The associative readout selects up to five episodes from the associative ranking. **(b)** The hybrid readout retains the dense top four and adds the first positive-scoring associative episode not already selected, with dense fallback. Each readout separately supplies the question and complete episodes to the same LLM answerer.

5 Experiments

We evaluate Memini's acquisition, retention and cross-arrival recombination on observation streams from three multi-hop QA benchmarks. Controlled interventions test how learned connections and reinforcement affect its performance. We also assess model sensitivity, scalability and computational cost.

5.1 Shared Experimental Setup

We use HippoRAG2's 1,000-question evaluation sets for MuSiQue, 2WikiMultiHopQA, and HotpotQA, each with a shared collection of supporting and distractor episodes (Table 1). MuSiQue

includes linear and branching chains, 2WikiMultiHopQA includes two-support questions and four-support bridge comparisons, and HotpotQA uses two supporting episodes. Gold annotations define schedules and scores but are not supplied to memory construction, retrieval, or answering.

Table 1: Multi-hop evaluation sets.

Dataset	Questions	Episodes	Required supports
MuSiQue (Trivedi et al., 2022)	1,000	11,656	2–4
2WikiMultiHopQA (Ho et al., 2020)	1,000	6,119	2 or 4
HotpotQA (Yang et al., 2018)	1,000	9,811	2

We compare Memini's associative and hybrid readouts with lexical BM25 (Robertson & Zaragoza, 2009), dense embedding retrieval (Dense), and HippoRAG2 (Gutiérrez et al., 2025). To compare evidence access and successful use under a common source-episode readout, all methods receive the same arriving sources and supply at most five complete episodes to the same GPT-4o-mini answerer, with identical prompts, parsing, and decoding settings. Memini, Dense, and HippoRAG2 share frozen NV-Embed-v2 (Lee et al., 2025). We exclude systems such as Panini because they supply generated QA pairs to the answerer, making answering depend additionally on how source information is reformulated (Rajesh et al., 2026).

Memini uses GPT-5.4 without additional reasoning for episode writing and cue extraction; HippoRAG2 uses GPT-4o-mini for triple extraction and query-time filtering. We vary Memini's writer and cue model separately (Section 5.5). At each arrival, we pass only new passages to HippoRAG2's released indexing routine on its existing graph. Model settings and prompts appear in Appendices A.1 and D.

We measure evidence access with Recall@5, the mean fraction of required supporting episodes retrieved per question, and successful use with answer exact match (EM) and token F1. We use two arrival schedules (Figure 3), one introducing successive question cohorts and the other separating dependent supporting episodes across arrivals.

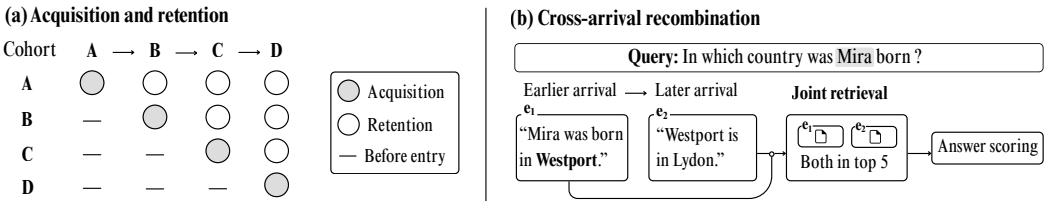


Figure 3: **(a)** We evaluate each cohort at entry for acquisition and after later updates for retention. **(b)** We measure joint evidence retrieval and answering after the later support arrives.

5.2 Acquisition and Retention Across Updates

We first ask whether Memini acquires usable knowledge from incoming observations and retains it through subsequent updates (Figure 3a). We divide each evaluation set into four cohorts of 250 questions and use four cyclic orders. After integrating each cohort's new episodes, we test its questions and retest those already evaluated. All orders reach the same final episode collection. Table 2 summarizes acquisition at entry and final retention for cohorts followed through subsequent updates. Complete@5 (C@5) measures how often all required supporting episodes are retrieved together.

On MuSiQue, combining learned associations with dense retrieval gives Memini's hybrid readout the highest retained R@5 and F1. Gains over Dense reach 5.20 R@5 and 5.03 F1 points, with 2.72 F1 points over HippoRAG2. Its acquisition F1 is also the highest observed, but its lead over HippoRAG2 has a paired 95% confidence interval including zero.

On 2Wiki, Memini's associative readout supports competitive acquisition and stronger evidence recovery and answering after updates. Complete support recovery reaches 80.30% versus HippoRAG2's 63.70%, with F1 gains of 3.33 points over HippoRAG2 and 8.54 over Dense. This advan-

Table 2: **Acquisition and retention (%).** Acquisition covers all four entry positions; retention excludes last-entry cohorts, which have no later updates. Subtracting these means does not give backward transfer (BWT). Bold and underlining mark best and second-best per dataset and column.

Dataset	Method	Acquisition ↑				Retention ↑			
		R@5	C@5	EM	F1	R@5	C@5	EM	F1
MuSiQue	BM25	41.48	10.18	19.83	30.10	40.60	9.90	19.20	29.34
	Dense	71.26	40.08	33.83	46.86	69.38	37.50	32.50	45.19
	HippoRAG2	76.26	**49.45**	35.75	49.92	73.04	44.23	33.57	47.50
	Memini (associative)	61.24	35.73	31.93	44.54	59.69	33.90	31.30	43.61
	Memini (hybrid)	**76.46**	49.35	**37.63**	**51.16**	**74.58**	**46.20**	**36.90**	**50.22**
2Wiki	BM25	64.48	31.88	42.60	48.26	63.93	31.20	42.40	48.14
	Dense	77.70	52.43	53.33	60.38	76.58	50.50	52.90	59.82
	HippoRAG2	**91.48**	76.95	**60.63**	**70.66**	84.81	63.70	56.33	65.04
	Memini (associative)	91.46	**81.58**	60.40	69.32	**90.85**	**80.30**	**59.70**	**68.37**
	Memini (hybrid)	89.98	71.20	58.55	68.10	88.80	69.00	57.80	67.23
HotpotQA	BM25	71.44	47.50	45.48	58.58	70.90	46.60	45.50	58.57
	Dense	95.00	90.23	58.63	73.12	94.40	89.10	58.70	73.01
	HippoRAG2	96.34	92.95	**58.93**	**73.97**	96.07	92.53	57.90	72.97
	Memini (associative)	86.48	77.48	54.25	69.02	85.55	75.80	53.90	68.56
	Memini (hybrid)	**96.59**	**93.48**	58.88	73.66	**96.15**	**92.60**	**59.40**	**73.92**

tage extends to the 235 four-support bridge-comparison questions requiring two evidence branches, with retained R@5 19.96 points above HippoRAG2 and F1 of 80.63 versus 74.02 (Table B8).

On HotpotQA, Dense recovers nearly all required evidence. Despite this limited retrieval headroom, Memini's hybrid readout reaches 96.15% retained R@5 versus Dense's 94.40% and improves retained EM over HippoRAG2 by 1.50 points, with both gains supported by paired 95% intervals.

Figure 4 tracks answering after acquisition. Episodes remain stored, but required evidence may fall outside the five-episode context or the LLM may use it incorrectly (e.g., Liu et al., 2024). Backward transfer (BWT) measures final minus entry F1; forgetting measures the decline from the best observed F1 to the final score. Both are averaged over matched question/order trajectories with later updates.

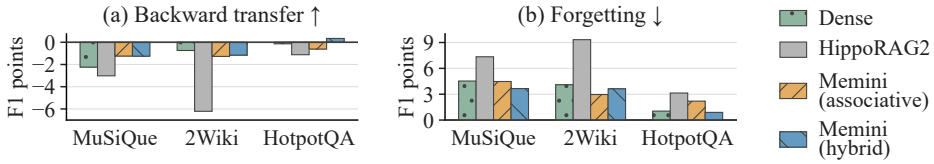


Figure 4: **Answering stability of Dense, HippoRAG2, and Memini.** Bars show point estimates in F1 points; higher BWT and lower forgetting are better. Paired uncertainty appears in Appendix B.2.

Memini's hybrid readout is more resistant to forgetting than HippoRAG2, with 50–72% lower mean F1 forgetting across all three datasets and paired intervals supporting each reduction. Its smaller observed forgetting reductions versus Dense have intervals including zero, while it retains higher F1 on MuSiQue and 2Wiki. Together with the retained scores, these results demonstrate strong answering after updates accompanied by limited loss of previously attained performance.

5.3 Cross-Arrival Recombination

We next test whether Memini recovers earlier evidence together with related evidence arriving later (Figure 3b). Each dataset follows one dependency-based schedule: four arrival stages for MuSiQue and two for 2Wiki and HotpotQA. Each separated dependency pair contributes one evaluation event when its later episode arrives. Joint@5 tests whether both episodes appear in the top five; EM/F1 score the original question. Other required supports may still be pending on MuSiQue.

Table 3: **Cross-arrival recombination (%).** Scores are averaged over events, so questions with more events contribute more. Paired comparisons appear in Appendix B.3. Bold and underlining mark best and second-best per dataset and metric.

Method	MuSiQue			2Wiki			HotpotQA		
	Joint@5	EM	F1	Joint@5	EM	F1	Joint@5	EM	F1
BM25	9.37	15.64	24.92	13.27	38.78	44.30	42.96	43.70	57.25
Dense	36.73	25.08	35.89	32.04	47.35	54.12	86.54	<u>58.02</u>	73.09
HippoRAG2	<u>44.46</u>	<u>27.42</u>	<u>39.28</u>	<u>66.12</u>	<u>52.76</u>	<u>62.31</u>	**91.85**	57.90	<u>73.97</u>
Memini (associative)	35.53	25.21	36.10	**80.31**	**56.73**	**65.00**	74.20	53.33	68.83
Memini (hybrid)	**45.54**	**28.12**	**39.75**	65.92	52.35	61.83	<u>90.86</u>	**59.01**	**74.52**

Combining associative recall with dense retrieval improves joint recovery and answering over Dense on all three datasets (Table 3). Hybrid gains of 1.44–7.71 F1 points are supported by paired 95% intervals; its F1 differences from HippoRAG2 have intervals including zero.

Memini's associative readout alone provides the clearest joint-recovery advantage on 2Wiki: it retrieves the pair in 80.31% of events, exceeding HippoRAG2 by 14.18 Joint@5 and 3.98 EM points, with both gains supported by paired intervals. Together, these gains show stronger cross-arrival evidence recovery alongside more exact answers.

5.4 Contributions of Associative Organization

Importance of learned connections. At each checkpoint, we shuffle the destination concepts of associative routes while preserving source episodes, concept–episode memberships and each route's original witness and quality. The reader uses each route's retained quality to activate its new destination concept, then follows that concept's existing memberships to score episodes. The modified route remains usable even when the new destination is absent from the original witness episode, so shuffling redirects associative access. Retention F1 falls by 31.29–56.68 points with associative recall and 1.96–8.59 with hybrid recall across the three datasets (Appendix B.4). The hybrid losses occur despite its four fixed dense selections, showing that Memini's learned connections contribute to retained answering both through associative recall and alongside semantic retrieval.

Additional contribution of reinforcement. We remove the activation boost from additional anchors while retaining the original concept connections, the strongest route per bridge and the pair qualities. Retention F1 changes by at most 0.71 points with associative recall and 0.22 with hybrid recall (Table 4). Removing reinforcement slightly improves scores on MuSiQue and 2Wiki and slightly reduces them on HotpotQA. The two examples in Appendix B.4 show how reinforcement can recover missing evidence or displace another required source. The aggregate results therefore show a small, mixed contribution from reinforcement beyond strongest-route recall.

Table 4: **Reinforcement in Memini.** Retention F1 (%), fixed graded quality; bold: column best.

Reinforcement	MuSiQue		2Wiki		HotpotQA	
	Associative	Hybrid	Associative	Hybrid	Associative	Hybrid
On	43.61	50.22	68.37	67.23	**68.56**	**73.92**
Off	**44.24**	**50.32**	**69.07**	**67.34**	68.40	73.70

5.5 Writer and Cue-Model Choice

We examine Memini's sensitivity to writer and cue-model choice. With GPT-5.4 writing fixed, GPT-4o-mini cues yield similar observed acquisition and retention performance across both readouts and all three datasets: R@5 changes by -0.80 to $+0.50$ points and F1 by -0.60 to $+1.22$ points. With GPT-4o-mini cues fixed, changing the writer leaves hybrid answering comparatively stable, with F1 changes of at most 0.41 points; standalone associative recall is more writer-sensitive (Appendix B.5). These results support lighter cue extraction for an existing memory and show that hybrid answering is less sensitive to writer choice than standalone associative recall.

5.6 Scalability and Cost

We measure the resources required to construct, store and query Memini. Episodes are written once on arrival, at a mean API cost of $0.0060–$0.0067 each with GPT-5.4 across the three datasets. As episodes accumulate, serialized memory size grows approximately in proportion to their count.

We measure retrieval and complete-answering latency separately. On a laptop with 32 GB RAM and an RTX 5070 Ti GPU, hybrid retrieval takes 63–110 ms median at final memories of 6,119–11,656 episodes, excluding cue extraction, question encoding and answering. Including these stages, hybrid answering with GPT-5.4 cues has lower observed median latency than HippoRAG2 across all three datasets: 3.02–3.20 s versus 4.62–5.22 s.

Lighter cues reduce recurring query API cost. In a separate paired experiment, replacing GPT-5.4 cues with GPT-4o-mini reduces hybrid complete-query API cost, including answering, by 80–82%, to $0.23–$0.24 per 1,000 queries; this is approximately half HippoRAG2's measured complete-query API cost. Appendix E details the measurements and protocols. Together with the similar observed acquisition and retention performance (Section 5.5), these measurements show that lighter cues enable substantially cheaper repeated access to the same written memory.

6 Discussion

Each system received the same sources and answered from retrieved source episodes using the same answerer. Differences in retrieval and answering distinguish preserving observations from keeping them accessible and usable as memory grows. Memini's hybrid readout retrieves more supporting evidence than dense retrieval after updates and combines strong retained answering with less forgetting than HippoRAG2 across all three datasets. The comparison with HippoRAG2 is informative because both systems maintain associative structure but differ in how it is formed and recalled.

