DISTRIBUTED EVIDENCE MEMORY: DEFERRING IRREVERSIBLE DECISIONS FOR RELIABLE LONG-TERM MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-term memory is essential for language-model agents operating beyond a single context window. Existing systems often compress past experiences into query-independent representations and use *central gating* to decide what enters the current context. We identify a broader irreversible evidence bottleneck in this early-decision structure: relevant episodes may be missed, conflicting versions discarded, and retrieval terminated before sufficient evidence is collected. We introduce **Distributed Evidence Memory (DEM)**, which moves these decisions to stages informed by the current query and, when needed, accumulated evidence. Raw-memory shards inspect original episodes and report query-conditioned evidence through a shared workspace; a Conflict Directed Acyclic Graph (CDAG) retains conflicting versions until query-time aggregation; and Good-Turing Coverage Estimation (GTC) provides an evidence-based adaptive stopping signal. DEM consistently outperforms strong retrieval and memory baselines on FactCon, LoCoMo, and LongMemEval. On controlled FactCon with DeepSeek-v4-flash, SubEM rises from 41.5 to 64.0; mechanism analyses attribute 68% of the gain to improved evidence coverage and show stronger conditional second-hop recall and reduced post-retrieval evidence loss under conflicting versions. These results support a broader principle: preserve potentially relevant evidence until its relevance, validity, or sufficiency can be judged with query-conditioned evidence.

1 INTRODUCTION

Large language model (LLM) agents increasingly operate over interaction histories that extend far beyond a single context window. As these histories grow, repeatedly exposing the entire interaction record to the model becomes infeasible, making selective long-term memory access necessary. A practical memory system must therefore decide which past experiences should become available for the current query. This decision is consequential: downstream reasoning can only use evidence that survives the memory-access stage.

Many existing memory systems organize past experiences into stored representations such as embeddings, summaries, extracted facts, notes, or graphs, and perform retrieval over these representations (Packer et al., 2023; Zhong et al., 2024; Chhikara et al., 2025; Xu et al., 2025; Rasmussen et al., 2025; Wang et al., 2026). In many such systems, an episode E_i is represented by a query-independent $Z_i = g(E_i)$ before a centralized mechanism decides whether the episode should enter the active context. We refer to this representation-level selection pattern as *central gating*. It raises a fundamental question: *Can a memory system reliably decide which memories matter when the decision is made from compressed representations rather than the original episodic evidence?* If relevant evidence is excluded before the corresponding episode is examined, stronger downstream reasoning cannot recover what is no longer available. More broadly, central gating is one instance of an *early-decision structure* in which memory decisions become irreversible before enough query-conditioned evidence is available. We call the resulting failure an *irreversible evidence bottleneck*. As illustrated in Figure 1, it can arise at three points: premature admission can exclude relevant episodes, premature version selection can discard useful conflicting evidence, and premature stopping can terminate retrieval before sufficient evidence has been collected.

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.

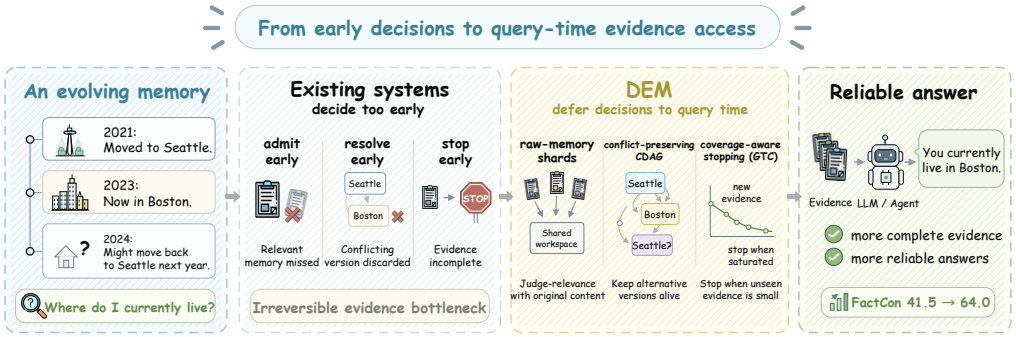


Figure 1: **From early selection to query-time evidence access.** Existing memory systems may make irreversible decisions before the relevant episodic evidence is fully available. DEM instead moves relevance judgment, conflict resolution, and retrieval stopping to stages informed by query-conditioned evidence.

Prior work has substantially improved memory representation, iterative retrieval, and conflict handling (Xu et al., 2025; Ji et al., 2026; Asai et al., 2024; Jiang et al., 2023; Chhikara et al., 2025; Rasmussen et al., 2025). These advances improve *what* is stored and *how* memory is accessed, but they do not by themselves resolve *when* a memory decision should become irreversible. Relevance, version validity, and evidence sufficiency may depend on information that becomes available only after the current query is known and retrieval has begun. This suggests that the timing of memory decisions is itself a first-class design problem.

We introduce **Distributed Evidence Memory (DEM)**, which defers each irreversible decision to a stage where it can exploit the current query and, when necessary, accumulated evidence. For admission, raw-memory shards inspect their original episodes and report query-conditioned evidence to a shared workspace, allowing evidence discovered in one round to determine subsequent retrieval targets. For version selection, a Conflict Directed Acyclic Graph (CDAG) preserves conflicting versions until query-time aggregation rather than resolving them at write time. For retrieval termination, Good-Turing Coverage Estimation (GTC) provides an evidence-based adaptive stopping signal. Together, these mechanisms keep potentially useful evidence available longer and make later decisions depend on evidence already collected.

Our contributions are:

- **An irreversible-evidence-bottleneck perspective.** We identify early irreversible memory decisions as a structural source of evidence loss and characterize three failure points: premature admission, premature version selection, and premature stopping.
- **Distributed Evidence Memory.** We introduce DEM, which addresses these failure points through distributed evidence access over raw-memory shards, conflict-preserving CDAG, and coverage-aware adaptive stopping with GTC.
- **Mechanistic and empirical validation.** Across three benchmarks, DEM consistently improves strong baselines, with robustness across seven backbones on FactCon and LoCoMo. Controlled analyses attribute most of the gain to evidence coverage and further verify conditional later-hop retrieval and conflict preservation.

2 Related Work

DEM is most closely related to long-term memory organization, iterative or adaptive retrieval, and conflict-aware memory updating; its distinction concerns where query-dependent memory decisions are made.

Long-term memory representation and organization. A growing body of work enables LLM agents to retain information across long interaction histories using memory streams, embeddings, summaries, extracted facts, structured notes, or temporal and relational graphs (Park et al., 2023;

Packer et al., 2023; Zhong et al., 2024; Chhikara et al., 2025; Xu et al., 2025; Rasmussen et al., 2025; Rezazadeh et al., 2025; Kang et al., 2025; Latimer et al., 2026; Wang et al., 2026). These approaches improve storage efficiency, memory organization, and retrieval quality. Many such systems retrieve over representations constructed before the future query is known (Zhong et al., 2024; Chhikara et al., 2025; Xu et al., 2025; Rasmussen et al., 2025). Even methods that reconstruct richer context after retrieval first select memory through an earlier retrieval stage (Wang et al., 2026). DEM is not primarily another memory representation. Instead, it keeps original episodic content accessible to local memory shards and evaluates relevance against the current query at read time.

