

Do Query-Agnostic Memory Structures Earn Their Budget?

A 10M-Token Audit with Exact Read Limits

Amit Maraj

Independent Researcher
Toronto, Canada
amit.maraj@gmail.com

Kannav Sethi

Independent Researcher
Toronto, Canada
kannavsethi02@gmail.com

Abstract

Long-term memory systems often organize experience before future queries are known, but whether this improves evidence access under a fixed read budget remains unclear. We evaluate inexpensive query-agnostic segmentation and hierarchical retrieval on AMA-Bench agent trajectories and BEAM conversations spanning 128K to 10M tokens. Experiments use exact serialized-context budgets, development-only tuning, two sentence encoders, and uncertainty estimates clustered by trajectory or conversation. In an exploratory post-plan analysis of 1,117 step-anchored trajectory questions, fixed three-turn windows outperform all tested alternative boundary methods. On BEAM, primary-encoder flat retrieval recall falls from 0.564 at 128K to 0.331 at 10M. Single-vector hierarchical routing further reduces recall; multiple local vectors recover part of this loss but do not surpass flat retrieval. Doubling the read budget from 8,000 to 16,000 characters improves evidence recall at every scale with both encoders. However, the 10M evaluation, comprising 130 questions from eight conversations, detects no corresponding QA improvement. Single-vector routing reduces QA accuracy, while multi-vector routing is not distinguishable from flat retrieval. These findings identify coarse representation quality as a bottleneck in the tested hierarchies and show that improved evidence access need not yield better answers.

1 Introduction

An external memory system stores more text than a model can read for each query. It must decide what to keep, how to divide it, and what to return. Many systems make the first two decisions at write time, before the future information need is known. Examples include episodic boundaries, topic groups, fixed-depth hierarchies, and graph neighborhoods. The practical assumption is that this organization makes relevant evidence easier to retrieve later.

That assumption is difficult to test with end-to-end ablations. Removing a segmentation component usually changes the number and size of retrievable units. A hierarchy can also change how much text reaches the reader. If the comparison does not hold granularity and serialized context length fixed, an apparent organization effect can be a unit-size or budget effect.

We evaluate these inexpensive mechanisms in two settings; our conclusions concern the methods tested here. AMA-Bench provides long agent trajectories with questions that often identify the relevant execution steps. It lets us compare boundary placement while matching the number of segments. BEAM provides conversations at several scales with evidence message identifiers. It lets us measure when flat retrieval degrades and whether cheap hierarchical routing closes that gap. The benchmark tier labels describe source-stream token scale; the tested read limits are exact serialized character budgets and are reported separately.

The experimental protocol matters to the conclusion. Our initial implementation admitted an oversized first unit and did not count every separator. We discarded those aggregates, replaced the selector with an exact serialized-context rule, stored every selected span, and reconstructed all manuscript values with a separate validator. We also reran the retrieval experiments with MiniLM and MPNet sentence encoders and repeated the 10M reader and judge evaluation on the corrected contexts.

We find three effects within this protocol. First, in the post-plan trajectory analysis, boundary placement does not improve evidence recall at matched granularity. Fixed windows are better than every tested placement method with both encoders. Second, scale creates a real retrieval deficit. Flat evidence recall is approximately 0.56 at 128K and 0.33 at 10M with MiniLM, and the corresponding MPNet values are 0.58 and 0.31. Third, single-

vector coarse routing loses evidence at 10M. Multi-vector routing reduces the loss and removes the detected QA deficit, but still does not beat flat retrieval. Doubling the budget improves retrieval at every scale, but we detect no corresponding QA improvement in the corrected 10M run.

Our contributions are:

1. an exploratory matched-granularity trajectory experiment showing that fixed windows outperform random, execution-based, embedding-cohesion, and LLM-selected boundaries under two encoders;
2. a strict-budget evidence-recall curve from 128K to 10M tokens, with development-only tuning and a second-encoder replication;
3. a controlled comparison showing that single-vector routing, multi-vector routing, and neighbor expansion do not improve flat retrieval, while local multi-vector block representations recover most of the downstream loss from a single coarse mean; and
4. a raw-record correction and validation package that reconstructs budgets, evidence recall, tuning, uncertainty, small-cluster checks, reader outputs, judge votes, and costs.

2 Related Work

Agent memory systems. Long-term agent memory has been implemented as virtual context management, stored reflections, extracted facts, temporal graphs, and multi-granularity stores. MemGPT treats memory management as an operating-system problem, while Generative Agents retrieve and synthesize stored observations (Packer et al., 2023; Park et al., 2023). More recent systems organize memories through adaptive distillation, multiple granularities, agentic linking, or temporal knowledge graphs (Ma et al., 2025; Xu et al., 2025a,b; Rasmussen et al., 2025). These systems motivate the question we study, but our experiments do not reproduce their complete pipelines. We isolate inexpensive, query-agnostic organization under a common retriever.