Episode-supported routes complement direct semantic access rather than replacing it. Standalone associative recall is strongest on 2Wiki, notably on questions requiring a separate evidence branch for each of two entities, while hybrid recall is more consistent across the three datasets. The allocation analysis in Appendix B.6 makes the trade-off explicit: assigning more of the five context positions to associative recall improves retained answering on 2Wiki, while the common 4+1 allocation has the lowest observed forgetting on MuSiQue and HotpotQA. These patterns suggest that how much of a limited budget should be spent on associative access depends on how the evidence a question needs is distributed across sources.

The specific concept connections formed as observations arrive help keep stored evidence usable: shuffling their destinations substantially degrades retained answering while preserving the source archive and the quality values attached to the routes. Additional reinforcement has smaller, mixed effects. Strengthening an association can recover missing evidence or displace another required source, since a meaningful connection is not always the one a question needs.

Limitations. Sequential chains remain difficult for every method we compare: on MuSiQue's four-step questions, hybrid recall retains the highest observed answering yet rarely retrieves all required supports, so its answer scores do not establish that the full chain was reconstructed (Appendix B.2). Recall follows one bounded step from cue to anchor to bridge, which suits parallel branches but limits evidence requiring several successive dependencies. Our controlled textual streams reach approximately 6,000–12,000 episodes, with storage growing as observations accumulate; evaluation and optimization for substantially longer streams remain future work. Earlier information is assumed to remain valid, leaving superseded facts outside the evaluated setting.

7 Conclusion

We introduced Memini, an external associative memory for continual learning from unlabeled textual observations with all pretrained components held fixed. Memini preserves source episodes and builds episode-supported concept connections; partial cues then activate bounded, parallel routes through these connections to recover complete source episodes. Across streams from three multi-hop QA benchmarks, it supports acquisition, retained answering and the joint use of evidence received at different times, through associative recall alone or combined with direct semantic retrieval.

AI USE STATEMENT

We used generative AI tools interactively to discover relevant literature and discuss research ideas and methods. We checked suggested ideas against the literature for originality and attribution and assessed proposed design choices before adopting them. We also used AI assistance for implementation and debugging, testing the resulting code before adoption. For manuscript preparation, we used iterative exchanges to draft and polish selected passages, which we reviewed and revised for accuracy, clarity, and consistency with the experimental evidence. We take responsibility for the final manuscript, code, and reported results.

ETHICS STATEMENT

This study uses existing public question-answering benchmarks and involves no new recruitment of human participants. Persistent retrieval systems can retain and reproduce errors, biases, or sensitive information from their sources. Deployment beyond these benchmarks requires appropriate controls over data access, retention, and deletion.

REPRODUCIBILITY STATEMENT

Section 4 and Appendix C specify memory construction and recall. Section 5 and Appendix A describe the datasets, arrival schedules, model settings, evaluation metrics, and uncertainty estimation. Appendices D and E provide the language-model prompts and resource-measurement protocols.

REFERENCES

Arslan Chaudhry, Puneet K. Dokania, Thalaiyasingam Ajanthan, and Philip H. S. Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Computer Vision – ECCV 2018*, volume 11215 of *Lecture Notes in Computer Science*, pp. 556–572, Cham, 2018. Springer. doi: 10.1007/978-3-030-01252-6_33. URL https://doi.org/10.1007/978-3-030-01252-6_33.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Gregory G. Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385, 2022. ISSN 0162-8828, 2160-9292, 1939-3539. doi: 10.1109/TPAMI.2021.3057446.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph RAG approach to query-focused summarization. arXiv preprint arXiv:2404.16130, February 2025. URL https://arxiv.org/abs/2404.16130.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. LightRAG: Simple and fast retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 10746–10761, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.568. URL https://aclanthology.org/2025.findings-emnlp.568/.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From RAG to memory: Non-parametric continual learning for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 21497–21515. PMLR, 2025.

Donald O. Hebb. *The Organization of Behavior: A Neuropsychological Theory*. John Wiley & Sons, New York, 1949.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.580.

Qisheng Hu, Quanyu Long, and Wenya Wang. When continual learning moves to memory: A study of experience reuse in LLM agents. arXiv preprint arXiv:2604.27003, April 2026. URL https://arxiv.org/abs/2604.27003.

Dharshan Kumaran, Demis Hassabis, and James L. McClelland. What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends in Cognitive Sciences*, 20(7):512–534, July 2016. doi: 10.1016/j.tics.2016.05.004.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. NV-Embed: Improved techniques for training LLMs as generalist embedding models. In *The Thirteenth International Conference on Learning Representations*, 2025. doi: 10.48550/arXiv.2405.17428.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33*, pp. 9459–9474. Curran Associates, Inc., 2020.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6467–6476. Curran Associates, Inc., 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/f87522788a2be2d171666752f97ddebb-Abstract.html.

James L. McClelland, Bruce L. McNaughton, and Randall C. O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419–457, 1995. doi: 10.1037/0033-295x.102.3.419.

German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, May 2019. ISSN 08936080. doi: 10.1016/j.neunet.2019.01.012.

Shreyas Rajesh, Pavan Holur, Mehmet Yigit Turali, Chenda Duan, and Vwani Roychowdhury. Panini: Continual learning in token space via structured memory. arXiv preprint arXiv:2602.15156, 2026. URL https://arxiv.org/abs/2602.15156.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 4(1-2):1–174, 2009. doi: 10.1561/1500000019.

Edmund T. Rolls. The mechanisms for pattern completion and pattern separation in the hippocampus. *Frontiers in Systems Neuroscience*, 7:74, October 2013. ISSN 1662-5137. doi: 10.3389/fnsys.2013.00074.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. RAPTOR: Recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations*, 2024. doi: 10.48550/arXiv.2401.18059.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. Continual learning of large language models: A comprehensive survey. *ACM Computing Surveys*, 58(5):1–42, 2025. ISSN 0360-0300, 1557-7341. doi: 10.1145/3735633.

Bernhard P. Staresina and Lila Davachi. Mind the gap: Binding experiences across space and time in the human hippocampus. *Neuron*, 63(2):267–276, July 2009. ISSN 0896-6273. doi: 10.1016/j.neuron.2009.06.024.

Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. Black-box tuning for language-model-as-a-service. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20841–20855, Baltimore, Maryland, USA, July 2022. PMLR. URL https://proceedings.mlr.press/v162/sun22e.html.

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*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475.

Tongtong Wu, Linhao Luo, Yuan-Fang Li, Shirui Pan, Thuy-Trang Vu, and Gholamreza Haffari. Continual learning for large language models: A survey. arXiv preprint arXiv:2402.01364, February 2024. URL https://arxiv.org/abs/2402.01364.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259.

APPENDICES

The appendices are organized as follows:

- **Appendix A: Experimental Protocols.** Model settings, arrival schedules, scoring and uncertainty estimation.
- **Appendix B: Additional Results and Analyses.** Supporting results, controlled interventions and illustrative examples.
- **Appendix C: Memory Formation and Recall Details.** Source checking, integration across arrivals, and associative and hybrid recall, with algorithms and a worked example.
- **Appendix D: Language-Model Prompts.** Instructions, input templates and output schemas for episode writing, cue extraction and answering.
- **Appendix E: Resource Accounting.** Storage, latency and cost measurements, with their evaluation conditions.
- **Appendix F: Comparison with Related Systems.** Memory design and reported evaluations of the approaches discussed in Related Work.

A EXPERIMENTAL PROTOCOLS

This appendix specifies the settings, arrival schedules and aggregation rules underlying Section 5.

A.1 MODEL AND IMPLEMENTATION SETTINGS

Table A1 specifies the model versions and request settings for the shared evaluation setup in Section 5.1.

Table A1: **Shared model settings.**

Component	Model	Request settings
Memini episode writer	GPT-5.4-2026-03-05	Reasoning effort `none`; temperature omitted; maximum 8,000 output tokens
Memini cue extractor	GPT-5.4-2026-03-05	Reasoning effort `none`; temperature omitted; maximum 2,000 output tokens
Shared encoder	NV-Embed-v2	Used by Memini, Dense and HippoRAG2
Shared answerer	GPT-4o-mini-2024-07-18	Temperature 0; seed 20260816; maximum 400 output tokens

We use NV-Embed-v2 revision `3fa59658547db50a1e8e3346cf057fd0c77ed6ef`, with a maximum input length of 32,768 tokens and unit-normalized embeddings. Appendix C.3 gives the encoder instructions and input formatting.

BM25 uses the `rank_bm25` defaults, $k_1 = 1.5$ and $b = 0.75$. HippoRAG2 uses the implementation released with its paper (Gutiérrez et al., 2025), with GPT-4o-mini-2024-07-18 and the shared encoder.[2] At each arrival, we pass only new passages to its indexing routine on the existing graph; previously supplied passages are not submitted again. The routine reconstructs entity and fact collections from saved extractions and appends fact, passage and synonym edges to the retained graph.

All methods supply up to five retrieved episodes, in their returned order, to the same answerer using HippoRAG2's released QA prompt and a shared answer parser. To control LLM variability across arrival orders, we reuse Memini's writer outputs and question cues, and HippoRAG2's entity and triple extractions, for the same episode or question. For all methods, we reuse answers when the

[2]Our results and implementation descriptions refer to the evaluated release; subsequent repository updates were not evaluated.

full prompt and generation settings are identical. EM compares normalized answers; token F1 measures token overlap after lowercasing, removing punctuation and English articles, and normalizing whitespace. Both take the highest score over accepted answer aliases.

A.2 Acquisition and retention

For the acquisition and retention evaluation in Section 5.2, each dataset's 1,000 questions form four cohorts of 250, denoted G_1, G_2, G_3, G_4, balanced by available hop count or question type using seed 20260713. Four cyclic orders place every cohort at each arrival position once (Table A2).

Table A2: **Cyclic cohort arrival orders.**

Order	First arrival	Second arrival	Third arrival	Fourth arrival
1	G_1	G_2	G_3	G_4
2	G_2	G_3	G_4	G_1
3	G_3	G_4	G_1	G_2
4	G_4	G_1	G_2	G_3

Each cohort introduces its supporting and distractor episodes. Shared episodes enter once, at their earliest occurrence in the arrival order. Each question enters evaluation with its cohort and is retested after subsequent arrivals.

Acquisition averages each question's first scheduled score across 1,000 questions and four orders, giving 4,000 scores per dataset. Retention averages only the final-checkpoint scores for questions entering in the first three arrivals, giving 3,000 scores. Questions entering last contribute only to acquisition because no subsequent update follows them.

Let $s_{q,o,t}$ denote the score of question q in order o at checkpoint t, on a 0–100 scale. Let $e_{q,o}$ be its first evaluation checkpoint and $T = 4$ the final checkpoint. For each question/order combination included in retention, backward transfer measures the change from first to final evaluation (Lopez-Paz & Ranzato, 2017):

$$\mathrm{BWT}_{q,o} = s_{q,o,T} - s_{q,o,e_{q,o}}.$$

Forgetting measures the decline from its highest observed score to its final score (Chaudhry et al., 2018):

$$\mathrm{Forgetting}_{q,o} = \max_{e_{q,o} \le t \le T} s_{q,o,t} - s_{q,o,T}.$$

Both measures are calculated per question/order combination before averaging over those included in retention. BWT therefore uses matched first and final scores, whereas acquisition also includes questions entering last. These aggregation rules apply to both retrieval and answering metrics.

A.3 Cross-arrival recombination

For the cross-arrival evaluation in Section 5.3, supporting-evidence annotations determine the episode pairs and their arrival order. MuSiQue uses dependencies between annotated reasoning steps, assigning their supporting episodes to successive dependency stages. Following annotated support order, 2Wiki pairs the first and second supports for compositional and inference questions, and the first with the third and the second with the fourth for bridge comparisons. HotpotQA bridge questions pair the first and second supports. Ordinary comparison questions combine independent facts and are excluded from this dependency-based evaluation.

Shared episodes enter once, at their earliest assigned stage. All pure distractors, episodes supporting none of the evaluation questions, are introduced initially, so later supporting evidence enters an already populated memory with competing passages. A pair is eligible only when its designated earlier episode arrives before its later episode. Each eligible pair contributes one evaluation event at the later arrival, yielding 1,579 events from 999 questions on MuSiQue, 980 from 747 on 2Wiki, and 810 from 810 on HotpotQA.

Joint@5 measures whether both designated supports occur within the top five retrieved episodes. EM and F1 measure exact and partial answer correctness for the original question using context

retrieved from the available prefix at that checkpoint, regardless of pair recovery. On MuSiQue, other required supports may not yet have arrived when the pair is evaluated. Scores are averaged over events, so each eligible pair contributes one retrieval score and the original question's EM and F1 at that checkpoint.

A.4 Paired confidence intervals

We compute 95% confidence intervals for differences between methods using 10,000 paired bootstrap resamples of questions drawn with replacement. For BWT and forgetting, each method's score is first calculated from its own question/order trajectory, then the paired difference between methods is formed. For acquisition, retention, BWT and forgetting, paired differences are averaged across each question's eligible arrival orders before resampling questions. For cross-arrival evaluation, questions are resampled with all their events and differences are averaged over events. Interval endpoints are the 2.5th and 97.5th percentiles of the bootstrap estimates. Intervals are unadjusted for multiple comparisons and conditional on the evaluated histories and model outputs.

B Additional Results and Analyses

This appendix provides additional evidence for Memini's continual retrieval and answering capabilities. Two worked examples illustrate retained access and joint use of evidence across arrivals, followed by detailed quantitative comparisons. We then examine the contributions of associative organization, sensitivity to the episode writer and cue extractor, and hybrid allocation.

B.1 Illustrative retrieval and answering examples

We present two examples taken from the evaluated runs in Sections 5.2 and 5.3, using their recorded retrieval results and generated answers. The first follows the same question through acquisition and subsequent updates; the second examines a question whose supporting evidence arrives at different times. Together, they connect evidence availability, retrieval and answering in the continual setting.