Iterative and adaptive retrieval. Dense and late-interaction retrievers improve query–document matching through finer-grained representations (Karpukhin et al., 2020; Khattab & Zaharia, 2020). More closely related to DEM, MRAgent iteratively adapts memory access using evidence accumulated during reasoning (Ji et al., 2026). Adaptive retrieval methods such as Self-RAG and FLARE instead decide when or what to retrieve during generation using reflection or model-confidence signals (Asai et al., 2024; Jiang et al., 2023). DEM does not claim iterative retrieval or query-conditioned matching itself as a contribution. Its structural difference is that query-relevant evidence is collected from raw episodic shards into a shared workspace, and evidence discovered in one round explicitly determines the retrieval targets of subsequent rounds. DEM also bases retrieval termination on statistics of the collected evidence, rather than on reader confidence or self-reflection.

Conflict-aware memory updating. Existing memory systems may update, replace, delete, or temporally invalidate inconsistent and evolving information (Chhikara et al., 2025; Rasmussen et al., 2025). Such strategies are appropriate when the objective is to maintain a compact estimate of the current state. DEM targets settings in which a future query may instead require an older value, a newer value, or the transition between them. It therefore keeps conflicting versions available in a Conflict Directed Acyclic Graph (CDAG) and defers query-specific version selection until aggregation over the current query and accumulated evidence.

3 METHOD

We present **Distributed Evidence Memory (DEM)**, whose overall framework is shown in Figure 2. DEM defers each irreversible memory decision to a stage where it can exploit the current query and, when necessary, accumulated evidence.

The framework is organized into two coupled functional stages. **Stage 1** performs distributed evidence access: raw-memory shards are screened against the current query, activated shards report evidence to a shared workspace, and accumulated evidence generates focused next-hop retrieval targets. **Stage 2** maintains the evidence state during this iterative process: CDAG preserves conflicting versions as new evidence arrives, while GTC provides an evidence-based adaptive stopping signal. The accumulated evidence and conflict structure are finally aggregated to produce the response.

3.1 CENTRAL GATING

Let $E = \{E_i\}_{i=1}^{N}$ denote the long-term memory and T the current task. For each episode, let $R_i = R(T, E_i) \in \{0, 1\}$ indicate whether E_i contains evidence relevant to the correct answer.

A common memory-access pattern first constructs a query-independent representation of each episode and then performs retrieval or admission over these representations:

$$Z_i = g(E_i), \qquad S = \pi(T, Z_{1:N}), \qquad \hat{y} = f(T, \{\phi(E_i) : i \in S\}). \tag{1}$$

Here, g denotes the query-independent memory representation, π the corresponding retrieval or admission rule, and $\phi(E_i)$ the content exposed to the downstream reader after admission. We refer to this representation-level selection pattern as *central gating*.

Because Z_i is constructed from E_i, the data-processing inequality (Cover & Thomas, 2006) gives

$$I(R_i; Z_i \mid T) \le I(R_i; E_i \mid T). \tag{2}$$

Equation 2 characterizes the information constraint induced by representation-level gating; it does not imply that compression is harmful in every setting. When Z_i is not sufficient for relevance under

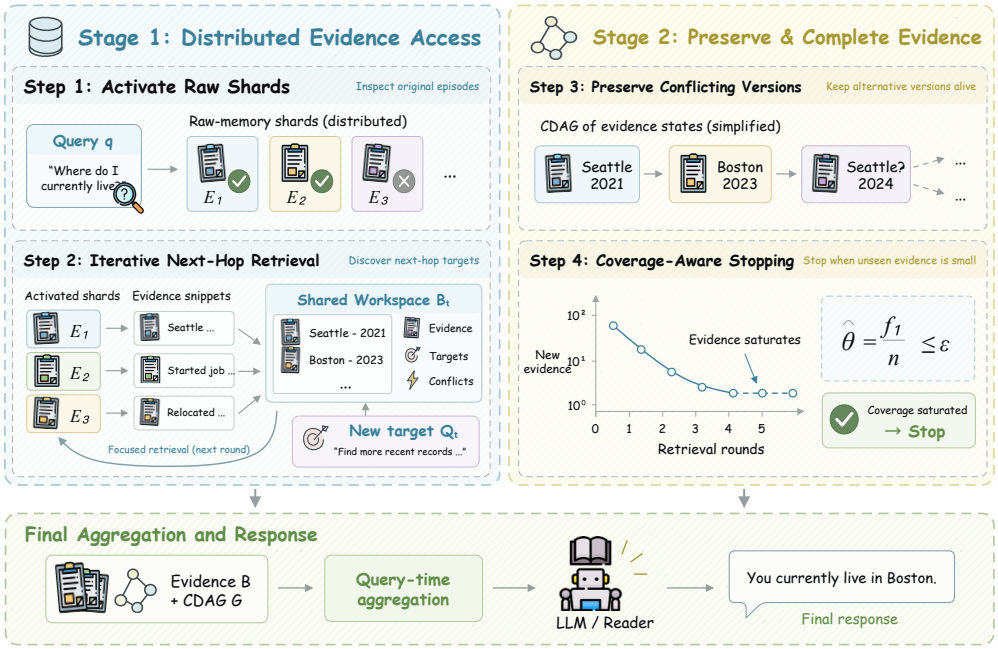


Figure 2: **Overall framework of Distributed Evidence Memory (DEM).** Stage 1 performs distributed evidence access: raw-memory shards inspect their original episodes, report query-conditioned evidence, and iteratively retrieve next-hop evidence through a shared workspace. Stage 2 maintains conflicting versions with CDAG and evaluates the GTC stopping signal as evidence accumulates. The accumulated evidence and conflict structure are then aggregated to produce the final response.

the current task, episodes that require different access decisions may become indistinguishable to a gate operating only on Z_i. A formal indistinguishability result is given in Appendix D.1.

DEM replaces representation-level centralized admission with query-time screening over original episodic content, as described next.

3.2 Distributed Evidence Access

Stage 1 of Figure 2 consists of shard activation followed by iterative evidence retrieval.

Step 1: Activate raw-memory shards. When a query arrives, each shard evaluates the current task against its own original content:

$$b_i = \beta(T, E_i), \qquad A = \{i : b_i \geq \tau_{\text{bid}}\}, \tag{3}$$

where A is the set of activated shards. Unlike a global top-k competition over compressed representations, each shard performs its local relevance judgment directly against the episodic content.

The activated set A remains fixed for the remainder of the current query. Only shards in A undergo the more expensive evidence-extraction stage. Let B_t denote the shared evidence workspace after retrieval round t. The initial workspace is

$$B_0 = \bigcup_{i \in A} \text{Extract}(T, E_i). \tag{4}$$

This separates lightweight screening over all shards from deeper processing over the activated subset.

Step 2: Iterative next-hop retrieval. The shared workspace allows evidence discovered in one round to determine what is searched for next. At round t, the current task, accumulated evidence, and conflict structure generate focused next-hop targets:

$$Q_t = h(T, B_t, G_t), \tag{5}$$

where G_t denotes the current conflict graph.

Each activated shard then searches a focused view of its original content conditioned on Q_t:

$$B_{t+1} = B_t \cup \bigcup_{i \in A} \mathrm{Extract}(T, \mathrm{Focus}(E_i, Q_t)) \,. \tag{6}$$

Thus, shard activation is performed once, while subsequent rounds repeatedly search the activated raw-memory shards using evidence-derived targets. This design is intended to improve conditional later-hop retrieval; we test this mechanism directly in Section 4.3.

3.3 Conflict-Preserving Memory with CDAG

DEM retains conflicting versions rather than globally invalidating one version at write time. A future query may require an older value, a newer value, or the transition between them, so query-specific version selection is deferred until evidence has been accumulated.