Evaluation and benchmarks. LongMemEval and LoCoMo study long-term conversational memory (Wu et al., 2024; Maharana et al., 2024). AMA-Bench extends evaluation to long-horizon agent trajectories (Zhao et al., 2026). BEAM provides

coherent conversations across several length tiers and probing questions with source-message references (Tavakoli et al., 2025). Recent analyses have also emphasized controlled baselines, scale-conditioned failure, and the distinction between stored, retrieved, and usable evidence (Jiang et al., 2026; Wang, 2026; Shao et al., 2026). Chain-of-Memory combines lightweight construction with query-time organization and also separates retrieval recall from reasoning accuracy (Xu et al., 2026). RaMem instead induces query-specific recall conditions and preserves episodic context through synthesis (Yang et al., 2026). Our paper focuses on pre-query organization and scale under a common retriever.

Segmentation and routing. Our cohesion boundary uses the local semantic-change intuition of TextTiling (Hearst, 1997), with Sentence-BERT embeddings (Reimers and Gurevych, 2019). We include it because it is deterministic, inexpensive, and gives query-agnostic segmentation a stronger signal than arbitrary cuts. The hierarchy uses either one mean-pooled block embedding or several local block embeddings as a coarse routing stage. AMA-Bench and BEAM also introduce AMA-Agent and LIGHT, respectively (Zhao et al., 2026; Tavakoli et al., 2025). We do not reproduce those end-to-end systems. This is a mechanism test of inexpensive pre-query organization, not a claim that learned, query-aware, hierarchical, or graph memories reduce to mean pooling.

For example, a question about an earlier booking prompts flat retrieval to search messages across the conversation. Hierarchical retrieval first selects coarse blocks, then searches messages only within those blocks. A relevant message in an excluded block cannot be retrieved.

3 Experimental Protocol

3.1 Retrieval units and exact budgets

Each stream contains ordered items: action-observation turns for AMA-Bench and messages for BEAM. A segmentation maps these items to contiguous units. We mean-pool the item embeddings inside each unit, rank units by cosine similarity to the question embedding, and serialize selected units with the separator `\n---\n`.

The primary budget is 8,000 characters and the budget control is 16,000 characters. Characters are used because they are deterministic across reader providers and tokenizers. The selector visits units

in rank order. A unit is admitted only if its complete serialized text, internal newlines, and preceding separator fit within the remaining budget. Units that do not fit are skipped. An oversized first unit is never admitted. The stored context length must equal the length reconstructed from the selected spans.

Units are selected in descending similarity order under the character budget. For BEAM QA, selected units are then arranged chronologically before presentation to the reader, with messages within each unit retaining their original conversation order.

Evidence recall for question q is

$$R_q = \frac{|G_q \cap S_q|}{|G_q|}, \quad (1)$$

where G_q is the set of annotated evidence items and S_q is the set covered by selected units. The metric does not require a reader and directly tests whether organization makes evidence available under the stated budget.

3.2 Tuning and uncertainty

We assign complete trajectories or conversations to development and test partitions with seed 13. No question from a development stream contributes to a test aggregate. AMA-Bench uses 30% of trajectories for development. BEAM uses 25% of eligible conversations from each tier for development and tunes one pooled configuration across tiers.

We report question-weighted means and 95% bootstrap intervals from 2,000 resamples of trajectories or conversations. Method contrasts are paired within questions and resampled by stream. The 128K versus 10M scale contrast is unpaired and resamples conversations within each tier. The 10M test set has only eight conversations, so we also report leave-one-conversation-out ranges and exact cluster-level sign-flip tests over all 256 assignments. Intervals are not adjusted for multiple comparisons. We use them to quantify the specified contrasts rather than claim familywise error control.

3.3 Encoders

The primary encoder is all-MiniLM-L6-v2. We repeat every reader-free experiment with all-mpnet-base-v2. Item embeddings are computed once and mean-pooled for larger units and blocks. This avoids penalizing larger units through

encoder-input truncation. Question and item encoders always match within a run.

Figure 1 summarizes the evaluation pipeline and controls.

4 Trajectory Boundary Placement

4.1 Data and evidence proxies

AMA-Bench contains 208 trajectories and 2,496 questions across software engineering, web, database, embodied-agent, general-assistant, and game tasks (Zhao et al., 2026). The primary exploratory analysis uses explicit step references in the question as gold evidence. This yields 1,563 questions, of which 1,117 from 139 trajectories fall in the test split.

The analysis plan specified finding evidence through exact or fixed-threshold fuzzy matches to the reference answer, with a minimum question coverage of 25%. This procedure identified evidence for only 89 of 2,496 questions (3.57%), including 53 test questions, and therefore failed the planned coverage requirement. We instead used the steps explicitly referenced in questions as a broader evidence proxy. Because this decision followed the analysis plan, we treat the step-reference results as exploratory and hypothesis-generating. We retain the original answer-match analysis as an exploratory robustness check.