Example 1 (Figure B1) examines retained access after updates. The 2Wiki question asks which of two films had the director who died later. The film names appear in both the question and episodes A and C, which identify the directors; biographies B and D supply their death dates, completing two evidence branches. All four episodes are available at acquisition, checkpoint 2, and the question is revisited at checkpoints 3 and 4. The figure shows these passages, F1 across checkpoints, and final retrieval, parsed answers and EM/F1 scores.

At final retention, the three baselines retrieve the film episodes but omit the biographies needed for comparison. Memini's associative readout recovers both complete evidence branches: the recurring directors connect each film episode to its biography. Complete recovery and exact answering persist through both updates within the five-episode budget.

Hybrid retrieves biography D but misses B, yet its parsed answer matches the reference. This example shows why we evaluate retrieval and answering together. We seek both recovery of the required evidence and correct answering from that evidence; EM/F1 alone cannot establish whether both occurred.

Question: Which film has the director who died later, 45 Calibre Echo or Bons Baisers De Hong Kong?
Reference answer: Bons Baisers De Hong Kong.

Supporting evidence

ID	Episode title	Relevant source excerpt
A	45 Calibre Echo	"a 1932 American western film directed by **Bruce M. Mitchell**"
B	Bruce M. Mitchell	Death date: "September 26, **1952**"
C	Bons Baisers de Hong Kong	"a 1975 French film directed by **Yvan Chiffre**."
D	Yvan Chiffre	Death date: "27 September **2016**"

Answer F1 at acquisition and after subsequent updates

Method	Acquisition Checkpoint 2	Retest Checkpoint 3	Final retention Checkpoint 4
BM25	0	0	0
Dense	0	0	0
HippoRAG2	66.67	100	0
Memini (associative)	100	100	100
Memini (hybrid)	100	100	100

Retrieved context and answer at final retention

	Retrieval rank							Scores (%)	
Method	1	2	3	4	5	Supporting episodes	Parsed answer	EM	F1
BM25	**C**	**A**	·	·	·	2/4	Cannot determine.	0	0
Dense	**A**	**C**	·	·	·	2/4	Cannot determine.	0	0
HippoRAG2	**A**	**C**	·	·	·	2/4	Cannot determine.	0	0
Memini (associative)	**C**	**A**	**D**	**B**	·	4/4	Bons Baisers de Hong Kong.	100	100
Memini (hybrid)	**A**	**C**	·	·	**D**	3/4	Bons Baisers de Hong Kong.	100	100

Columns 1–5 show retrieval order. Letters identify the supporting episodes above; · denotes another retrieved episode. Supporting episodes are reported as retrieved/required within the top-five context; the corresponding percentage is R@5.

Figure B1: **Example 1: Retained access on 2Wiki.** Supporting passages, answer F1 across three checkpoints, and final top-five contexts and parsed answers for the first of the four evaluated arrival orders. Associative recall recovers both evidence branches; hybrid matches the reference answer despite missing one required biography. EM and F1 are reported as percentages.

Example 2 (Figure B2) examines a complementary requirement: bringing together evidence received at different times. The MuSiQue question asks for the child of the *Victim of Romance* performer. Episode A arrives at stage 0 and identifies Michelle Phillips as the performer; episode B arrives at stage 1 and identifies her daughter, Chynna Phillips. Evaluation occurs at stage 1, when both episodes are available to every method. The figure shows their evidence and arrival stages, followed by retrieval, parsed answers, Joint@5 and EM/F1.

Question: Who is the child of the Victim of Romance performer?
Reference answer: Chynna Phillips.

Supporting evidence and arrival stages

ID	Stage	Episode title	Relevant source excerpt
A	0	Victim of Romance	"Victim of Romance is singer and songwriter **Michelle Phillips**' first and only solo album"
B	1	Wilson Phillips	"Wilson Phillips is an American vocal group consisting of Carnie Wilson, Wendy Wilson, and **Chynna Phillips**, the daughters, respectively, of Brian Wilson of The Beach Boys and of John and **Michelle Phillips** of The Mamas & the Papas."

Retrieved context and answer at stage 1

	Retrieval rank						Scores (%)		
Method	1	2	3	4	5	Parsed answer	Joint@5	EM	F1
BM25	**A**	·	·	·	·	Not specified in the text.	0	0	0
Dense	**A**	·	·	·	·	Not specified.	0	0	0
HippoRAG2	**A**	·	·	·	·	Not specified.	0	0	0
Memini (associative)	**A**	**B**				Chynna Phillips.	100	100	100
Memini (hybrid)	**A**	·	·	·	**B**	Chynna Phillips.	100	100	100

Columns 1 to 5 show retrieval order. Letters identify the supporting episodes above; · indicates another retrieved episode. Associative recall returns only two episodes in this example, so positions 3 to 5 are empty. Joint@5 is 100 when both supporting episodes occur in the top-five context, and 0 otherwise.

The later supporting episode B ranks 143 for Dense and 8 for HippoRAG2, and is absent from BM25's saved top 200. Memini places it second with associative recall and fifth with hybrid recall. Hybrid's first four retrieved episodes are identical to Dense's first four.

Figure B2: **Example 2: Joint use across arrivals on MuSiQue.** Episodes A and B arrive at stages 0 and 1; retrieval and answering are evaluated when B arrives. Both Memini readouts recover the pair and answer correctly; hybrid adds the later episode to Dense's first four selections. Parsed answers are shown, with Joint@5, EM and F1 reported as percentages.

All methods retrieve the earlier album episode, but the three baselines omit the later *Wilson Phillips* episode from their top-five contexts. Michelle Phillips connects the album to the later family information. Associative recall retrieves both episodes in its first two positions; hybrid preserves Dense's first four selections and adds the missing episode fifth. Both answer "Chynna Phillips" exactly. Associative access thus supplies evidence outside Dense's leading results, making information received at different times usable together within the same context budget.

B.2 Acquisition and retention

This section extends the acquisition and retention results in Section 5.2. We first report evidence recovery and answering at both stages, with paired uncertainty for the comparisons between Memini and the baselines. We then report BWT and forgetting to show how retrieval and answering change after acquisition. Finally, we examine two question types: parallel evidence branches on 2Wiki, where associative recovery is particularly strong, and sequential four-step chains on MuSiQue, where complete recovery remains difficult.

Evidence recovery and answering. Tables B1–B4 extend main Table 2 with paired differences against each baseline for R@5, C@5, EM and F1. Acquisition covers all four arrival orders; retention covers the three orders in which each question is followed by further updates. Scores are percentages; differences are percentage points, calculated before rounding. Positive performance differences favor Memini. Brackets give paired 95% confidence intervals following Appendix A.4. Tables B1 and B2 distinguish recovery of individual supports from recovery of the complete evidence set.

Table B1: **Supporting-episode recall (R@5) at acquisition and retention.** Scores (%) are shown as baseline → Memini. Δ is the Memini score minus the baseline score, in percentage points. Brackets give the paired 95% confidence interval for this difference. Positive differences are bold, highlighting where Memini scores higher.

	Acquisition			Retention		
Baseline	Scores (%)	Δ	95% CI	Scores (%)	Δ	95% CI
MuSiQue: Memini (associative)						
BM25	41.48 → 61.24	**+19.76**	[17.58, 22.05]	40.60 → 59.69	**+19.09**	[16.75, 21.43]
Dense	71.26 → 61.24	−10.02	[−12.16, −7.85]	69.38 → 59.69	−9.69	[−11.92, −7.46]
HippoRAG2	76.26 → 61.24	−15.02	[−17.04, −12.97]	73.04 → 59.69	−13.35	[−15.49, −11.19]
MuSiQue: Memini (hybrid)						
BM25	41.48 → 76.46	**+34.99**	[33.19, 36.75]	40.60 → 74.58	**+33.98**	[32.14, 35.86]
Dense	71.26 → 76.46	**+5.20**	[4.15, 6.29]	69.38 → 74.58	**+5.20**	[4.08, 6.38]
HippoRAG2	76.26 → 76.46	**+0.20**	[−0.91, 1.30]	73.04 → 74.58	**+1.54**	[0.37, 2.65]
2Wiki: Memini (associative)						
BM25	64.48 → 91.46	**+26.98**	[25.14, 28.80]	63.93 → 90.85	**+26.93**	[24.98, 28.85]
Dense	77.70 → 91.46	**+13.76**	[12.06, 15.45]	76.58 → 90.85	**+14.28**	[12.53, 16.03]
HippoRAG2	91.48 → 91.46	−0.03	[−1.44, 1.38]	84.81 → 90.85	**+6.04**	[4.48, 7.54]
2Wiki: Memini (hybrid)						
BM25	64.48 → 89.98	**+25.49**	[23.99, 27.01]	63.93 → 88.80	**+24.88**	[23.33, 26.45]
Dense	77.70 → 89.98	**+12.28**	[11.12, 13.44]	76.58 → 88.80	**+12.23**	[11.00, 13.45]
HippoRAG2	91.48 → 89.98	−1.51	[−2.49, −0.54]	84.81 → 88.80	**+3.99**	[2.93, 5.03]
HotpotQA: Memini (associative)						
BM25	71.44 → 86.48	**+15.04**	[12.94, 17.19]	70.90 → 85.55	**+14.65**	[12.45, 16.90]
Dense	95.00 → 86.48	−8.53	[−10.23, −6.86]	94.40 → 85.55	−8.85	[−10.65, −7.05]
HippoRAG2	96.34 → 86.48	−9.86	[−11.50, −8.29]	96.07 → 85.55	−10.52	[−12.27, −8.87]
HotpotQA: Memini (hybrid)						
BM25	71.44 → 96.59	**+25.15**	[23.41, 26.91]	70.90 → 96.15	**+25.25**	[23.40, 27.10]
Dense	95.00 → 96.59	**+1.59**	[0.90, 2.31]	94.40 → 96.15	**+1.75**	[1.00, 2.50]
HippoRAG2	96.34 → 96.59	**+0.25**	[−0.44, 0.91]	96.07 → 96.15	**+0.08**	[−0.63, 0.78]

Table B2: **Complete support recovery (C@5) at acquisition and retention.** Scores (%) are shown as baseline → Memini. Δ is the Memini score minus the baseline score, in percentage points. Brackets give the paired 95% confidence interval for this difference. Positive differences are bold, highlighting where Memini scores higher.

	Acquisition			Retention		
Baseline	Scores (%)	Δ	95% CI	Scores (%)	Δ	95% CI
MuSiQue: Memini (associative)						
BM25	10.18 → 35.73	**+25.55**	[22.60, 28.45]	9.90 → 33.90	**+24.00**	[20.90, 27.10]
Dense	40.08 → 35.73	−4.35	[−7.53, −1.25]	37.50 → 33.90	−3.60	[−6.90, −0.40]
HippoRAG2	49.45 → 35.73	−13.73	[−16.53, −10.90]	44.23 → 33.90	−10.33	[−13.30, −7.33]
MuSiQue: Memini (hybrid)						
BM25	10.18 → 49.35	**+39.18**	[36.18, 42.20]	9.90 → 46.20	**+36.30**	[33.30, 39.40]
Dense	40.08 → 49.35	**+9.28**	[7.15, 11.50]	37.50 → 46.20	**+8.70**	[6.40, 11.00]
HippoRAG2	49.45 → 49.35	−0.10	[−2.38, 2.08]	44.23 → 46.20	**+1.97**	[−0.27, 4.23]
2Wiki: Memini (associative)						
BM25	31.88 → 81.58	**+49.70**	[46.58, 52.83]	31.20 → 80.30	**+49.10**	[45.90, 52.30]
Dense	52.43 → 81.58	**+29.15**	[25.98, 32.35]	50.50 → 80.30	**+29.80**	[26.50, 33.20]
HippoRAG2	76.95 → 81.58	**+4.63**	[1.95, 7.35]	63.70 → 80.30	**+16.60**	[13.53, 19.67]
2Wiki: Memini (hybrid)						
BM25	31.88 → 71.20	**+39.33**	[36.50, 42.25]	31.20 → 69.00	**+37.80**	[34.70, 40.90]
Dense	52.43 → 71.20	**+18.78**	[16.40, 21.15]	50.50 → 69.00	**+18.50**	[16.00, 21.00]
HippoRAG2	76.95 → 71.20	−5.75	[−8.05, −3.53]	63.70 → 69.00	**+5.30**	[3.17, 7.37]

Table B2 (continued): Complete support recovery (C@5)

	Acquisition			Retention		
Baseline	Scores (%)	Δ	95% CI	Scores (%)	Δ	95% CI
HotpotQA: Memini (associative)						
BM25	47.50 → 77.48	**+29.98**	[26.48, 33.50]	46.60 → 75.80	**+29.20**	[25.40, 32.90]
Dense	90.23 → 77.48	−12.75	[−15.53, −10.00]	89.10 → 75.80	−13.30	[−16.20, −10.40]
HippoRAG2	92.95 → 77.48	−15.48	[−18.00, −12.95]	92.53 → 75.80	−16.73	[−19.47, −14.00]
HotpotQA: Memini (hybrid)						
BM25	47.50 → 93.48	**+45.98**	[42.85, 49.18]	46.60 → 92.60	**+46.00**	[42.80, 49.20]
Dense	90.23 → 93.48	**+3.25**	[1.90, 4.63]	89.10 → 92.60	**+3.50**	[2.00, 5.00]
HippoRAG2	92.95 → 93.48	**+0.53**	[−0.80, 1.83]	92.53 → 92.60	**+0.07**	[−1.33, 1.43]

Hybrid recall retains more individual supports and complete evidence sets than Dense on all three datasets. On 2Wiki, associative recall is stronger still: its retained R@5 of 90.85 exceeds every baseline, and complete recovery reaches 80.30% versus HippoRAG2's 63.70%, with a paired interval supporting the 16.60-point difference. These results show two useful forms of access: recovering supporting evidence through learned associations alone and supplying additional evidence alongside dense selections within the same five-episode budget. On HotpotQA, hybrid and HippoRAG2 both recover complete evidence in about 93% of cases, and their small difference remains uncertain.

Tables B3 and B4 examine whether the retrieved evidence supports correct answering under the shared answerer. EM measures exact match, while F1 gives partial credit for token overlap with the reference answer.

Table B3: **Answer exact match (EM) at acquisition and retention.** Scores (%) are shown as baseline → Memini. Δ is the Memini score minus the baseline score, in percentage points. Brackets give the paired 95% confidence interval for this difference. Positive differences are bold, highlighting where Memini scores higher.