Each extracted evidence item retains its source provenance and is converted into one or more structured facts $v = (s, r, o, \ell)$, where s, r, o, and ℓ denote the subject, relation, object, and temporal or interaction index. Facts sharing the same (s, r) pair represent alternative versions of the same relation. Two versions $v_a = (s, r, o_a, \ell_a)$ and $v_b = (s, r, o_b, \ell_b)$ with $o_a \neq o_b$ are treated as conflicting. For $\ell_a < \ell_b$, CDAG records their temporal relation as

$$v_a \xrightarrow{d_{a,b}} v_b. \tag{7}$$

Because edges follow temporal order, the resulting graph is acyclic. The edge label $d_{a,b}$ records the observed transition without declaring either endpoint globally invalid, so both versions remain available to later retrieval and reasoning. The graph is updated as new evidence enters the shared workspace.

Only after evidence accumulation does DEM perform query-specific selection, $v^\star = \mathrm{Select}(T, B, G)$. CDAG therefore separates *recording that a change occurred* from *deciding which version answers the current query*. A formal analysis of write-time single-version selection is provided in Appendix D.2.

3.4 Coverage-Aware Stopping with GTC

The remaining decision is when iterative evidence collection should stop. Rather than relying only on a fixed retrieval depth or model-confidence signal, GTC estimates whether substantial unseen evidence is likely to remain from evidence rediscovery statistics.

Each extracted evidence item is assigned a canonical record/fact ID. Duplicate hits within a round are deduplicated, while repeated hits across rounds are treated as rediscoveries. Let n be the total number of canonical evidence observations and f_1 the number of evidence identities observed in exactly one retrieval round. Following the Good-Turing unseen-mass intuition (Good, 1953), DEM estimates

$$\hat{\theta} = \frac{f_1}{n}. \tag{8}$$

Retrieval satisfies the GTC stopping condition when $\hat{\theta} \leq \epsilon_{\mathrm{gtc}}$, with $\epsilon_{\mathrm{gtc}} = 0.08$. The full procedure terminates when no further next-hop target is generated, the GTC condition is met, or the configured round limit is reached. GTC therefore provides an evidence-level adaptive stopping signal rather than an estimate of answer confidence.

3.5 Aggregation

After retrieval terminates, the accumulated workspace B and conflict graph G are passed to a query-time aggregator that performs reasoning and version selection to produce $\hat{y}$. Shard-level operations

within a round are parallelizable, whereas retrieval rounds remain sequential because later targets depend on previously collected evidence. The complete procedure and cost decomposition are provided in Appendix Algorithm 1 and Appendix E.4; empirical compute is reported in Section 4.4.

4 EXPERIMENTS

We evaluate DEM on overall performance, mechanism behavior, and robustness/efficiency through three-benchmark comparisons, controlled ablations, and backbone-, memory-size-, and compute-level analyses.

4.1 EXPERIMENTAL SETUP

Benchmarks and metrics. We evaluate on three complementary long-term memory benchmarks. **FactCon**, the conflict-resolution track of MemoryAgentBench, contains 100 single-hop and 100 multi-hop questions; we use Substring Exact Match (SubEM) as the primary metric (Hu et al., 2026). **LoCoMo** evaluates conversational memory across multi-hop, temporal, open-domain, and single-hop questions, for which we report F1 over Categories 1–4 (Maharana et al., 2024). **LongMemEval** evaluates long-range memory under single-hop, multi-hop, temporal, knowledge-update, and abstention settings; we report LLM-as-a-judge accuracy on both LongMemEval-S and the MemoryAgentBench reformulation LME(S*) (Wu et al., 2025; Hu et al., 2026).

Baselines and backbones. We compare against Full-Context, BM25, Naive RAG, and representative long-term memory systems including MemGPT, Mem0, A-Mem, Zep, MemoryBank, and E-mem (Robertson & Zaragoza, 2009; Lewis et al., 2020; Packer et al., 2023; Chhikara et al., 2025; Xu et al., 2025; Rasmussen et al., 2025; Zhong et al., 2024; Wang et al., 2026). The cross-system comparison uses GPT-4o-mini. Controlled mechanism analyses use DeepSeek-v4-flash, with E-mem as the primary memory-system reference. We additionally compare E-mem and DEM across seven backbones on FactCon and LoCoMo: DeepSeek-v4-flash, Qwen3.5-4B, Qwen3.5-9B, Qwen3.5-27B, MiniMax-M3, Claude 5 Sonnet, and GPT-5.5.

DEM configuration. Unless otherwise stated, DEM uses approximately 4k-token raw-memory shards, a shard bid threshold of $\tau_{\text{bid}} = 0.55$, and a GTC threshold of $\epsilon_{\text{gtc}} = 0.08$. All ablations modify only the target mechanism while keeping the remaining configuration fixed. Additional implementation details, retrieval limits, and evaluation protocols are provided in Appendix E.

4.2 MAIN RESULTS

Table 1 shows that DEM consistently outperforms the strongest reported memory baseline across FactCon, LoCoMo, and both LongMemEval settings under GPT-4o-mini. The same pattern holds in the controlled DeepSeek-v4-flash evaluation (Appendix Table 3), where DEM improves FactCon SubEM from 41.5 with E-mem to 64.0. Improvements also persist on multi-hop questions, motivating the controlled mechanism analysis below.

4.3 ABLATION AND MECHANISM ANALYSIS

The ablations in Table 2 confirm that all three design choices contribute. Removing focused next-hop retrieval or conflict preservation causes the largest degradations, while replacing GTC with stopping at the first zero-new-evidence round yields a smaller but consistent drop. Complete ablations across both DeepSeek-v4-flash and Qwen3.5-9B are reported in Appendix Table 8.

Table 2: FactCon ablation with DeepSeek-v4-flash. Parentheses show ΔSubEM from full DEM.

Configuration	SubEM ↑	SH	MH
Full DEM	**64.0**	92.0	36.0
Distributed retrieval			
w/o focused next-hop search	53.5 (-10.5)	87.0	20.0
Conflict preservation			
write-time LLM resolution	55.8 (-8.2)	85.0	26.5
last-write-wins	50.8 (-13.2)	79.0	22.5
Coverage-aware stopping			
w/o GTC	61.5 (-2.5)	91.0	32.0

Table 1: Main results under GPT-4o-mini. FactCon reports SubEM, LoCoMo reports F1, and LongMemEval reports LLM-as-a-judge accuracy; SH/MH denote single-/multi-hop performance. Best overall results are **bold**; strongest baselines are underlined.

	FactCon			LoCoMo			LME-S	LME(S*)
Method	**SubEM ↑**	SH	MH	**F1 ↑**	SH	MH	**Acc. ↑**	**Acc. ↑**
Long context								
Full-Context	25.0	45.0	5.0	37.31	46.68	29.23	55.4	30.7
Classical retrieval								
BM25	25.5	48.0	3.0	38.41	48.73	21.86	50.6	45.3
Naive RAG	15.5	28.0	3.0	44.73	52.45	27.50	53.8	48.3
Memory systems								
MemGPT	15.5	28.0	3.0	–	41.04	26.65	–	32.0
Mem0	10.0	18.0	2.0	45.10	47.65	38.72	–	36.0
A-Mem	–	–	–	39.65	44.65	27.02	–	–
Zep	5.0	7.0	3.0	–	35.74	19.37	63.8	38.3
MemoryBank	–	–	–	–	6.61	5.00	–	–
E-mem	29.0	52.0	6.0	54.17	59.23	42.64	65.3	52.6
Ours								
DEM	**53.5**	84.0	23.0	**58.82**	62.23	53.81	**69.2**	**57.7**

Evidence coverage explains most of the improvement. Figure 3(a) decomposes the 22.5-point E-mem-to-DEM gain. A Shapley decomposition (Shapley, 1953) attributes 15.4 points (68%) to increased evidence coverage; the remainder comes primarily from better use of fully covered evidence. Complete conditional accuracies and attribution results are in Appendix Tables 9 and 10.