	Acquisition			Retention		
Baseline	Scores (%)	Δ	95% CI	Scores (%)	Δ	95% CI
MuSiQue: Memini (associative)						
BM25	19.83 → 31.93	**+12.10**	[9.60, 14.65]	19.20 → 31.30	**+12.10**	[9.40, 14.90]
Dense	33.83 → 31.93	−1.90	[−4.35, 0.55]	32.50 → 31.30	−1.20	[−4.00, 1.60]
HippoRAG2	35.75 → 31.93	−3.83	[−6.18, −1.53]	33.57 → 31.30	−2.27	[−4.93, 0.40]
MuSiQue: Memini (hybrid)						
BM25	19.83 → 37.63	**+17.80**	[15.25, 20.33]	19.20 → 36.90	**+17.70**	[15.00, 20.50]
Dense	33.83 → 37.63	**+3.80**	[2.23, 5.43]	32.50 → 36.90	**+4.40**	[2.60, 6.30]
HippoRAG2	35.75 → 37.63	**+1.88**	[0.15, 3.58]	33.57 → 36.90	**+3.33**	[1.33, 5.33]
2Wiki: Memini (associative)						
BM25	42.60 → 60.40	**+17.80**	[15.18, 20.43]	42.40 → 59.70	**+17.30**	[14.40, 20.20]
Dense	53.33 → 60.40	**+7.08**	[4.83, 9.38]	52.90 → 59.70	**+6.80**	[4.20, 9.50]
HippoRAG2	60.63 → 60.40	−0.23	[−2.05, 1.63]	56.33 → 59.70	**+3.37**	[0.97, 5.77]
2Wiki: Memini (hybrid)						
BM25	42.60 → 58.55	**+15.95**	[13.47, 18.40]	42.40 → 57.80	**+15.40**	[12.70, 18.10]
Dense	53.33 → 58.55	**+5.23**	[3.50, 6.98]	52.90 → 57.80	**+4.90**	[2.80, 7.00]
HippoRAG2	60.63 → 58.55	−2.08	[−3.73, −0.43]	56.33 → 57.80	**+1.47**	[−0.47, 3.40]
HotpotQA: Memini (associative)						
BM25	45.48 → 54.25	**+8.78**	[6.43, 11.15]	45.50 → 53.90	**+8.40**	[5.80, 11.00]
Dense	58.63 → 54.25	−4.38	[−6.43, −2.35]	58.70 → 53.90	−4.80	[−7.10, −2.60]
HippoRAG2	58.93 → 54.25	−4.68	[−6.58, −2.83]	57.90 → 53.90	−4.00	[−6.17, −1.93]
HotpotQA: Memini (hybrid)						
BM25	45.48 → 58.88	**+13.40**	[11.18, 15.70]	45.50 → 59.40	**+13.90**	[11.60, 16.40]
Dense	58.63 → 58.88	**+0.25**	[−0.95, 1.48]	58.70 → 59.40	**+0.70**	[−0.60, 2.10]
HippoRAG2	58.93 → 58.88	−0.05	[−1.43, 1.35]	57.90 → 59.40	**+1.50**	[0.13, 2.93]

Table B4: **Answer F1 at acquisition and retention.** Scores (%) are shown as baseline → Memini. Δ is the Memini score minus the baseline score, in percentage points. Brackets give the paired 95% confidence interval for this difference. Positive differences are bold, highlighting where Memini scores higher.

	Acquisition			Retention		
Baseline	Scores (%)	Δ	95% CI	Scores (%)	Δ	95% CI
MuSiQue: Memini (associative)						
BM25	30.10 → 44.54	**+14.45**	[11.97, 16.94]	29.34 → 43.61	**+14.27**	[11.53, 17.08]
Dense	46.86 → 44.54	−2.32	[−4.75, 0.11]	45.19 → 43.61	−1.57	[−4.34, 1.18]
HippoRAG2	49.92 → 44.54	−5.38	[−7.63, −3.09]	47.50 → 43.61	−3.88	[−6.47, −1.31]
MuSiQue: Memini (hybrid)						
BM25	30.10 → 51.16	**+21.06**	[18.62, 23.56]	29.34 → 50.22	**+20.87**	[18.23, 23.59]
Dense	46.86 → 51.16	**+4.29**	[2.72, 5.88]	45.19 → 50.22	**+5.03**	[3.19, 6.87]
HippoRAG2	49.92 → 51.16	**+1.23**	[−0.46, 2.90]	47.50 → 50.22	**+2.72**	[0.73, 4.68]
2Wiki: Memini (associative)						
BM25	48.26 → 69.32	**+21.06**	[18.44, 23.70]	48.14 → 68.37	**+20.23**	[17.38, 23.09]
Dense	60.38 → 69.32	**+8.94**	[6.63, 11.27]	59.82 → 68.37	**+8.54**	[6.00, 11.04]
HippoRAG2	70.66 → 69.32	−1.34	[−3.21, 0.51]	65.04 → 68.37	**+3.33**	[0.98, 5.61]
2Wiki: Memini (hybrid)						
BM25	48.26 → 68.10	**+19.84**	[17.41, 22.31]	48.14 → 67.23	**+19.09**	[16.40, 21.82]
Dense	60.38 → 68.10	**+7.72**	[5.89, 9.57]	59.82 → 67.23	**+7.41**	[5.32, 9.54]
HippoRAG2	70.66 → 68.10	−2.56	[−4.22, −1.00]	65.04 → 67.23	**+2.19**	[0.26, 4.11]
HotpotQA: Memini (associative)						
BM25	58.58 → 69.02	**+10.43**	[8.22, 12.65]	58.57 → 68.56	**+9.98**	[7.59, 12.44]
Dense	73.12 → 69.02	−4.10	[−5.88, −2.35]	73.01 → 68.56	−4.45	[−6.39, −2.52]
HippoRAG2	73.97 → 69.02	−4.96	[−6.64, −3.26]	72.97 → 68.56	−4.41	[−6.24, −2.64]
HotpotQA: Memini (hybrid)						
BM25	58.58 → 73.66	**+15.08**	[13.01, 17.18]	58.57 → 73.92	**+15.35**	[13.21, 17.58]
Dense	73.12 → 73.66	**+0.54**	[−0.49, 1.60]	73.01 → 73.92	**+0.91**	[−0.26, 2.09]
HippoRAG2	73.97 → 73.66	−0.31	[−1.45, 0.86]	72.97 → 73.92	**+0.95**	[−0.27, 2.15]

Hybrid recall has the highest observed acquisition F1 on MuSiQue and retains higher EM and F1 than Dense and HippoRAG2, with paired intervals supporting the retained gains. Its acquisition F1 difference from HippoRAG2 remains uncertain. On 2Wiki, both readouts retain higher F1 than every baseline; associative recall also retains higher EM. On HotpotQA, hybrid improves retained EM over HippoRAG2 by 1.50 points, while its 0.95-point F1 difference remains uncertain. Standalone associative recall retains lower answering scores than Dense and HippoRAG2 on MuSiQue and HotpotQA. Together, retrieval and answering identify useful retained access: the gains concern both evidence entering the context and the answers generated from it.

Changes after acquisition. Tables B5–B7 extend Figure 4 with retrieval and answering BWT, forgetting, and paired F1-forgetting comparisons. BWT measures final minus entry performance; forgetting measures the loss from the best score along each question/order trajectory with subsequent updates, following Appendix A.2.

Small losses must be interpreted alongside retained performance. On MuSiQue, BM25 has slightly less F1 forgetting than associative recall (4.39 versus 4.48 points), but much lower retained F1 (29.34 versus 43.61). The relevant combination is strong retained performance with limited loss after acquisition.

Table B5: **Retrieval stability across updates.** Backward transfer (BWT) and forgetting in percentage points, averaged over matched question/order histories with subsequent updates. Positive BWT indicates improvement after acquisition; negative BWT indicates decline. Lower forgetting is better.

	R@5		C@5	
Method	BWT ↑	Forgetting ↓	BWT ↑	Forgetting ↓
MuSiQue				
BM25	-1.17	1.75	-0.37	0.97
Dense	-2.51	2.51	-3.43	3.43
HippoRAG2	-4.11	4.88	-6.60	7.77
Memini (associative)	-2.07	2.75	-2.43	2.90
Memini (hybrid)	-2.51	2.69	-4.20	4.43
2Wiki				
BM25	-0.74	1.11	-0.90	1.53
Dense	-1.50	1.50	-2.57	2.57
HippoRAG2	-7.43	7.67	-14.87	15.30
Memini (associative)	-0.81	0.98	-1.70	2.03
Memini (hybrid)	-1.57	1.67	-2.93	3.10
HotpotQA				
BM25	-0.72	1.35	-1.20	2.30
Dense	-0.80	0.80	-1.50	1.50
HippoRAG2	-0.35	1.18	-0.53	2.13
Memini (associative)	-1.23	1.47	-2.23	2.60
Memini (hybrid)	-0.58	0.58	-1.17	1.17

Table B6: **Answering stability across updates.** Backward transfer (BWT) and forgetting in percentage points, averaged over matched question/order histories with subsequent updates. Positive BWT indicates improvement after acquisition; negative BWT indicates decline. Lower forgetting is better.

	EM		F1	
Method	BWT ↑	Forgetting ↓	BWT ↑	Forgetting ↓
MuSiQue				
BM25	-0.83	3.50	-1.01	4.39
Dense	-1.77	3.83	-2.24	4.53
HippoRAG2	-2.77	6.53	-3.02	7.35
Memini (associative)	-0.83	4.03	-1.24	4.48
Memini (hybrid)	-0.97	3.20	-1.25	3.65
2Wiki				
BM25	-0.27	3.13	-0.16	3.36
Dense	-0.57	3.47	-0.74	4.10
HippoRAG2	-4.80	8.20	-6.21	9.33
Memini (associative)	-0.93	3.00	-1.28	2.96
Memini (hybrid)	-1.00	3.40	-1.16	3.63
HotpotQA				
BM25	0.03	1.53	-0.01	1.80
Dense	0.10	1.03	-0.14	1.04
HippoRAG2	-1.20	3.57	-1.12	3.14
Memini (associative)	-0.47	2.23	-0.61	2.20
Memini (hybrid)	0.70	1.03	0.34	0.89

Table B7: **Paired differences in answer-F1 forgetting.** Δ is Memini's forgetting minus the baseline's forgetting, in F1 points. Brackets give paired 95% confidence intervals. Negative differences are bold, highlighting less forgetting for Memini. Absolute forgetting values appear in Table B6.

	F1 forgetting difference (points) ↓					
	BM25		Dense		HippoRAG2	
Memini readout	Δ	95% CI	Δ	95% CI	Δ	95% CI
MuSiQue						
Associative	$+0.09$	$[-1.17, 1.34]$	$\mathbf{-0.05}$	$[-1.32, 1.24]$	$\mathbf{-2.87}$	$[-4.26, -1.50]$
Hybrid	$\mathbf{-0.74}$	$[-1.93, 0.43]$	$\mathbf{-0.88}$	$[-1.86, 0.05]$	$\mathbf{-3.70}$	$[-4.94, -2.47]$
2Wiki						
Associative	$\mathbf{-0.39}$	$[-1.50, 0.71]$	$\mathbf{-1.14}$	$[-2.36, 0.09]$	$\mathbf{-6.37}$	$[-7.99, -4.78]$
Hybrid	$+0.27$	$[-0.88, 1.45]$	$\mathbf{-0.48}$	$[-1.68, 0.68]$	$\mathbf{-5.71}$	$[-7.19, -4.21]$
HotpotQA						
Associative	$+0.40$	$[-0.56, 1.33]$	$+1.16$	$[0.40, 1.94]$	$\mathbf{-0.94}$	$[-1.90, 0.01]$
Hybrid	$\mathbf{-0.90}$	$[-1.64, -0.21]$	$\mathbf{-0.15}$	$[-0.56, 0.27]$	$\mathbf{-2.25}$	$[-3.05, -1.48]$

On 2Wiki, associative recall combines the highest retained evidence recovery with smaller retrieval losses than Dense and HippoRAG2. Hybrid retains more evidence than Dense across datasets, although its retrieval forgetting is slightly higher on MuSiQue and 2Wiki. For answering, hybrid combines strong retained performance with lower F1 forgetting than HippoRAG2 on all three datasets, with paired intervals supporting each reduction. It also shows positive EM and F1 backward transfer on HotpotQA. Associative recall's forgetting reductions relative to HippoRAG2 are supported on MuSiQue and 2Wiki but remain uncertain on HotpotQA. Hybrid's smaller observed reductions relative to Dense also remain uncertain. The strongest combination of retained answering and stability comes from associative recall on 2Wiki and hybrid recall across datasets.

Parallel evidence branches. Bridge-comparison questions require following a separate evidence branch for each of two entities and comparing the resulting facts. In Example 1, each film is linked to its director, whose biography supplies the death date needed for the comparison. Table B8 examines this structure across the 235 four-support 2Wiki bridge-comparison questions, reporting evidence recovery and answering at acquisition and retention.

Table B8: **Two-branch recovery on 2Wiki.** Acquisition and retention scores (%) for 235 questions (940 acquisition and 705 retention question/order observations). Bold and underlining mark the best and second-best value in each column.

	Acquisition				Retention			
Method	R@5	C@5	EM	F1	R@5	C@5	EM	F1
BM25	48.91	1.06	66.06	66.83	48.19	0.85	67.23	67.94
Dense	57.23	1.81	71.91	72.63	56.28	1.28	72.34	73.36
HippoRAG2	<u>81.62</u>	<u>37.66</u>	<u>82.77</u>	<u>84.00</u>	67.06	8.94	73.19	74.02
Memini (associative)	**88.38**	**62.45**	**84.04**	**84.47**	**87.02**	**59.15**	**80.43**	**80.63**
Memini (hybrid)	76.54	15.85	75.64	76.73	<u>75.11</u>	<u>13.62</u>	<u>73.62</u>	<u>74.65</u>

Associative recall has the highest observed scores across all four metrics at acquisition and retention. It retains all four supports in 59.15% of question/order cases versus HippoRAG2's 8.94%, and reaches 80.63 retained F1. Its 19.96-point retained R@5 advantage over HippoRAG2 supports the subgroup finding in Section 5.2. Hybrid also retains more supporting evidence than every baseline, although its complete recovery is lower than associative recall's. These results extend Example 1 beyond a single question, identifying strong retained access to both evidence branches.

Sequential evidence chains. Sequential-chain questions require each intermediate answer to supply the entity or information needed by the next step. The 108 MuSiQue `4hop1` questions form four-step chains, $q_1 \rightarrow q_2 \rightarrow q_3 \rightarrow q_4$. Table B9 examines their evidence recovery and answering under the same continual acquisition and retention protocol.

Table B9: **Sequential-chain recovery on MuSiQue.** Acquisition and retention scores (%) for 108 questions (432 acquisition and 324 retention question/order observations). Bold and underlining mark the best and second-best value in each column.

	Acquisition				Retention			
Method	R@5	C@5	EM	F1	R@5	C@5	EM	F1
BM25	22.05	0.00	22.92	32.65	21.30	0.00	22.22	31.89
Dense	46.35	1.62	<u>29.63</u>	<u>42.55</u>	44.44	0.93	<u>26.85</u>	<u>39.71</u>
HippoRAG2	<u>47.86</u>	<u>3.24</u>	28.94	40.16	<u>44.68</u>	<u>1.23</u>	26.23	37.01
Memini (associative)	33.80	0.46	18.75	28.98	33.10	0.93	20.37	30.29
Memini (hybrid)	**50.81**	**3.47**	**31.48**	**43.45**	**48.38**	**2.78**	**31.48**	**43.48**

Complete recovery remains difficult for every method. Hybrid achieves the highest observed scores across all four metrics at acquisition and retention, reaching 43.48 retained F1 but recovering all required supports in only 2.78% of retention cases. Its higher answer score therefore does not establish that the full chain was reconstructed.

Standalone associative recall has lower R@5 and answering scores than Dense and HippoRAG2 on these questions. Successive dependencies can challenge bounded associative traversal; hybrid combines dense selections with associative evidence and improves both recovery and answering over Dense and standalone associative recall. Together, the two breakdowns identify strong associative access to parallel evidence branches and the complementary value of hybrid access on sequential chains, where complete recovery remains difficult.

B.3 Cross-arrival recombination

This section extends Section 5.3 by examining whether evidence received at different times is recovered and used together. We first compare joint recovery of the designated earlier and later supporting episodes, then answering of the original question. Tables B10–B12 give the scores from main Table 3, with paired differences and 95% confidence intervals for its baseline comparisons (Appendix A.4). Differences are calculated before rounding.

Each event is evaluated when its later episode arrives, with the same source availability across methods. Both designated episodes are then available; failure to recover the pair means that at least one available episode was missed. Joint@5 measures whether both enter the top-five context. EM and F1 score the original question using that context; on MuSiQue, other required supports may not yet have arrived. Joint recovery therefore concerns the designated pair rather than the full evidence set. All three metrics are averaged over the same events, following Appendix A.3.

Table B10: **Joint recovery across arrivals (Joint@5).** Scores (%) are shown as baseline → Memini. Δ is the Memini score minus the baseline score, in percentage points. Brackets give the paired 95% confidence interval for this difference. Positive differences are bold, highlighting where Memini scores higher.

Readout	Baseline	Scores (%)	Δ	95% CI
MuSiQue				
Associative	BM25	9.37 → 35.53	**+26.16**	[23.53, 28.80]
	Dense	36.73 → 35.53	−1.20	[−4.28, 1.81]
	HippoRAG2	44.46 → 35.53	−8.93	[−11.85, −6.08]
Hybrid	BM25	9.37 → 45.54	**+36.16**	[33.46, 38.90]
	Dense	36.73 → 45.54	**+8.80**	[6.84, 10.82]
	HippoRAG2	44.46 → 45.54	**+1.08**	[−1.14, 3.30]

Table B10 (continued): Joint recovery across arrivals (Joint@5)

Readout	Baseline	Scores (%)	Δ	95% CI
2Wiki				
Associative	BM25	13.27 → 80.31	**+67.04**	[63.98, 70.01]
	Dense	32.04 → 80.31	**+48.27**	[44.69, 51.78]
	HippoRAG2	66.12 → 80.31	**+14.18**	[10.87, 17.44]
Hybrid	BM25	13.27 → 65.92	**+52.65**	[49.90, 55.37]
	Dense	32.04 → 65.92	**+33.88**	[31.24, 36.45]
	HippoRAG2	66.12 → 65.92	−0.20	[−3.07, 2.65]
HotpotQA				
Associative	BM25	42.96 → 74.20	**+31.23**	[27.16, 35.31]
	Dense	86.54 → 74.20	−12.35	[−15.80, −8.89]
	HippoRAG2	91.85 → 74.20	−17.65	[−20.74, −14.57]
Hybrid	BM25	42.96 → 90.86	**+47.90**	[44.44, 51.48]
	Dense	86.54 → 90.86	**+4.32**	[2.47, 6.30]
	HippoRAG2	91.85 → 90.86	−0.99	[−2.72, 0.74]

These results highlight the distinction between evidence being available and being recovered together within the five-episode context. Memini's learned associations support this joint access, both through associative recall and alongside dense retrieval.

Tables B11 and B12 examine whether the resulting contexts also support correct answers to the original questions.

Table B11: **Exact answering across arrivals (EM).** Scores (%) are shown as baseline → Memini. Δ is the Memini score minus the baseline score, in percentage points. Brackets give the paired 95% confidence interval for this difference. Positive differences are bold, highlighting where Memini scores higher.

Readout	Baseline	Scores (%)	Δ	95% CI
MuSiQue				
Associative	BM25	15.64 → 25.21	**+9.56**	[7.07, 12.07]
	Dense	25.08 → 25.21	**+0.13**	[−2.38, 2.64]
	HippoRAG2	27.42 → 25.21	−2.22	[−4.65, 0.19]
Hybrid	BM25	15.64 → 28.12	**+12.48**	[10.22, 14.85]
	Dense	25.08 → 28.12	**+3.04**	[1.40, 4.74]
	HippoRAG2	27.42 → 28.12	**+0.70**	[−1.13, 2.58]
2Wiki				
Associative	BM25	38.78 → 56.73	**+17.96**	[13.99, 21.79]
	Dense	47.35 → 56.73	**+9.39**	[5.53, 13.21]
	HippoRAG2	52.76 → 56.73	**+3.98**	[0.62, 7.26]
Hybrid	BM25	38.78 → 52.35	**+13.57**	[9.79, 17.32]
	Dense	47.35 → 52.35	**+5.00**	[1.95, 8.14]
	HippoRAG2	52.76 → 52.35	−0.41	[−3.52, 2.77]
HotpotQA				
Associative	BM25	43.70 → 53.33	**+9.63**	[6.67, 12.59]
	Dense	58.02 → 53.33	−4.69	[−7.16, −2.22]
	HippoRAG2	57.90 → 53.33	−4.57	[−6.91, −2.22]
Hybrid	BM25	43.70 → 59.01	**+15.31**	[12.72, 18.02]
	Dense	58.02 → 59.01	**+0.99**	[−0.49, 2.59]
	HippoRAG2	57.90 → 59.01	**+1.11**	[−0.62, 2.96]

Table B12: **Answer quality across arrivals (F1).** Scores (%) are shown as baseline → Memini. Δ is the Memini score minus the baseline score, in percentage points. Brackets give the paired 95% confidence interval for this difference. Positive differences are bold, highlighting where Memini scores higher.

Readout	Baseline	Scores (%)	Δ	95% CI
MuSiQue				
Associative	BM25	24.92 → 36.10	**+11.18**	[8.76, 13.62]
	Dense	35.89 → 36.10	**+0.22**	[−2.23, 2.65]
	HippoRAG2	39.28 → 36.10	−3.18	[−5.49, −0.92]
Hybrid	BM25	24.92 → 39.75	**+14.83**	[12.62, 17.13]
	Dense	35.89 → 39.75	**+3.87**	[2.21, 5.56]
	HippoRAG2	39.28 → 39.75	**+0.47**	[−1.41, 2.32]
2Wiki				
Associative	BM25	44.30 → 65.00	**+20.70**	[16.75, 24.58]
	Dense	54.12 → 65.00	**+10.88**	[7.11, 14.74]
	HippoRAG2	62.31 → 65.00	**+2.68**	[−0.58, 5.98]
Hybrid	BM25	44.30 → 61.83	**+17.53**	[13.75, 21.30]
	Dense	54.12 → 61.83	**+7.71**	[4.63, 10.85]
	HippoRAG2	62.31 → 61.83	−0.49	[−3.50, 2.52]
HotpotQA				
Associative	BM25	57.25 → 68.83	**+11.58**	[8.81, 14.30]
	Dense	73.09 → 68.83	−4.26	[−6.48, −2.08]
	HippoRAG2	73.97 → 68.83	−5.14	[−7.22, −3.09]
Hybrid	BM25	57.25 → 74.52	**+17.28**	[14.82, 19.77]
	Dense	73.09 → 74.52	**+1.44**	[0.14, 2.77]
	HippoRAG2	73.97 → 74.52	**+0.56**	[−0.97, 2.12]

Hybrid improves cross-arrival F1 over Dense on all three datasets, with gains of 1.44–7.71 points supported by paired 95% intervals. Its F1 differences from HippoRAG2 have intervals including zero. Standalone associative recall's clearest advantage is joint recovery on 2Wiki, exceeding HippoRAG2 by 14.18 Joint@5 points and also improving EM by 3.98 points, with both intervals above zero. These results show complementary gains: more successful answering than Dense through hybrid access, and stronger joint evidence recovery on 2Wiki through associative recall.

B.4 Contributions of associative organization

This section details the shuffling of route destinations described in Section 5.4 and its effect on retained answering, then uses two worked examples to explain reinforcement's mixed contribution.

Learned connections. At each checkpoint, we shuffle the destination concepts of the directed route records. Each record retains its source concept, original anchor (witness) episode, source-to-anchor membership strength and pair-quality value. Each concept keeps the same number of outgoing and incoming route records, counting separate witnesses separately. Reassigned connections remain reciprocal, and self-links and duplicate source–destination–witness records are excluded. Source episodes and all concept–episode memberships remain unchanged. The reported answering comparison uses shuffle seed 41210.

The reader scores each modified route using its retained membership and quality, with question focus evaluated for the new destination. It then follows that destination's existing memberships to target episodes. The modified route remains usable even when the new destination is absent from the retained witness; its quality is reused without recomputation. Shuffling therefore redirects associative access through different destination concepts while retaining the original route evidence.

Table B13: **Learned connections: retained answer F1 (%).** Δ is the original score minus the shuffled score, in percentage points. Brackets give the paired 95% confidence interval. Positive differences are bold, highlighting higher scores with the learned connections. Differences are calculated before rounding.

Dataset	Readout	Original	Shuffled	Δ	95% CI
MuSiQue	Associative	43.61	12.32	$+\mathbf{31.29}$	[28.40, 34.18]
	Hybrid	50.22	45.01	$+\mathbf{5.21}$	[3.29, 7.21]
2Wiki	Associative	68.37	11.68	$+\mathbf{56.68}$	[53.71, 59.70]
	Hybrid	67.23	58.64	$+\mathbf{8.59}$	[6.43, 10.71]
HotpotQA	Associative	68.56	16.00	$+\mathbf{52.55}$	[49.62, 55.52]
	Hybrid	73.92	71.96	$+\mathbf{1.96}$	[0.55, 3.36]

Table B13 shows lower retained F1 after shuffling for both readouts across all three datasets; all six paired 95% intervals exclude zero. The losses support the contribution of the specific learned connections to answering after updates. They persist with the hybrid's four dense selections fixed, showing that learned associative access complements dense retrieval.

Reinforcement and evidence selection. Reinforcement adds a bounded activation boost when distinct anchor episodes support reaching the same bridge, supplementing its strongest route. Table 4 tests this contribution while preserving the learned memory, pair qualities and strongest-route recall. The examples below explain its small, mixed effects through changes in the five-episode context. Each figure shows source excerpts, ranked contexts and generated answers with reinforcement off and on.

Example 3 (Figure B3) shows reinforcement completing an evidence chain in the hybrid readout on HotpotQA. Answering requires identifying SS501's main rapper and then the station that aired his 2014 television series. The first four dense selections remain fixed, so the comparison isolates the episode contributed in fifth position.

Question: On which station did the 2014 South Korean television series, starring the main rapper of boyband SS501, air?
Reference answer: KBS2.

Supporting evidence and the displaced episode

Episode title	Relevant source excerpt
Kim Hyun-joong	"the leader and **main rapper of boyband SS501**."
Inspiring Generation	"a 2014 South Korean television series starring **Kim Hyun-joong**"; "It aired on **KBS2**"
Heo Young-saeng	"the **main vocalist of boyband SS501**."

Retrieved context and generated answer

Retrieval rank	Reinforcement off	Reinforcement on
1	Kim Hyun-joong	Kim Hyun-joong
2	Baro (singer)	Baro (singer)
3	Kim Hyung-jun	Kim Hyung-jun
4	Ok Taec-yeon	Ok Taec-yeon
5	**Heo Young-saeng**	**Inspiring Generation**
Generated answer	The information is insufficient to determine the station.	KBS2.

Bold episode titles mark the episode entering or leaving the context.

Figure B3: **Example 3: reinforcement recovers missing evidence.** At final retention on HotpotQA, the hybrid's four dense selections stay fixed. Reinforcement retrieves *Inspiring Generation* in fifth position, supplying the missing station and changing the generated answer from insufficient information to KBS2.

Additional anchors increase activation of the Kim Hyun-joong bridge, bringing *Inspiring Generation* into fifth position in place of *Heo Young-saeng*. The displaced episode identifies another member of SS501 but supplies no television station. The series episode completes the connection from SS501's

main rapper to his 2014 series and KBS2. With the four dense selections unchanged, this single replacement changes the generated answer from insufficient information to the correct station. Here, support from additional episodes brings required evidence into the context that the strongest route alone misses, enabling the correct answer.

Example 4 (Figure B4) shows a different consequence in associative recall on 2Wiki. The nationality comparison requires both Gus Triandos's and Jen Armbruster's biographies. Reinforcement instead promotes *Myers Field*, an episode connected to Triandos through baseball.