The relative retrieval gain is larger at the second hop. For two-hop questions, let $r_2 = P(\text{retrieve hop 2} \mid \text{hop 1 retrieved})$ denote conditional second-hop recall. Figure 3(b) shows that E-mem obtains $r_1 = 0.75$ and $r_2 = 0.24$, whereas DEM obtains $r_1 = 0.96$ and $r_2 = 0.47$. The corresponding relative gains are $1.28\times$ at the first hop and $1.96\times$ at the conditional second hop. The same later-hop pattern persists across all seven backbones and also appears on LoCoMo (Appendix Figure 5 and Table 11). This supports the intended role of the shared workspace: previously retrieved evidence changes what the system searches for next.

Conflict preservation reduces post-retrieval evidence loss. Figure 3(c) holds retrieval coverage fixed while varying only the conflict policy. All three policies achieve identical single-hop/multi-hop evidence coverage of $96\%/45\%$. Nevertheless, multi-hop accuracy conditioned on complete coverage decreases from 67.8 with CDAG to 47.0 with write-time LLM resolution and 38.0 with last-write-wins. The loss therefore occurs after the relevant evidence has already been retrieved, supporting the decision to preserve conflicting versions until query-time aggregation. Complete controlled results are reported in Appendix Table 14.

Figure 3 isolates the distributed-access and conflict-preservation mechanisms. The contribution of GTC is characterized by the stopping ablation above and the accuracy–compute analysis in Section 4.4.

4.4 Robustness and Efficiency

Robustness across backbones. Across seven backbones, DEM improves over E-mem by 19.5–29.0 FactCon SubEM points, with the same direction on LoCoMo. Full results, including the three-seed DeepSeek-v4-flash evaluation, are in Appendix Tables 4 and 6.

Robustness to memory size. As N grows from 8 to 512, E-mem first-hop recall falls from 96.0% to 34.0%, whereas DEM remains at 92.5–94.0%; the average deeply processed shards increase only from $K = 2.4$ to 11.8 (Figure 4(a)). Because DEM still screens all N shards, this measures retrieval robustness rather than sublinear total cost. Complete results are in Appendix Table 12.

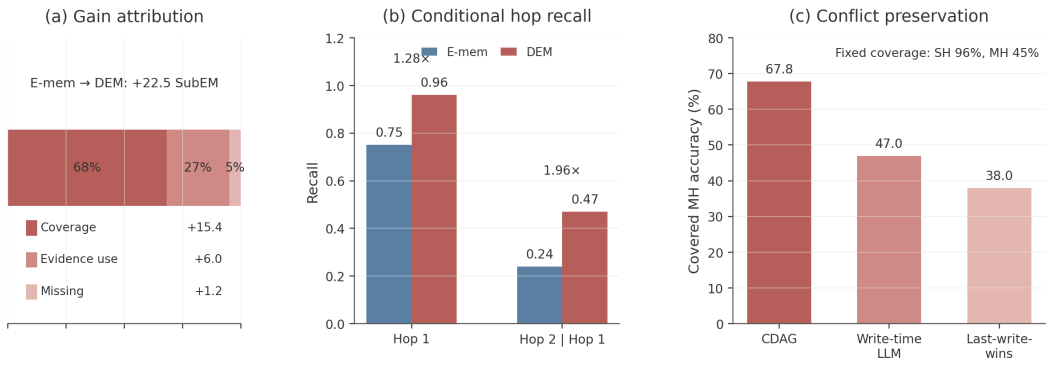


Figure 3: **Mechanism analysis on FactCon with DeepSeek-v4-flash.** (a) Improved evidence coverage explains most of the E-mem-to-DEM gain. (b) DEM yields a larger relative improvement in conditional second-hop recall than in first-hop recall. (c) Under identical retrieval coverage, write-time conflict resolution substantially reduces multi-hop accuracy, showing the benefit of preserving conflicting versions until query-time aggregation.

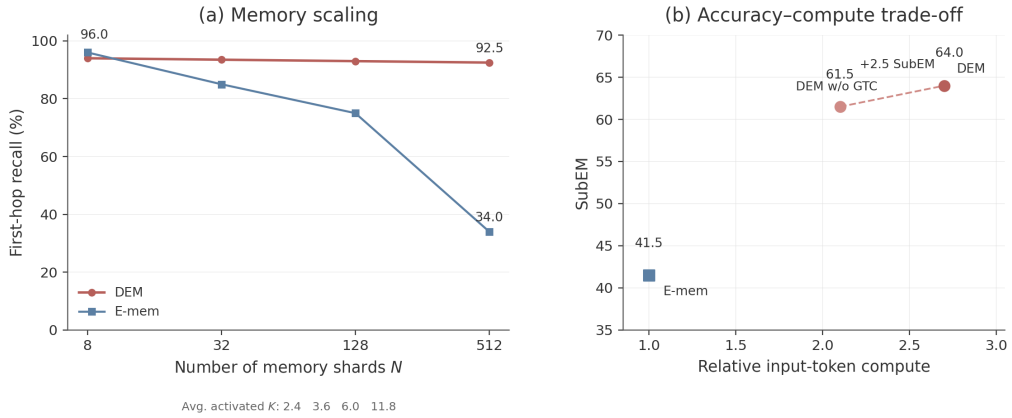


Figure 4: **Memory-size robustness and efficiency analysis.** (a) DEM maintains stable first-hop recall as the number of memory shards increases, whereas E-mem degrades sharply under a fixed retrieval budget. (b) DEM trades additional input-token compute for higher accuracy, while GTC provides further gains beyond the cheaper zero-new-evidence stopping variant.

Accuracy–compute trade-off. With DeepSeek-v4-flash, DEM uses 14.6M input tokens (2.7× E-mem's 5.4M), while zero-new-evidence stopping uses 11.3M (2.1×) but reduces SubEM from 64.0 to 61.5 (Figure 4(b)). Thus, GTC trades additional retrieval for accuracy. Detailed statistics and accounting are in Appendix Table 13 and Appendix E.4.

5 Conclusion

We introduced **Distributed Evidence Memory (DEM)** around a simple design principle: irreversible memory decisions should be deferred until they can exploit the current query and, when necessary, accumulated evidence. DEM combines query-time evidence access over raw-memory shards, conflict-preserving CDAG, and evidence-based adaptive stopping with GTC. Across FactCon, LoCoMo, and LongMemEval, DEM consistently improves strong retrieval and memory baselines. Controlled analyses attribute most of the improvement to higher evidence coverage and further show stronger conditional later-hop retrieval and reduced post-retrieval loss under conflicting versions. These gains require additional query-time computation, revealing an explicit accuracy–

efficiency trade-off. More broadly, our results support preserving potentially relevant and conflicting evidence and delaying irreversible commitments until their relevance, validity, or sufficiency can be assessed using query-conditioned evidence.