Question: Do Jen Armbruster and Gus Triandos have the same nationality?
Reference answer: Yes.

Supporting evidence and the promoted episode

Episode title	Relevant source excerpt
Gus Triandos	"was a Greek **American** professional baseball player and scout."; "a catcher and a first baseman for the **New York Yankees**"
Jen Armbruster	"is an **American** goalball player."
Myers Field	"the Norfolk Tars, a **New York Yankees** minor league affiliate"; "**Gus Triandos** … called Myers Field home."

Retrieved context and generated answer

Retrieval rank	Reinforcement off	Reinforcement on
1	Gus Triandos	Gus Triandos
2	Don Leppert (second baseman)	Don Leppert (second baseman)
3	Jack Spring	**Myers Field**
4	**Jen Armbruster**	Jimmy Paredes
5	Jimmy Paredes	Jack Spring
Generated answer	Yes.	Cannot determine.

Bold episode titles mark the episode entering or leaving the context.

Figure B4: **Example 4: reinforcement displaces required evidence.** At final retention on 2Wiki, Memini's associative readout promotes the related baseball episode *Myers Field*, while *Jen Armbruster* leaves the five-episode context. Her nationality is then missing from the retrieved evidence, and the generated answer changes from Yes to Cannot determine.

The *Gus Triandos* episode connects him to the New York Yankees. Two other episodes support this connection: *Don Leppert (second baseman)* describes a Yankees trade involving Triandos, while *Myers Field* describes the Yankees affiliate for which he played. These episodes provide converging support for the New York Yankees bridge. Reinforcement boosts its activation, bringing *Myers Field* into the context and displacing *Jen Armbruster*. Her biography remains stored, but the answer changes from Yes to Cannot determine because her nationality is now missing. The baseball connection is meaningful, but its additional support gives related evidence priority over the biography needed to establish the other person's nationality.

These examples explain the small, mixed effects in Table 4. The connections are learned from observations before future questions are known, so a meaningful association does not necessarily lead to the evidence required by a particular question. Reinforcement can recover missing evidence or displace another required source.

B.5 Writer and cue-model choice

Following Sections 5.1 and 5.5, we examine model choice in memory formation and repeated access. We vary the cue model with the writer fixed, then the writer with cues fixed, measuring evidence recall and answer F1 at acquisition and retention.

Cheaper cues for repeated access. Memini writes each episode once but extracts cues for every new question. As its memory is reused, cue extraction therefore becomes a recurring cost. We test GPT-4o-mini as a cheaper way to access that memory while retaining the stronger model for episode writing. Table B14 compares evidence recall and answer F1 with the two cue models on the same written memory, at acquisition and retention.

Table B14: **Cue-model substitution with GPT-5.4 writing fixed.** Acquisition and retention R@5 and answer F1 (%). Scores show GPT-5.4 → GPT-4o-mini cues.

	Acquisition		Retention	
Readout	R@5	F1	R@5	F1
MuSiQue				
Associative	61.24 → 61.45	44.54 → 44.17	59.69 → 59.90	43.61 → 43.36
Hybrid	76.46 → 76.20	51.16 → 51.09	74.58 → 74.25	50.22 → 50.07
2Wiki				
Associative	91.46 → 91.86	69.32 → 70.42	90.85 → 91.35	68.37 → 69.58
Hybrid	89.98 → 90.24	68.10 → 68.47	88.80 → 89.13	67.23 → 67.76
HotpotQA				
Associative	86.48 → 85.73	69.02 → 68.42	85.55 → 84.75	68.56 → 68.01
Hybrid	96.59 → 96.48	73.66 → 73.59	96.15 → 96.00	73.92 → 73.77

Across both readouts and all three datasets, replacing GPT-5.4 cues with GPT-4o-mini changes acquisition and retention scores by only -0.80 to $+0.50$ points in R@5 and -0.60 to $+1.22$ points in answer F1. In a separate paired experiment on the hybrid readout, this substitution reduces complete-query API cost, including answering, by 80–82% (Appendix E.3). Together, these results support using a lighter cue model for repeated access, preserving similar observed retrieval and answering performance while substantially reducing recurring API cost.

Writer choice. Having examined cheaper access to an existing memory, we next vary the model used to form it. Table B15 holds GPT-4o-mini cues fixed and changes the episode writer, testing its effect on evidence recovery and answering with both readouts.

Table B15: **Writer substitution with GPT-4o-mini cues fixed.** Acquisition (Acq.) and retention (Ret.) R@5 and answer F1 (%). Memini's Table 2 results use GPT-5.4 cues.

	MuSiQue				2Wiki				HotpotQA			
	Acq.		Ret.		Acq.		Ret.		Acq.		Ret.	
Episode writer	R@5	F1	R@5	F1	R@5	F1	R@5	F1	R@5	F1	R@5	F1
Associative												
GPT-5.4	61.45	44.17	59.90	43.36	91.86	70.42	91.35	69.58	85.73	68.42	84.75	68.01
GPT-5.4-mini	60.74	43.39	58.93	41.94	90.91	70.22	90.20	69.61	84.46	67.69	83.75	67.10
GPT-4o-mini	58.19	41.59	56.76	40.44	81.55	63.89	81.10	62.86	82.51	66.81	81.60	65.94
Hybrid												
GPT-5.4	76.20	51.09	74.25	50.07	90.24	68.47	89.13	67.76	96.48	73.59	96.00	73.77
GPT-5.4-mini	76.26	50.72	74.39	49.67	90.04	68.60	89.05	67.94	96.24	73.88	95.85	73.71
GPT-4o-mini	76.31	51.26	74.43	50.07	89.15	68.21	88.18	67.41	96.24	73.71	95.85	73.37

Standalone associative recall is more sensitive to writer choice than to cue substitution. GPT-5.4-mini produces relatively small changes in retrieval and answering, while GPT-4o-mini writing produces larger reductions, particularly on 2Wiki, where retained F1 falls from 69.58 to 62.86.

Hybrid retrieval and answering remain comparatively stable across the tested writers. Relative to GPT-5.4 writing, R@5 changes by at most 0.24 points with GPT-5.4-mini and 1.10 points with GPT-4o-mini across datasets and stages; answer F1 changes by at most 0.41 points with either writer. On 2Wiki, GPT-4o-mini writing changes hybrid retained F1 only from 67.76 to 67.41, compared with the larger decline for standalone associative recall.

Together, these results support lighter cues for repeated access with similar observed performance and show that hybrid answering is less sensitive to writer choice than standalone associative recall.

B.6 HYBRID ALLOCATION

We examine how dividing the five-episode context between dense and associative retrieval affects performance after updates. The 4+1 allocation was fixed across all main experiments before this sensitivity study. We compare it with 3+2 and 2+3, keeping the written memory, GPT-5.4 cues, answerer and continual evaluation protocol fixed. Each allocation preserves its dense prefix and fills the remaining positions with distinct associative selections, using dense fallback when needed. Table B16 reports retained performance and answer-F1 forgetting.

Table B16: **Hybrid allocation under continual updates.** Allocations specify dense + associative selections. Retained scores are percentages; F1 forgetting is measured in points. Higher retained scores and lower forgetting are better.

Dataset	Allocation	R@5 ↑	C@5 ↑	F1 ↑	F1 forgetting ↓
MuSiQue	4+1	74.58	46.20	50.22	3.65
	3+2	73.42	44.90	50.26	4.54
	2+3	71.52	42.10	48.40	4.67
2Wiki	4+1	88.80	69.00	67.23	3.63
	3+2	92.13	79.00	70.35	2.40
	2+3	93.20	82.20	71.21	2.52
HotpotQA	4+1	96.15	92.60	73.92	0.89
	3+2	96.15	92.60	73.61	1.19
	2+3	94.90	90.10	73.52	1.51

All three allocations retain higher observed recall, completeness and answer F1 than Dense across all three datasets. Additional associative slots are particularly effective on 2Wiki, consistent with its strong associative recovery in Appendix B.2. Relative to 4+1, 3+2 and 2+3 increase retained F1 by 3.12 [1.54, 4.76] and 3.98 [2.19, 5.75] points, respectively; their F1-forgetting differences are -1.22 [-2.22, -0.22] and -1.11 [-2.18, -0.04] points. Brackets give paired 95% intervals from 10,000 question resamples, with arrival orders kept together; differences are alternative minus 4+1, calculated before rounding.

On MuSiQue, 4+1 retains more evidence, while 3+2 achieves similar retained F1. On HotpotQA, 3+2 matches its retained retrieval scores and answering remains close across allocations. The 4+1 allocation has the lowest observed F1 forgetting on both datasets, while 2+3 achieves the highest retained answering on 2Wiki. Thus, the common allocation fixed for the main experiments captures the benefit of combining the two access mechanisms, while their balance affects evidence recovery, retained answering and forgetting differently across datasets.

C MEMORY FORMATION AND RECALL DETAILS

This appendix details how Memini forms and uses the external memory introduced in Section 4. We describe how source episodes become checked local evidence, how that evidence is integrated as observations arrive, and how a question produces the associative and hybrid readouts. A worked example then connects the stored records to the retrieval process.

C.1 FROM SOURCE EPISODES TO CHECKED LOCAL EVIDENCE

Memory formation begins by identifying reusable concepts while preserving the source of each observation. The episode writer receives only the arriving episode's title and body. It proposes local concepts, each with a name, a coarse type and a group of coreferent mentions. The deterministic checker resolves each proposed mention to the declared title or body. It first searches for an exact match; if none is found, it applies conservative normalization for matching, handling Unicode presentation, case, quotation marks, dashes and combining accents while retaining token boundaries. This repair preserves the original source text and offsets. Table C1 summarizes the acceptance and rejection conditions; the writer's instructions and output format appear in Appendix D.1.

Table C1: **Admission of source evidence.** Checking resolves proposed mentions against the unchanged episode. Accepted mentions retain their source text, title/body provenance and character offsets.

Source condition	Checker action
Exact match in the declared title or body	Accept token-bounded occurrences; repeated names and acronyms can supply several mentions.
Repair matches one distinct source string	Accept its occurrence or occurrences, preserving the actual source text and repair status.
Match exists only in the other source	Reject if found only in the title instead of the declared body, or vice versa; do not change its attribution.
No match, or repair matches distinct source strings	Reject the mention: it is absent or cannot be matched unambiguously.
Repeated reference cannot be uniquely located	Reject repeated pronouns or generic descriptions under exact matching or repair.
Local concept has no surviving mentions or selectable name	Drop the local concept from the checked record.

For each proposed local concept, the checker selects one canonical name from its accepted mentions, which may contain several candidate names. The following rules apply in order; the first applicable rule determines the selection, with ties resolved as specified:

1. **Prefer a qualified title mention.** Select an accepted title mention with a parenthetical or comma qualifier, such as "Mercury (planet)" or "Carson, California", subject to the checker's title-pattern exclusions. Among these candidates, prefer the earliest start offset, then the longer span, then an exact rather than repaired match.
2. **Prefer the writer's proposed name.** Otherwise, look for an accepted mention matching the proposed name after repair normalization. If the proposed name is classified as a pronoun, require a matching title mention; prefer the earliest start offset, then the shorter span. If not, prefer an exact match over a repaired match, then the earliest start offset.
3. **Select another accepted mention.** Otherwise, exclude pronoun mentions and prefer candidates that are not acronyms, when available. Order the candidates by title before body, longer span, earlier start offset and case-folded lexical order. Drop the concept if no candidate remains.

These checks establish where a mention occurs; the writer's within-episode coreference grouping remains a semantic decision.

For each accepted local concept, Memini records its mention count, title presence and body-sentence occurrences. Within that concept, accepted mentions with identical source (title or body), start and end offsets are counted once. Overlapping spans with different offsets remain separate mentions. Title presence indicates whether any accepted mention occurs in the title.

For the body, each concept records the set of sentences containing its accepted mentions, assigning each mention by its start offset. Sentence boundaries come from the same deterministic sentence splitter across all three datasets: it splits after a period, exclamation mark or question mark followed by whitespace and an uppercase ASCII letter or digit, allowing adjacent ASCII quotes or brackets. Case-sensitive guards prevent splits after Mr, Mrs, Ms, Dr, St, No, vs and etc. Each sentence is recorded once per concept, regardless of how many of its mentions occur there.

Memini also records every unordered pair of distinct local concepts in the episode, identifying shared body sentences by intersecting their sentence sets. The pair is recorded even when the concepts share no body sentence, since both occur in the same episode. Thus, the same source supplies both concept–episode evidence and finer information about where concepts occurred together. These records preserve co-observation without requiring the writer to assign a named relation. Their local identities can now be connected to concepts encountered in earlier episodes.

C.2 INTEGRATING EVIDENCE ACROSS ARRIVALS

Integration connects recurring concepts while keeping their episode-specific evidence distinct. A checked name is normalized using Unicode NFKC, case folding, removal of surrounding punctuation and whitespace collapse. Equal normalized names map to the same global concept; coarse types do not alter this identity rule. This provides deterministic links across episodes from checked names, without an additional semantic identity decision. No stemming is applied. The displayed name is the most frequent checked local name, with lexical tie breaking. Checked name and acronym surfaces remain available for query lookup.

When several local concepts in one episode map to the same global concept, their mention counts and title indicators are consolidated before computing the membership weight in Equation 1. Consolidation adds the mention counts already recorded for each local concept and sets the title indicator if any has an accepted title mention. The weight's logarithmic term lets repeated mentions strengthen membership with diminishing influence, while title presence provides a prominence signal.

Mapped pairs are consolidated similarly: self-pairs are removed, shared-sentence sets are united, and title support is retained if either endpoint has an accepted title mention. Each episode contributes at most one observation to an unordered global pair. The episode's number of distinct global concepts supplies the size term in Equation 4.

Algorithm C1 Integrate newly arriving episodes

Input: current archive D_t, associative state A_t, and new episodes Δ.
Output: updated archive and associative state.