AI USE STATEMENT

Generative AI tools were used during manuscript preparation to assist with language polishing, structural editing, LaTeX formatting, and bibliography organization. They were not treated as authoritative sources. All technical claims, reported experimental results, citations, and final manuscript content were reviewed and verified by the authors, who take full responsibility for the paper.

REPRODUCIBILITY STATEMENT

We provide complete benchmark-level results, full ablations, mechanism analyses, formal analyses, and the end-to-end retrieval procedure in the appendix. Appendix E documents the principal DEM settings used in the reported experiments, including shard size and overlap, shard-bid threshold, focused-retrieval limit, GTC threshold, retrieval-round limits, evaluation protocols, and compute-accounting conventions. We also report model-call counts, token usage, and retrieval-depth statistics for the controlled comparisons.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. In *ECAI 2025 - 28th European Conference on Artificial Intelligence*, volume 413 of *Frontiers in Artificial Intelligence and Applications*, pp. 2993–3000. IOS Press, 2025. doi: 10.3233/FAIA251160.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2 edition, 2006. doi: 10.1002/047174882X.

I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264, 1953. doi: 10.1093/biomet/40.3-4.237.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in LLM agents via incremental multi-turn interactions. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=DT7JyQC3MR.

Shuo Ji, Yibo Li, and Bryan Hooi. Memory is reconstructed, not retrieved: Graph memory for LLM agents. *CoRR*, abs/2606.06036, 2026. doi: 10.48550/arXiv.2606.06036.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7969–7992. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.495.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. Memory OS of AI agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25961–25970. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1318.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 6769–6781. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.550.

Omar Khattab and Matei Zaharia. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 39–48. Association for Computing Machinery, 2020. doi: 10.1145/3397271.3401075.

Christopher Latimer, Nicolò Boschi, Andrew Neeser, Chris Bartholomew, Gaurav Srivastava, Xuan Wang, and Naren Ramakrishnan. Hindsight: Structured agent memory that retains, recalls, and reflects. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 275–285. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-demo.27.

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*, volume 33, pp. 9459–9474, 2020.

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

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *CoRR*, abs/2310.08560, 2023. doi: 10.48550/arXiv.2310.08560.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22. Association for Computing Machinery, 2023. doi: 10.1145/3586183.3606763.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: A temporal knowledge graph architecture for agent memory. *CoRR*, abs/2501.13956, 2025. doi: 10.48550/arXiv.2501.13956.

Alireza Rezazadeh, Zichao Li, Wei Wei, and Yujia Bao. From isolated conversations to hierarchical schemas: Dynamic tree memory representation for LLMs. In *International Conference on Learning Representations*, 2025.

Stephen E. Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019.

Lloyd S. Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games II*, volume 28 of *Annals of Mathematics Studies*, pp. 307–317. Princeton University Press, 1953.

Kaixiang Wang, Yidan Lin, Jiong Lou, Zhaojiacheng Zhou, Bunyod Suvonov, and Jie Li. E-mem: Multi-agent based episodic context reconstruction for LLM agent memory. *CoRR*, abs/2601.21714, 2026. doi: 10.48550/arXiv.2601.21714.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations*, 2025.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-Mem: Agentic memory for LLM agents. In *Advances in Neural Information Processing Systems*, volume 38, pp. 20004–20031, 2025. doi: 10.52202/085713-0593.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. MemoryBank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19724–19731, 2024. doi: 10.1609/aaai.v38i17.29946.

A Additional Benchmark Results

This appendix provides the complete numerical results omitted from the main paper for space. We first report additional benchmark-level comparisons and cross-backbone results, followed by full ablations and detailed mechanism, robustness, and efficiency analyses.

A.1 FactCon Results

We report the controlled DeepSeek-v4-flash comparison and the complete cross-backbone robustness results on FactCon.

Table 3: Controlled FactCon results with DeepSeek-v4-flash. All methods use the same local evaluation pipeline. SubEM is the primary metric; SH/MH denote single-/multi-hop SubEM.

Group	**Method**	**SubEM ↑**	EM	F1	SH	MH
Long context	Full-Context	48.0	41.0	44.0	82.0	14.0
Classical retrieval	BM25	36.0	31.0	33.0	64.0	8.0
	Naive RAG	27.5	24.0	25.5	48.0	7.0
Memory system	E-mem	41.5	35.5	38.5	72.0	11.0
Ours	**DEM**	**64.0**	57.0	59.5	92.0	36.0

Table 4: FactCon robustness across seven backbones. E-mem and DEM are evaluated under the same pipeline. SubEM is the primary metric; DeepSeek-v4-flash 3-seed results are reported as mean ± standard deviation.

Backbone	**Method**	**SubEM ↑**	EM	F1	SH	MH
Qwen3.5-4B	E-mem	33.0	31.5	32.7	60.0	6.0
	DEM	**62.0**	24.0	36.4	98.0	26.0
Qwen3.5-9B	E-mem	37.0	31.0	33.6	71.0	3.0
	DEM	**63.5**	10.0	25.0	97.0	30.0
Qwen3.5-27B	E-mem	40.0	34.0	36.0	72.0	8.0
	DEM	**65.0**	22.0	38.0	97.0	33.0
DeepSeek-v4-flash	E-mem	41.5	35.5	38.5	72.0	11.0
	DEM	**64.0**	57.0	59.5	92.0	36.0
DeepSeek-v4-flash (3 seeds)	E-mem	41.2 ± 0.8	35.2 ± 0.9	38.2 ± 0.8	71.7 ± 1.2	10.7 ± 1.5
	DEM	$\mathbf{63.7 \pm 1.1}$	56.6 ± 1.4	59.1 ± 1.2	91.7 ± 1.0	35.7 ± 2.0
MiniMax-M3	E-mem	39.5	34.0	36.3	72.0	7.0
	DEM	**59.0**	47.0	49.8	89.0	29.0
Claude 5 Sonnet	E-mem	43.0	40.0	41.5	74.0	12.0
	DEM	**66.0**	62.0	63.5	94.0	38.0
GPT-5.5	E-mem	44.0	41.0	42.0	75.0	13.0
	DEM	**67.0**	63.0	64.5	94.0	40.0

A.2 LoCoMo Results

We report category-level results under GPT-4o-mini and the corresponding cross-backbone robustness results on LoCoMo.

Table 5: LoCoMo category-level results under GPT-4o-mini. Scores are F1 (%); Overall is the weighted average over Categories 1–4, excluding the adversarial category. A dash indicates an unreported result.

Group	**Method**	**Overall ↑**	**Category F1 (%)**				
			SH	MH	Temporal	Open-dom.	Adv.
Long context	Full-Context	37.31	46.68	29.23	25.97	16.87	–
Classical retrieval	BM25	38.41	48.73	21.86	–	–	–
	Naive RAG	44.73	52.45	27.50	46.07	23.23	–
Memory systems	MemGPT	–	41.04	26.65	25.52	9.15	43.29
	Mem0	45.10	47.65	38.72	48.93	28.64	–
	A-Mem	39.65	44.65	27.02	45.85	12.14	50.03
	Zep	–	35.74	19.37	42.00	49.56	–
	MemoryBank	–	6.61	5.00	9.68	5.56	7.36
	E-mem	54.17	59.23	42.64	59.82	24.89	–
Ours	**DEM**	**58.82**	62.23	53.81	63.54	27.88	52.4

Table 6: LoCoMo cross-backbone category-level results. Scores are F1 (%); Overall is the weighted average over Categories 1–4.