1. Reject episode identifiers already in D_t or repeated within Δ.
2. For each new episode, preserve its source text and obtain checked local evidence as in Appendix C.1.
3. Map checked names to existing or new global concepts; retain the local-to-global mapping.
4. Consolidate concept–episode evidence and compute one membership weight per global concept and episode.
5. Consolidate mapped pairs, discard self-pairs, and add the new episode as one witness to each remaining pair.
6. Incrementally update concept name/type counts, concept–episode membership maps, pair-witness records and counts, and checked-surface mention counts; retain the checked evidence.

Algorithm C1 distinguishes repetition within an observation from an additional episode. In the evaluated streams, episode identifiers combine the dataset name with a hash of the episode's exact title and body. Identical title–body units within a dataset therefore share one identifier and enter the archive once. More mentions can change a membership weight, but only a new episode identifier adds another pair witness; questions do not increment witness counts. Shared concepts therefore connect earlier and newly arriving episodes without losing the provenance of each observation.

Earlier episodes are not passed through the writer again. Before querying each checkpoint, the reader reconstructs compact membership and anchor–bridge arrays from the retained records, computing normalized membership shares and local pair qualities for the current memory. Sorted checked-surface lookups are formed from the accumulated mention counts. This preparation reuses stored evidence rather than re-ingesting historical sources. The resulting state provides the memberships and witnessed routes used by the next question.

C.3 QUERY PREPARATION AND ASSOCIATIVE RECALL

A question first finds entry concepts in the current memory, then prioritizes episodes reachable through their associations. We specify cue matching, route combination and the two readouts below, using the scoring equations in Section 4.3. These operations leave the persistent memory unchanged.

Preparing the cues. The cue extractor identifies concepts named in the question (Appendix D.2). Each cue uses the same name normalization as integration and the first acceptable mapping tier in Table C2. This ordering prioritizes direct identity matches, then accommodates surface variation within explicit ambiguity limits. When several cues reach one global concept, Memini retains the largest assigned mass. Normalizing the positive masses to sum to one gives $c_q(g)$; repeated cues therefore do not accumulate extra mass on the same concept.

Table C2: **Mapping one cue to the current memory.** Tiers are tried in order. The masses below precede normalization across the question's mapped concepts.

Matching tier	Acceptance and selection	Assigned mass
Exact concept name	An existing global name matches after normalization.	1.0 to that concept.
Checked surface	An exact normalized name/acronym surface maps to at most 50 concepts. Keep at most three, ordered by accepted mention count, then global identifier.	Total 0.8, divided in proportion to the retained counts.
Contiguous name match	The cue's tokens occur contiguously within 1–20 global names. More than 20 matches rejects this tier.	Total 0.4, divided equally among the matches.

In the checked-surface tier, the count is the number of accepted name/acronym mentions of that normalized surface assigned to a global concept across stored episodes, excluding pronouns and generic nominals. It counts mention records, including repeated occurrences within one episode, rather than distinct witnessing episodes.

Only when extraction returns no cues does Memini select the three currently visible concept names most similar to the question. This fallback uses the encoder's concept-query instruction, with ties following the fixed concept-inventory order. The selected names enter the exact-name tier with equal mass 1.0, giving $1/3$ each after normalization; similarity determines selection, not cue mass. If the extractor returns cues but none maps to a concept in memory, associative scores remain zero.

To prioritize reachable evidence, the complete question is embedded using the encoder's passage-query instruction for dense ranking and associative focus. Focus uses cosine similarities to episode and concept-name embeddings, setting negative values to zero. Appendix A.1 gives model and encoder settings. The exact query instructions are:

- **Passage query:** "Given a question, retrieve relevant documents that best answer the question."
- **Concept query:** "Given a question, retrieve relevant phrases that are mentioned in this question."

Each query is formatted as `Instruct:` followed by the instruction, a newline, then `Query:` followed by the complete question, with one space after each colon. Episode texts (title, newline, body) and concept names are embedded without an instruction.

Combining witnessed routes. A mapped cue reaches a bridge concept through a pair observation in an anchor episode. Membership normalization distributes activation across all currently visible episodes containing that cue concept in proportion to their membership strengths (Equation 3). A concept occurring in many episodes therefore does not assign its full activation independently to each.

Local pair quality distinguishes how the two concepts were observed together. Same-sentence occurrences receive greater weight because they share a narrower textual context; pairs without this support retain weaker episode-level support. An accepted title mention provides a prominence signal. The episode-size discount tempers pair strengths when many concepts co-occur and generate many pair observations. These factors grade stored co-observation independently of the question. Cue mass and question focus then determine which witnessed routes contribute to recall and how strongly (Equation 5).

If several cues reach the same bridge through the same anchor, only the largest contribution is retained. Distinct anchors with positive contributions determine $k_q(b)$ in Equation 6. This quantity can combine anchors reached from different cues, whereas $n_t(a, b)$ counts stored witnesses for one concept pair. For each bridge, reinforcement counts distinct contributing anchor episodes; multiple cue routes through the same anchor count only once. A single active anchor supplies the base activation without a reinforcement boost.

Algorithm C2 Recall source episodes from the current memory

Input: question q, current archive D_t, associative state A_t, and frozen encoders.
Output: associative and hybrid episode selections, each with at most five episodes.

1. Extract cues, apply the empty-extraction rule if needed, and map them using Table C2. Retain maximum masses and normalize to obtain c_q.
2. Compute question focus for episodes and concept names, and the independent dense episode ranking.
3. For each witnessed cue–anchor–bridge route, evaluate Equation 5. Retain the maximum for each anchor–bridge pair.
4. For each bridge with positive contributions, combine distinct anchors using Equation 6.
5. Compute episode scores through bridge memberships using Equation 7. Retain the strongest contribution per episode, normalize positive scores and rank them.
6. **Associative:** select the five highest-ranked positive-scoring episodes, or all if fewer exist.
7. **Hybrid:** preserve the dense top four; add the highest-ranked positive-scoring associative episode not already selected. If none is available, fill the fifth slot with the highest-ranked remaining dense episode.

Returning the answering context. Each target episode receives its strongest bridge contribution, so overlapping routes through different bridges do not accumulate in its score. Positive episode scores are normalized to sum to one; this rescaling preserves their order. Equal episode scores follow the fixed corpus order. The associative readout has no dense fill rule and can return fewer than five episodes. The hybrid preserves four dense selections to retain direct semantic matches, while the remaining slot allows associative recall to supply additional evidence. Algorithm C2 specifies duplicate handling and dense fallback.

These operations perform bounded parallel recall: several bridges can recover several related episodes in one traversal, without recursively expanding new search chains. The selected episodes retain their complete original text and provenance when supplied with the question to the shared answerer (Appendix D.3). Persistent evidence determines which routes exist; the question determines their temporary activation and selection. Neither retrieval nor the generated answer updates that evidence.

C.4 Worked example: from stored evidence to retrieved episodes

We reuse the 2Wiki question from Example 1 in Appendix B.1 to show how stored evidence supports recall. Answering requires linking each film to its director and comparing their death dates. Figure C1 shows selected records, routes to the biographies and the associative ranking at checkpoint 4, after all episodes have arrived, for the first of the four evaluated arrival orders.

The same director concept connects each film episode to its biography. Mention counts record repeated occurrences within an episode; total concept counts determine the episode-size discount.

The film-name cues reach the director concepts through episodes A and C. These shared concepts then provide access to biographies B and D, even though the biographies do not mention the films. The displayed routes explain how the missing biographical evidence is recovered. The final ranking also includes contributions from other concepts: *From Hong Kong with Love*, an alternative film name recorded in C, provides that episode's strongest contribution.

Writing connects recurring concepts while preserving episode-specific evidence; recall uses those connections to recover original sources. Here, the biographies supply the dates needed for the saved answer, *Bons Baisers de Hong Kong*. The learned organization enables access, and the recovered episodes provide the evidence for the comparison.

Question: Which film has the director who died later, 45 Calibre Echo or Bons Baisers De Hong Kong?
Reference answer: Bons Baisers de Hong Kong.

Stored evidence: selected records from the four supporting episodes

ID	Episode title and exact source excerpt	Relevant concepts and checked evidence
A	45 Calibre Echo "directed by Bruce M. Mitchell"	**45 Calibre Echo; Bruce M. Mitchell.** They share a body sentence; the film has title support. 7 concepts in the episode.
B	Bruce M. Mitchell "September 26, 1952"	**Bruce M. Mitchell.** Four accepted mentions of this concept, including the title.
C	Bons Baisers de Hong Kong "also known as From Hong Kong with Love"; "directed by Yvan Chiffre"	**Bons Baisers de Hong Kong; From Hong Kong with Love; Yvan Chiffre.** All three share a body sentence; the first film name has title support. 14 concepts in the episode.
D	Yvan Chiffre "27 September 2016"	**Yvan Chiffre.** Three accepted mentions of this concept, including the title.

Recall: the two routes recovering the director biographies

The question yields two film-name cues, **45 Calibre Echo** and **Bons Baisers de Hong Kong**. After mapping to memory and normalization, each receives cue mass 0.5. The two routes below use film episodes A and C as anchors, the corresponding director concepts as bridges, and biographies B and D as target episodes.

Component	A → B via Bruce M. Mitchell	C → D via Yvan Chiffre
Cue membership share $m(a, e)$	1.000000	0.654313
Anchor focus $g_q(e)$	0.431843	0.392369
Pair quality $q_{\text{pair}}(e, a, b)$	0.473197	0.375000
Bridge focus $g_q(b)$	0.056114	0.087352
Route contribution $r_q(a, e, b)$	0.005733	0.004205
Biography membership share $m(b, u)$	0.550197	0.687252
Biography focus $g_q(u)$	0.114167	0.159657
Biography raw score $S_q(u)$	0.000360	0.000461

Multiply each column's first four factors by cue mass 0.5 to obtain the route contribution. Each director has one contributing anchor, $k_q(b) = 1$, so there is no reinforcement boost and $\alpha_q(b) = r_q(a, e, b)$. Multiplying this activation by the biography membership share and focus gives its raw score.

Final associative ranking from the full memory

Retrieval rank	Retrieved episode	Strongest bridge	Normalized score
1	**C**: Bons Baisers de Hong Kong	From Hong Kong with Love	0.566638
2	**A**: 45 Calibre Echo	Bruce M. Mitchell	0.072023
3	**D**: Yvan Chiffre	Yvan Chiffre	0.059676
4	**B**: Bruce M. Mitchell	Bruce M. Mitchell	0.046581
5	Won in the Clouds	Bruce M. Mitchell	0.014943

Scores are unitless and rounded to six decimals. The normalization denominator is 0.007731365, summed over all positive episode scores.

Figure C1: **Worked example of source-grounded associative recall.** At checkpoint 4 of the first evaluated 2Wiki arrival order, film cues reach director concepts through co-observations in A and C, then recover biographies B and D through their memberships. Selected stored records and route contributions explain this recovery; the final ranking uses the full memory. The four supporting episodes supply both directors' death dates for comparison.

D LANGUAGE-MODEL PROMPTS

The evaluated instructions, input templates and output schemas are shown below; omitted demonstrations are marked. Model settings appear in Appendix A.1.

D.1 EPISODE WRITER

Episode writer prompt

Instructions (developer message)

You extract reusable passage-local concepts from one titled passage.

This is passage ingestion before any question is known. Use only the supplied title and passage body. Never use world knowledge to add entities, aliases, dates, values, or relations that are not written in the title or body.

A local concept is a specific entity or salient concept explicitly named in the title or body that could later serve as a graph node. Do not output relationships, facts, or cross-passage nodes. Assign each concept exactly one type from this list:

- PERSON: a named individual.
- ORG: a named organization or institution (company, agency, team, label, observatory, university).
- LOCATION: a named place or physical/geographic feature (city, country, building, river, bridge).
- WORK: a named creative work (book, film, album, song, series, painting).
- EVENT: a named event (battle, draft, tournament, election, ceremony).
- GROUP: a named set of people that is not a formal organization (band, duo, family, dynasty, ethnic or national group).
- DATE: a central date or time expression.
- VALUE: a central number or measurement.
- OTHER: a central reusable concept that fits none of the above; use rarely.

When an entity is both an institution and a place (observatory, university, airport, stadium), use ORG if the passage treats it as an actor (it operates, funds, was founded, employs) and LOCATION if it treats it as a place (located in, near, at).

Aim for useful coverage of meaningful named concepts without exhaustive noun-phrase extraction. When unsure whether something is a reusable concept, leave it out. An empty passage may yield no concepts.

Title entity: when the title names a specific person, organization, group, location, work, event, or office, include it as a concept with a title-source mention. For a list/topic title such as "List of Bright Star episodes", you may extract an embedded work, but the full list title should not be listed as a mention of that work. A disambiguated title like "River Light (album)" is the same concept as shorter body mentions of the same thing; attach the title span and let later steps pick the canonical name.

Identity-safe coreference:

- Same referent in this passage -> one concept.
- Different referents -> separate concepts.
- Merely related things are not coreference. Do not merge a city with its airport, a country with its national team, a work with its creator or a character in it, an event with its location, or co-located entities in a title such as "Northland at the Solar Cup" (that is "Northland" and "Solar Cup", not one concept).
- Pronouns and generic nominals ("she", "the album") are evidence only: list them as mentions under a named concept, never as a concept name.

Roles, offices, and kinship: extract the named entity, not the relation phrase. From "Mara Solis was president of Westland" extract "Mara Solis" and "Westland", not "president of

Episode writer prompt (continued)

Westland". Phrases like "his father" or "the founder" are not concepts when the actual person or organization is named. The office itself is a concept only when the passage is about that office (e.g. a page titled "President of Westland" about the office).

Tables and lists: Do not exhaustively extract row/cell entities. Skip per-row players, teams, episodes, picks, scores, and repeated years; extract only the central event or organizing entity, and a row entity only when prose makes it central.

DATE and VALUE: extract only when central or likely answer-bearing, not every year, score, runtime, capacity, or coordinate. Keep the meaningful source phrase: "early in the 20th century" should not be shortened to "20th century". Do not invent normalized forms. Use OTHER rarely, only for a central reusable concept, never as a bucket for generic nouns, roles, or relation words.

Each mention must be copied verbatim from the title or body (the entity itself, not surrounding words or descriptive wrappers), with source "title" for title text and "body" for body text.

Return only what is named in the title or body. If nothing qualifies, return {"concepts": []}.

[Four fixed demonstrations omitted.]

Return valid JSON only. No markdown, comments, explanations, or extra fields.

Input template (user message)

```
Title:
{title}

Passage body:
{text}
```

Output format (strict JSON schema)

```
{"type": "object", "additionalProperties": false,
  "properties": {"concepts": {"type": "array",
      "items": {"type": "object", "additionalProperties": false,
        "properties": {"name": {"type": "string"},
          "type": {"type": "string",
            "enum": ["PERSON", "ORG", "LOCATION", "WORK", "EVENT",
              "GROUP", "DATE",
              "VALUE", "OTHER"]},
          "mentions": {"type": "array",
            "items": {"type": "object",
              "additionalProperties": false,
              "properties": {"text": {"type": "string"},
                "source": {"type": "string", "enum": ["title",
                  "body"]}},
              "required": ["text", "source"]}}},
        "required": ["name", "type", "mentions"]}}},
  "required": ["concepts"]}
```

D.2 CUE EXTRACTOR

Cue extractor prompt

Instructions (developer message)

Extract the cue concepts from one multi-hop question.

List every specific entity, creative-work title, organization, place, event, group, or central date/value that is NAMED in the question text, copied verbatim (keep casing). Do NOT answer the question. Do NOT add entities that are merely implied or would come from world knowledge. Descriptive references that name nothing specific ("the company", "the director") are not concepts. Assign each concept one type: PERSON, ORG, LOCATION, WORK, EVENT, GROUP, DATE, VALUE, OTHER.

Input template (user message)

```
{question_text}
```

Output format (strict JSON schema)

```
{"type": "object", "additionalProperties": false,
  "required": ["concepts"],
  "properties": {"concepts": {"type": "array",
      "items": {"type": "object", "additionalProperties": false,
        "required": ["name", "type"],
        "properties": {"name": {"type": "string"},
          "type": {"type": "string",
            "enum": ["PERSON", "ORG", "LOCATION", "WORK", "EVENT",
              "GROUP", "DATE",
              "VALUE", "OTHER"]}}}}}}
```

D.3 SHARED ANSWERER

All methods use HippoRAG2's released QA prompt and demonstration.

Shared answerer prompt

Instructions (system message)

As an advanced reading comprehension assistant, your task is to analyze text passages and corresponding questions meticulously. Your response start after "Thought: ", where you will methodically break down the reasoning process, illustrating how you arrive at conclusions. Conclude with "Answer: " to present a concise, definitive response, devoid of additional elaborations.

[Fixed user/assistant demonstration omitted.]

Input template (user message)

```
{retrieved_episodes}
Question: {question}
Thought:
```

Episodes are inserted in their returned order, each formatted as:

```
Wikipedia Title: {title}
{body}
```

E Resource Accounting

This appendix details the resource measurements in Section 5.6. Both readouts share one stored memory; model and request settings follow Appendix A.1.

Local integration (E.1), loaded retrieval (E.2) and complete answering (E.3) are measured on an Intel Core Ultra 9 275HX laptop with 32 GB RAM and an RTX 5070 Ti Laptop GPU, running Windows with one BLAS/OpenMP thread. LLM inference uses remote APIs. Loaded retrieval in E.2 uses synchronized CUDA float32 similarities with TF32 disabled and precomputed query vectors. The original live-answering study in E.3 includes fresh NV-Embed-v2 encoding in float16 with mixed CPU/GPU placement. Throughout this appendix, p95 denotes the 95th percentile of the stated sample.

E.1 Memory formation and storage

Figure E1 shows serialized memory size and hybrid retrieval latency as episodes accumulate. The retrieval measurements follow the protocol in Section E.2.

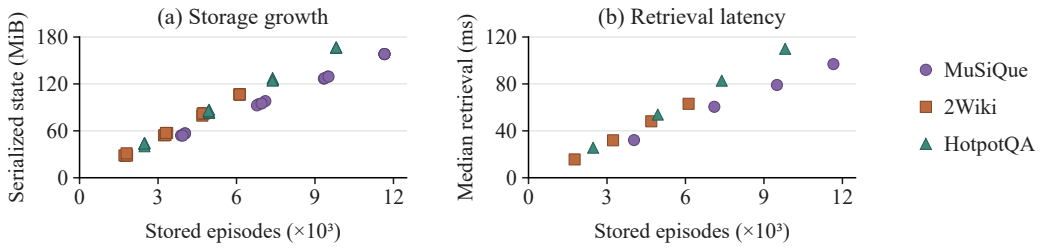


Figure E1: **Memini's storage growth and retrieval latency.** (a) Both readouts share state; embeddings/model weights excluded. (b) Hybrid retrieval excludes LLM calls and question encoding.

Figure E2 shows how concepts, memberships and pair observations accumulate across three datasets, four arrival orders and four checkpoints. With k_e selected concepts in episode e, local pair observations total $\sum_e k_e\,(k_e-1)/2$. Their count grows at most linearly with episode count when selected concept counts are bounded. The writer imposes no fixed cap: endpoint p95 concept counts are 20–27, and the largest 1% by concept count contribute 10.81–17.87% of pair observations.

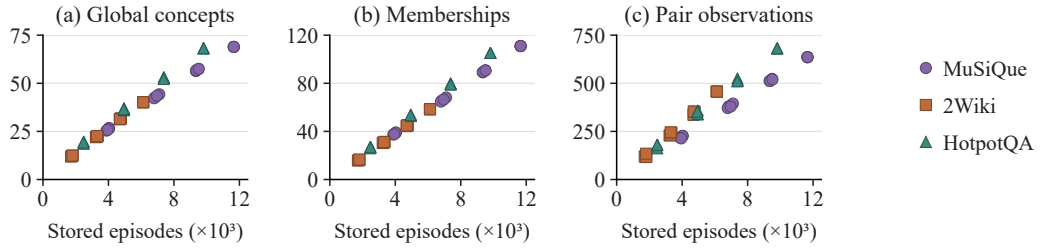


Figure E2: **Growth of Memini's stored associative evidence.** All panels show counts in thousands across 48 states; overlapping observations are retained. Memberships link concepts to episodes; pair observations count episode-local evidence.

Table E1 distinguishes mutable update checkpoints, prepared reader serialization and vectors. Writer costs use recorded successful token usage at the historical cached and noncached rates. Local integration is measured separately on 20 fixed new episodes after checkpoint 3, taking each episode's median over three repetitions. Local integration timings exclude writer inference and full reader rebuilding.

Table E1: **Stored memory, writing and local integration.** Final memories use the first arrival order. Each cost is per episode; storage representations are not additive.

Quantity	MuSiQue	2Wiki	HotpotQA
Stored episodes	11,656	6,119	9,811
Mutable checkpoint (MiB)	158.15	106.55	166.46
Prepared reader serialization (MiB)	109.86	76.55	116.70
Episode/concept vectors (GiB)	1.229	0.706	1.190
Writer input tokens: mean	2,203.9	2,198.9	2,220.1
Writer output tokens: mean	299.9	315.0	344.7
Writer API cost: mean (USD)	0.005978	0.006196	0.006693
Writer API cost: p95 (USD)	0.010452	0.014481	0.011466
Local integration: median (ms)	246.94	179.74	387.77
Local integration: p95 (ms)	454.65	310.99	644.72

E.2 Loaded retrieval

Table E2 reports both readouts at all four checkpoints of the first arrival order. The same 200 questions per dataset are selected by a fixed ID hash, independently of answerability or speed. After five warm-ups, each question is timed three times. We report the median and p95 of question-level repeat medians. Timing includes fresh similarities, seeding, query focus, associative scoring and ranking, with resolved cues and question vectors already available. Cue extraction, question encoding, answering and startup are excluded.

Table E2: **Loaded retrieval across memory sizes.** Median and p95 latencies in milliseconds for the two readouts of the shared memory.

		Associative readout		Hybrid readout	
Dataset	Episodes	Median	p95	Median	p95
MuSiQue	4,033	32.35	198.94	32.09	198.75
	7,103	60.82	66.09	60.47	65.42
	9,497	79.51	84.11	79.08	83.38
	11,656	97.37	104.20	96.89	103.07
2Wiki	1,762	15.67	19.87	15.62	19.82
	3,232	32.02	37.81	31.92	37.66
	4,693	48.27	62.09	48.11	61.93
	6,119	63.30	68.89	63.05	68.65
HotpotQA	2,470	25.72	29.42	25.63	29.21
	4,942	54.06	72.42	53.85	72.16
	7,388	83.09	88.88	82.77	88.31
	9,811	110.40	119.31	109.94	118.42

E.3 Complete answering and query costs

We measure the recurring cost of using stored memory to answer new questions. The evaluation uses 30 fixed questions per dataset, three repetitions and the final checkpoint of the first arrival order. Each complete query includes fresh question encoding, retrieval, answer generation and parsing, together with cue extraction for Memini or triple filtering for HippoRAG2. Stored memories and corpus vectors are reused; startup and memory construction are excluded. Latency is summarized by the median and p95 of question-level repeat medians.

Complete answering. Table E3 summarizes complete-query latency and API cost for Dense, HippoRAG2 and both Memini readouts. The hybrid readout is shown with GPT-5.4 and GPT-4o-mini cues to show how using a lighter cue model changes the total cost of answering a question.

Table E3: **Complete answering and query API costs.** Latencies are in seconds; API costs are USD per 1,000 queries, including answering and observed provider caching. Memory construction is excluded.

			Latency (s)		API cost (USD)
Dataset	Method	Cue model	Median	p95	per 1,000 queries
MuSiQue	Dense	—	1.839	2.364	0.1991
	HippoRAG2	—	5.213	5.863	0.4798
	Memini (associative)	GPT-5.4	3.052	3.829	1.2515
	Memini (hybrid)	GPT-5.4	3.044	3.812	1.2435
	Memini (hybrid)†	GPT-4o-mini	3.628	4.448	0.2372
2Wiki	Dense	—	1.759	2.508	0.2065
	HippoRAG2	—	4.624	5.536	0.4656
	Memini (associative)	GPT-5.4	2.975	3.672	1.2452
	Memini (hybrid)	GPT-5.4	3.016	3.520	1.2229
	Memini (hybrid)†	GPT-4o-mini	3.427	4.359	0.2334
HotpotQA	Dense	—	1.798	2.151	0.1734
	HippoRAG2	—	5.223	5.890	0.4517
	Memini (associative)	GPT-5.4	3.207	3.651	1.2864
	Memini (hybrid)	GPT-5.4	3.201	3.617	1.2787
	Memini (hybrid)†	GPT-4o-mini	3.795	4.824	0.2297

†These rows come from a separate experiment (Table E4), where both cue models were measured together. Their latencies should not be directly compared with the earlier GPT-5.4 measurements above.

Both Memini and HippoRAG2 perform additional query processing beyond dense retrieval alone, introducing a trade-off between serving overhead and the retrieval and answering benefits reported in Sections 5.2 and 5.3. Memini extracts question cues and follows learned connections without per-hop LLM calls, whereas HippoRAG2 uses an LLM to filter retrieved triples before graph retrieval. Both Memini readouts show lower observed latency than HippoRAG2 across all three datasets and have similar API costs with GPT-5.4 cues. Those costs exceed HippoRAG2's, motivating the following experiment on whether lighter cues can reduce recurring expense while preserving retrieval and answering performance.

Lighter cues for repeated access. Following Section 5.5, we compare GPT-5.4 and GPT-4o-mini cues in a separate paired experiment with the hybrid readout. Both cue models use the same questions and evaluation setup, while the writer, stored memory and answerer remain fixed. Table E4 reports the complete-query costs and reductions.

Table E4: **Complete-query API cost under paired cue-model substitution.** Costs are USD per 1,000 queries, including cue extraction and answering with observed provider caching; episode writing is excluded. GPT-5.4 costs are remeasured for this comparison.

Dataset	GPT-5.4 cues	GPT-4o-mini cues	Cost reduction
MuSiQue	1.2406	0.2372	80.9%
2Wiki	1.1888	0.2334	80.4%
HotpotQA	1.2815	0.2297	82.1%

GPT-4o-mini cues reduce complete-query API cost by 80.4–82.1%, with paired 95% intervals supporting the reduction on every dataset. The resulting cost is approximately \$0.23–\$0.24 per 1,000 queries, around half HippoRAG2's measured query cost. The separate continual evaluation in Appendix B.5 (Table B14) shows that, at acquisition and retention, the hybrid readout's scores change by at most 0.34 points in R@5 and -0.15 to $+0.54$ points in answer F1.

Together with the similar observed acquisition and retention performance in Appendix B.5, these measurements show that lighter cues enable substantially cheaper repeated access to the same written memory.

F Comparison with Related Systems

Table F1 summarizes the memory design and reported evaluation of the approaches discussed in Section 3. In the reported-evaluation columns, a dash indicates that the paper does not report the corresponding evaluation under the criteria specified in the caption; it does not imply that the method cannot support it.

Table F1: **Memory design and reported evaluation of related approaches.** *Frozen models* remain fixed during memory updates. *Observation-only writing* excludes task objectives, outcome feedback, and evaluation supervision. *Memory organization* is persistent structure available before queries. *Accumulated support* indicates whether repeated relations or co-observations strengthen connections. *LLM context* is the evidence supplied for answering. *Acquisition and retention* separately track retrieval and answering on questions supported by incoming observations and revisited after further updates. *Cross-arrival recombination* tests joint recovery and use of supporting evidence assigned to different arrivals. ✓: present; ✗: absent; —: not reported.

	Memory design					**Reported evaluation**	
Method	Frozen models	Observation-only writing	Memory organization	Accumulated support	LLM context	Acquisition and retention	Cross-arrival recombination
RAG	✓	✓	Passage index	✗	Source passages	—	—
RAPTOR	✓	✓	Summary hierarchy	✗	Source passages + summaries	—	—
GraphRAG (global)	✓	✓	Relation graph + summaries	✓	Partial answers (from reports)	—	—
LightRAG	✓	✓	Relation graph	✓	Source passages + descriptions	—	—
HippoRAG2	✓	✓	Phrase–passage graph	✓	Source passages	—	—
Panini	✓	✓	Document workspaces	✗	Generated QAs	—	—
Memini	✓	✓	Episode-supported concept pairs	✓	Source passages	✓	✓