Backbone	**Method**	**Overall ↑**	**Category F1 (%)**				
			SH	MH	Temporal	Open-dom.	Adv.
Qwen3.5-4B	E-mem	48.6	54.0	36.0	54.0	20.0	42.0
	DEM	**55.6**	60.0	48.0	60.0	24.0	48.0
Qwen3.5-9B	E-mem	52.6	58.0	40.5	58.0	23.5	45.0
	DEM	**57.8**	61.5	52.0	62.5	26.5	50.0
Qwen3.5-27B	E-mem	54.7	59.8	43.5	60.0	25.5	47.0
	DEM	**59.2**	62.5	54.5	63.8	28.0	52.0
DeepSeek-v4-flash	E-mem	55.3	60.2	44.5	60.5	26.0	48.0
	DEM	**59.7**	63.0	55.5	64.2	28.5	53.0
MiniMax-M3	E-mem	54.3	59.5	43.0	59.5	25.0	46.0
	DEM	**57.1**	60.8	51.5	61.5	26.5	50.0
Claude 5 Sonnet	E-mem	56.3	61.0	46.0	61.5	27.5	50.0
	DEM	**60.4**	63.5	56.5	65.0	29.5	54.0
GPT-5.5	E-mem	56.9	61.5	47.0	62.0	28.0	51.0
	DEM	**61.0**	64.0	57.5	65.5	30.0	55.0

A.3 LongMemEval-S Results

We report the available category-level LongMemEval-S results; methods without reliable category-level values are omitted from the breakdown.

Table 7: LongMemEval-S category-level results under GPT-4o-mini. Scores are LLM-as-a-judge accuracy (%). Methods without reliable category-level results are omitted; their overall scores are reported in Table 1.

Method	**Overall ↑**	**Category accuracy (%)**				
		SH	MH	Temporal	Knowledge update	Abstention
Full-Context	55.4	81.4	40.6	36.5	76.9	–
E-mem	65.3	90.1	46.2	48.5	76.4	40.0
DEM	**69.2**	93.2	52.6	56.8	80.5	46.7

B Full Ablation Results

We provide the complete FactCon ablations for Qwen3.5-9B and DeepSeek-v4-flash, including both performance and retrieval statistics.

Table 8: Full FactCon ablation results across Qwen3.5-9B and DeepSeek-v4-flash. Values in parentheses denote the SubEM change relative to full DEM. "w/o GTC" replaces GTC with stopping at the first zero-new-evidence round. A dash indicates an unreported auxiliary statistic.

Configuration	**Performance**			**Retrieval statistics**	
	SubEM ↑	SH	MH	Avg. rounds	Avg. evidence
Qwen3.5-9B					
Full DEM	**63.5**	97.0	30.0	3.1	8.2
w/o focused next-hop search	52.5				
,(-11.0)	91.0	14.0	–	–	
write-time LLM resolution	56.0				
,(-7.5)	90.0	22.0	2.8	7.2	
last-write-wins	50.5				
,(-13.0)	83.0	18.0	2.6	6.6	
w/o GTC	60.5				
,(-3.0)	96.0	25.0	2.3	6.1	
DeepSeek-v4-flash					
Full DEM	**64.0**	92.0	36.0	2.6	7.8
w/o focused next-hop search	53.5				
,(-10.5)	87.0	20.0	–	–	
write-time LLM resolution	55.8				
,(-8.2)	85.0	26.5	2.3	6.8	
last-write-wins	50.8				
,(-13.2)	79.0	22.5	2.2	6.3	
w/o GTC	61.5				
,(-2.5)	91.0	32.0	2.1	6.4	

C Mechanism, Robustness, and Efficiency Analyses

We next provide the detailed analyses underlying the mechanism, memory-size robustness, and efficiency results summarized in the main paper.

C.1 Coverage and Accuracy Decomposition

We first separate gains due to complete evidence coverage from gains due to better use of retrieved evidence. Table 9 reports conditional accuracies, while Table 10 gives the corresponding Shapley attribution.

Table 9: Evidence coverage, conditional accuracy, and gold-evidence bypass on FactCon. c_{SH} and c_{MH} denote complete evidence coverage; a_{cov} denotes accuracy conditioned on complete coverage, and a_{miss} denotes accuracy when some required evidence is missing. Gold bypass directly supplies the gold evidence to the downstream reader.

Backbone	**Method**	**Evidence coverage (%)**		**Conditional accuracy (%)**			**Gold bypass SubEM**
		c_{SH}	c_{MH}	$a_{\mathrm{cov}}^{\mathrm{SH}}$	$a_{\mathrm{cov}}^{\mathrm{MH}}$	a_{miss}	
DeepSeek-v4-flash	E-mem	75	18	94.0	33.8	6.0	65.5
	DEM	96	45	95.4	67.8	10.0	76.9
Qwen3.5-9B	E-mem	74	8	94.7	–	3.5	64.0
	DEM	99	40	97.9	67.5	5.0	77.7

Table 10: Shapley decomposition of the DeepSeek-v4-flash improvement from E-mem to DEM on FactCon. Values are attributed SubEM gains in points; Share denotes each source's contribution to the overall gain. Overall attributions are rounded independently and may not sum exactly to the total improvement.

Source	**Attributed SubEM gain (points)**			**Share**
	SH	MH	Overall	
Coverage c	+18.6	+12.1	+15.4	68%
Covered accuracy a_{cov}	+1.2	+10.7	+6.0	27%
Missed-evidence accuracy a_{miss}	+0.2	+2.2	+1.2	5%

C.2 Multi-Hop Retrieval

We examine conditional hop recall on FactCon and then verify the same single-hop versus multi-hop pattern on LoCoMo.

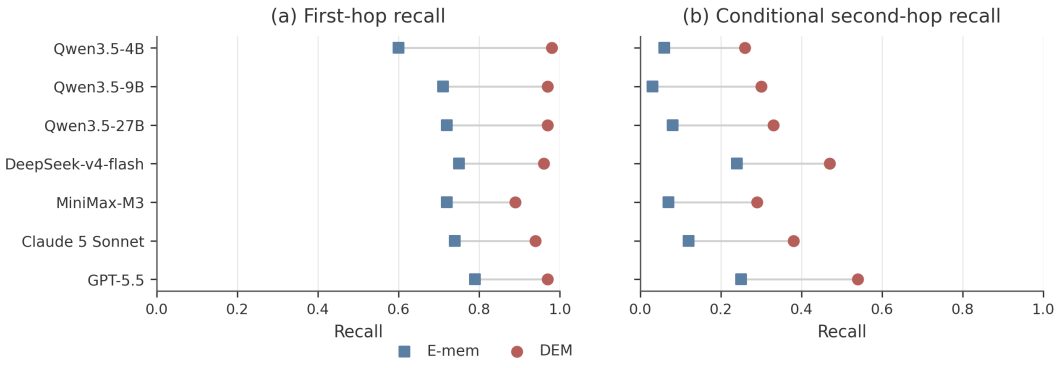


Figure 5: **Cross-backbone conditional hop recall on FactCon.** (a) First-hop recall r_1 and (b) conditional second-hop recall $r_2 = P(\text{retrieve hop 2} \mid \text{hop 1 retrieved})$ for E-mem and DEM across seven backbones. DEM consistently improves both quantities, with larger relative gains in conditional second-hop retrieval.

Table 11: LoCoMo single-hop and multi-hop gains from E-mem to DEM. DEM yields substantially larger improvements on multi-hop questions under both the GPT-4o-mini and controlled DeepSeek-v4-flash settings.

Setting	Single-hop F1 (%)			Multi-hop F1 (%)		
	E-mem	**DEM**	ΔSH	E-mem	**DEM**	ΔMH
GPT-4o-mini	59.23	**62.23**	+3.00	42.64	**53.81**	**+11.17**
DeepSeek-v4-flash	60.2	**63.0**	+2.8	44.5	**55.5**	**+11.0**

C.3 Memory-Size Robustness

We provide the complete numerical results corresponding to the memory-size robustness analysis in the main paper.

Table 12: Memory-size robustness results on a 40-question FactCon slice with DeepSeek-v4-flash. N is the number of memory shards, c denotes overall complete-evidence coverage over the slice, and K is the average number of deeply activated DEM shards. For single-hop questions, complete coverage equals r_1; for two-hop questions, it equals $r_1 r_2$.

N	E-mem				DEM	
	r_1 (%)	r_2 (%)	Coverage c (%)	SubEM	r_1 (%)	Avg. K
8	96.0	31.0	62.9	54.0	94.0	2.4
32	85.0	28.0	54.4	47.5	93.5	3.6
128	75.0	24.0	46.5	41.5	93.0	6.0
512	34.0	12.0	19.0	20.5	92.5	11.8

C.4 Computational Cost

We report the detailed retrieval statistics and input-token cost underlying the accuracy–compute trade-off.

Table 13: Accuracy and computational cost on FactCon. Avg. rounds and Avg. evidence are per-query averages; Calls and Tokens are aggregate counts for each reported configuration. Relative compute is measured by input-token consumption relative to E-mem under the same backbone.

Backbone	Configuration	Performance		Retrieval statistics		Compute		
		SubEM ↑	MH	Avg. rounds	Avg. evidence	Calls	Tokens	Rel. compute
Qwen3.5-9B	E-mem	37.0	3.0	1.0	4.6	1160	4.9M	1.0×
	DEM w/o GTC	60.5	25.0	2.3	6.1	2090	8.8M	1.8×
	DEM	63.5	30.0	3.1	8.2	2780	11.8M	2.4×
DeepSeek-v4-flash	E-mem	41.5	11.0	1.0	4.8	1180	5.4M	1.0×
	DEM w/o GTC	61.5	32.0	2.1	6.4	2480	11.3M	2.1×
	DEM	64.0	36.0	2.6	7.8	3190	14.6M	2.7×

C.5 Conflict Preservation under Fixed Coverage

We control retrieval coverage within each backbone and vary only the conflict policy, isolating post-retrieval evidence loss.

Table 14: Controlled conflict-retention analysis on FactCon. All policies use identical retrieval coverage within each backbone, isolating the effect of conflict handling after evidence retrieval. a_{cov} denotes accuracy conditioned on complete evidence coverage.

Conflict policy	Fixed coverage (%) SH / MH	Conditional accuracy (%)		Performance		
		$a_{\text{cov}}^{\text{SH}}$	$a_{\text{cov}}^{\text{MH}}$	SubEM	SH	MH
Qwen3.5-9B						
CDAG, retain all	99 / 40	97.9	67.5	63.5	97.0	30.0
Write-time LLM	99 / 40	90.5	47.0	56.0	90.0	22.0
Last-write-wins	99 / 40	84.0	38.0	50.5	83.0	18.0
DeepSeek-v4-flash						
CDAG, retain all	96 / 45	95.4	67.8	64.0	92.0	36.0
Write-time LLM	96 / 45	88.0	47.0	55.8	85.0	26.5
Last-write-wins	96 / 45	82.0	38.0	50.8	79.0	22.5

D Formal Analysis

This section formalizes two structural failure modes discussed in Section 3: indistinguishability under compressed central gating and irreversible information loss under write-time single-version conflict resolution. These results characterize failure modes that DEM is designed to avoid; they do not imply that compression or conflict resolution is necessarily harmful in every setting.

D.1 Indistinguishability under Central Gating

Recall the central-gating formulation in Section 3.1. Each episode E_i is first mapped to a query-independent representation

$$Z_i = g(E_i),$$

and the admission rule decides whether the episode becomes available to the downstream reader using the current task T and the compressed representation Z_i.

Equation 2 gives

$$I(R_i; Z_i \mid T) \leq I(R_i; E_i \mid T),$$

where $R_i = R(T, E_i)$ indicates whether episode E_i contains evidence relevant to the current task. Equality requires Z_i to be sufficient for predicting R_i given T. When this condition fails, distinct episodes may be indistinguishable to the admission rule even though their relevance to the current task differs.

Proposition 1 (Indistinguishability under compressed gating). Let $Z = g(E)$ be a query-independent representation of an episode, and let $\pi(T, Z)$ be a deterministic admission rule that observes the current task T and the representation Z, but not the original episode E. Suppose there exist two episodes E^+ and E^- such that

$$g(E^+) = g(E^-),$$

while

$$R(T, E^+) = 1, \qquad R(T, E^-) = 0.$$

Then π cannot simultaneously admit E^+ and reject E^-.

Proof. Let

$$Z = g(E^+) = g(E^-).$$

Because the admission rule observes only (T, Z), its inputs are identical for the two episodes. Therefore,

$$\pi(T, g(E^+)) = \pi(T, g(E^-)).$$

Hence the rule must make the same admission decision for both episodes. It therefore cannot deterministically admit the relevant episode E^+ while rejecting the irrelevant episode E^-. □

Discussion. The proposition does not state that memory compression is inherently lossy for every future task. If Z is sufficient for predicting relevance under T, compressed gating can preserve the information required for the admission decision. The failure arises when two episodes that require different decisions under the current task collapse to the same query-independent representation. Once a necessary episode is excluded, the downstream reader cannot directly inspect the missing evidence. DEM avoids this particular bottleneck by moving relevance judgment back to the original shard content, as defined in Equation 3.

D.2 Irreversibility of Write-Time Conflict Resolution

Section 3.3 considers conflicting memory versions

$$v_a = (s, r, o_a, \ell_a), \qquad v_b = (s, r, o_b, \ell_b),$$

with $o_a \neq o_b$ and $\ell_a < \ell_b$. A write-time single-version policy resolves the conflict before the future task is known by permanently retaining only one of the two versions. The following observation formalizes the corresponding information-loss constraint.

Proposition 2 (Irreversibility of write-time single-version selection). Consider two conflicting versions v_a and v_b of the same subject–relation pair. Suppose there exist two possible future tasks T_a and T_b such that answering T_a requires evidence from v_a, whereas answering T_b requires evidence from v_b. Then any query-independent write-time policy that permanently retains only one element of $\{v_a, v_b\}$ necessarily removes evidence required by at least one of the two tasks.

Proof. A single-version policy has only two possible outcomes.

If the policy retains v_a and permanently discards v_b, then the evidence required by T_b is no longer available.

Conversely, if the policy retains v_b and permanently discards v_a, then the evidence required by T_a is no longer available.

Thus, irrespective of which version is selected at write time, at least one possible future task loses access to evidence that it requires. □

Discussion. This result concerns irreversible *single-version* resolution before the future query is known. It does not imply that preserving all versions guarantees correct downstream reasoning. Rather, preserving both versions removes this specific write-time information-loss constraint. CDAG implements this principle by recording the transition between conflicting versions without declaring either version universally invalid. The final version decision is deferred until query-time aggregation, as described in Section 3.3. The controlled results in Appendix C.5 isolate this effect by comparing CDAG with write-time LLM resolution and last-write-wins under fixed retrieval coverage.

E Implementation and Evaluation Details

This section provides additional implementation and evaluation details for the experiments in Section 4. We describe the DEM configuration, iterative retrieval procedure, evaluation protocols, and compute-accounting conventions used throughout the experiments.

E.1 DEM Configuration

Unless otherwise stated, the interaction history is divided into approximately 4k-token raw-memory shards with an overlap ratio of 0.1. The shards retain the original episodic content; no summary construction, fact extraction, or conflict resolution is required before they are stored.

At query time, each shard performs a local relevance check against the current task. A shard is activated when its bid score satisfies

$$b_i \geq \tau_{\text{bid}}, \tag{9}$$

where the bid threshold is

$$\tau_{\text{bid}} = 0.55. \tag{10}$$

Only activated shards proceed to the more expensive evidence-extraction stage.

For focused next-hop retrieval, each activated shard searches its original content using the current next-hop targets Q_t. The implementation retains at most

$$L = 30 \tag{11}$$

focused lines from each activated shard before applying evidence extraction to the resulting local view.

For coverage-aware stopping, we use

$$\epsilon_{\text{gtc}} = 0.08. \tag{12}$$

Evidence identity is defined by the canonical record/fact ID. Duplicate hits to the same identity within one retrieval round are deduplicated, whereas repeated hits across different rounds are counted as rediscoveries for GTC.

The local implementation allows at most five iterative retrieval rounds, whereas the API configuration allows at most three rounds. A single optional follow-up round is enabled before final aggregation.

All ablation experiments modify only the target mechanism while keeping the remaining DEM configuration fixed.

E.2 End-to-End Retrieval Procedure

Algorithm 1 summarizes the complete query-time DEM procedure. Shard activation is performed once for each query. Subsequent retrieval rounds repeatedly search the same activated shard set using targets generated from the accumulated evidence and conflict structure.

Here, $\mathcal{O}_t$ denotes the raw evidence observations returned in retrieval round t, while H_t contains their unique canonical record/fact identities. Multiple hits to the same identity within a round are counted once. If the same canonical identity is observed again in a later round, its observation count is incremented and the event is treated as rediscovery.

The workspace B stores the unique accumulated evidence used for downstream reasoning. The counter C separately maintains cross-round observation frequencies used by GTC. After each round, $\text{UpdateCDAG}(B)$ converts retrieved evidence into fact tuples (s, r, o, ℓ), groups facts sharing the same (s, r) pair, and adds temporally directed edges between conflicting versions.

Focused retrieval is restricted to the initially activated shard set A. The procedure terminates when no next-hop target is generated, when the GTC estimate falls below $\epsilon_{\text{gtc}} = 0.08$, or when the configured retrieval-round limit is reached. The local implementation allows at most five iterative rounds, while the API configuration allows at most three, with a single optional follow-up round before final aggregation.

E.3 Evaluation Protocols

FactCon. We evaluate on the conflict-resolution track of MemoryAgentBench. The evaluation set contains 200 questions, consisting of 100 single-hop and 100 multi-hop examples under the `factconsolidation {sh,mh} 32k` protocol. Substring Exact Match (SubEM) is used as the primary metric, and single-hop and multi-hop SubEM are additionally reported.

LoCoMo. LoCoMo evaluates conversational long-term memory across single-hop, multi-hop, temporal, open-domain, and adversarial categories. We report F1 and compute the overall score over Categories 1–4, excluding the adversarial category from the aggregate.

Algorithm 1 Distributed Evidence Memory

Require: Task T, memory shards $\{E_i\}_{i=1}^{N}$
Ensure: Final answer $\hat{y}$

$A \leftarrow \{i : \beta(T, E_i) \geq \tau_{\text{bid}}\}$ ▷ fixed activated set
$C \leftarrow \emptyset$ ▷ evidence observation counts
$\mathcal{O}_0 \leftarrow \text{Concat}_{i \in A}\, \textsc{Extract}(T, E_i)$
$H_0 \leftarrow \textsc{RoundUniqueIDs}(\mathcal{O}_0)$
$B \leftarrow \textsc{UniqueEvidence}(\mathcal{O}_0)$
$C \leftarrow \textsc{UpdateObservationCounts}(C, H_0)$
$G \leftarrow \textsc{UpdateCDAG}(B)$
for $t = 1, \ldots, R_{\max}$ **do**
  $Q_t \leftarrow \textsc{NextHop}(T, B, G)$
  if $Q_t = \emptyset$ **then**
    break
  end if
  $\mathcal{O}_t \leftarrow \text{Concat}_{i \in A}\, \textsc{FocusedRetrieve}(T, E_i, Q_t)$
  $H_t \leftarrow \textsc{RoundUniqueIDs}(\mathcal{O}_t)$
  $B \leftarrow B \cup \textsc{UniqueEvidence}(\mathcal{O}_t)$
  $C \leftarrow \textsc{UpdateObservationCounts}(C, H_t)$
  $G \leftarrow \textsc{UpdateCDAG}(B)$
  if $\text{GTC}(C) \leq \epsilon_{\text{gtc}}$ **then**
    break
  end if
end for
$\hat{y} \leftarrow \textsc{Aggregate}(T, B, G)$
return $\hat{y}$

LongMemEval. LongMemEval evaluates long-range memory under single-hop, multi-hop, temporal reasoning, knowledge-update, and abstention settings. We report LLM-as-a-judge accuracy on the original LongMemEval-S setting and also report the MemoryAgentBench reformulation, denoted LME(S*).

Backbones and comparison protocols. The cross-system comparison uses GPT-4o-mini. Controlled local experiments use DeepSeek-v4-flash as the anchor backbone, with the data, prompts, and evaluation pipeline kept fixed across compared methods. We additionally evaluate E-mem and DEM with Qwen3.5-4B, Qwen3.5-9B, Qwen3.5-27B, MiniMax-M3, Claude 5 Sonnet, and GPT-5.5. For DeepSeek-v4-flash, we additionally run three random seeds on FactCon.

E.4 Evidence-Coverage and Compute Accounting

Evidence coverage. For the FactCon retrieval analyses, evidence coverage is defined as the event that all gold records required for a question enter the E-mem activated context or the DEM shared workspace.

For a single-hop question, coverage is equivalent to first-hop recall r_1. For a two-hop question, the complete-evidence coverage can be written as

$$c_{\text{MH}} = r_1 r_2, \tag{13}$$

where

$$r_2 = P(\text{retrieve hop 2} \mid \text{hop 1 retrieved}) \tag{14}$$

is the conditional second-hop recall.

Compute accounting. Let c_s denote one shard-screening cost, $K = |A|$, $c_{d,t}$ the average deeper-processing cost per activated shard in round t, and R_q the number of evidence-extraction rounds. A

coarse query-time cost decomposition is

$$C_{\mathrm{DEM}} = Nc_s + K \sum_{t=0}^{R_q-1} c_{d,t} + C_{\mathrm{aux}}, \tag{15}$$

where C_{aux} includes workspace maintenance, CDAG updates, and final aggregation.

Shard-level relevance checks, initial evidence extraction, and shard operations within a retrieval round can be executed in parallel. Retrieval rounds themselves remain sequential because later next-hop targets and the stopping state depend on evidence collected in previous rounds.

We therefore report model-call counts and input-token consumption in addition to retrieval-depth statistics. Relative compute is measured using input-token consumption relative to E-mem with the same backbone, as reported in Appendix Table 13.