4.2 Boundary arms

Fixed windows place a boundary every N turns; for $N = 3$, cuts occur after turns 3, 6, and so on. Random boundaries use the same number of cuts but choose their locations randomly. Execution boundaries prioritize gaps where the first action token changes, indicating a possible change in activity, and fill any remaining positions with fixed gaps. Cohesion boundaries select the largest adjacent embedding drops as potential semantic transitions between turns. N is selected independently for each arm from $\{3, 6, 12, 24\}$ on development data; every arm selects $N = 3$.

The LLM arm asks Gemini 2.5 Flash-Lite to identify new subtask or phase starts in 40-turn windows. Because it chooses its own number of boundaries, we compare it both with the tuned three-turn fixed baseline and with evenly spaced boundaries that have exactly the LLM-selected segment count. Cached boundary calls cover 142 trajectories and contain 653,435 input plus 5,603 output tokens.

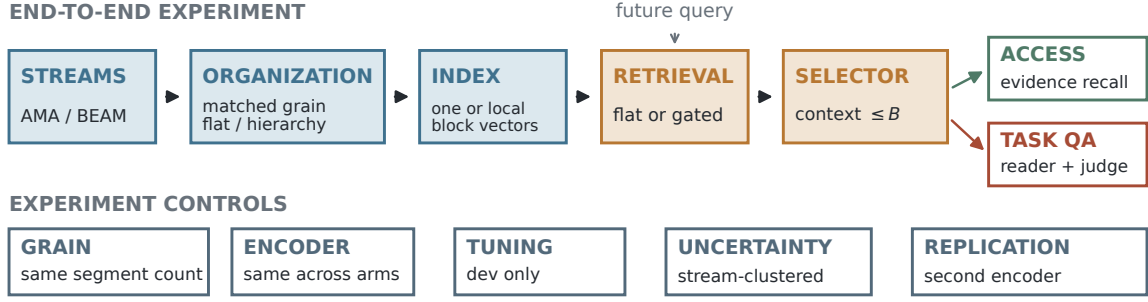


Figure 1: Architecture of the organization audit. Streams pass through matched-grain organization, a common encoder, query retrieval, and exact-budget selection before access and QA are measured separately. The lower lane lists controls held outside the tested mechanism.

4.3 Results

Within this post-plan analysis, Table 1 shows that fixed windows lead every comparison. The paired interval for each alternative is below zero under both encoders.

Figure 2a shows the paired contrasts under both encoders. The result is not an artifact of question phrases such as “Step 8.” When explicit step references are removed before encoding the query, MiniLM fixed-window recall is 0.365, compared with 0.323 for random, 0.330 for execution, and 0.314 for cohesion boundaries. The MPNet ordering is the same.

The LLM arm separates grain from placement. Its placement is 0.024 better than even spacing at the same self-chosen segment count with MiniLM, but the paired interval $[-0.018, 0.065]$ includes zero. Its coarser grain loses more than that possible placement gain. Under the under-covered registered answer-match proxy, no placement arm separates from fixed at the tuned three-turn grain. That small subset neither confirms nor overturns the step-reference analysis.

5 Conversation Scale and Organization

5.1 Data and scale curve

BEAM contains conversations at four nominal tiers and probing questions across ten memory abilities (Tavakoli et al., 2025). We call the first tier 128K, following the benchmark paper and dataset card; its source split is named 100K. Abstention questions are excluded because they have no answerable evidence. Four 1M conversations are excluded because duplicate message identifiers make their evidence pointers ambiguous. Table 2 gives the test samples.

Flat retrieval groups adjacent messages into units of size $\{1, 2, 4, 8\}$. Pooled development recall selects individual messages for both encoders. The unpaired 128K minus 10M difference is 0.233 $[0.153, 0.311]$ for MiniLM and 0.268 $[0.189, 0.345]$ for MPNet. The loss is therefore not specific to one sentence encoder.

Figure 2b puts the two encoder curves on the same scale and exposes the 0.23–0.27 loss between the shortest and longest tiers.

5.2 Organization arms

The hierarchy divides a conversation into coarse blocks, ranks a query-agnostic block index, and permits fine-message retrieval only inside the top E blocks. We compare fixed and cohesion-placed blocks represented by one mean vector. A stronger multi-vector arm uses fixed blocks, represents each with means of its non-overlapping 32-message sub-blocks, and scores the coarse block by its best local match. This spends more index storage and routing comparisons without changing the serialized reader budget. Block size $M \in \{32, 128, 512\}$ and $E \in \{2, 4, 8\}$ are selected on pooled development data. All three variants select $M = 512$ and $E = 8$. At the three shorter tiers this configuration admits the complete stream and is exactly equivalent to flat retrieval. At 10M it filters the candidate pool.

Neighbor expansion retrieves a message together with its immediate neighbors under the same final budget. The budget arm keeps flat retrieval but raises the limit from 8,000 to 16,000 characters.

Table 3 shows that no hierarchy improves evidence recall. The fixed-block loss has a paired interval below zero for MiniLM and overlaps zero for MPNet. Cohesion routing is worse with both encoders. The multi-vector arm reduces the MiniLM

Arm	MiniLM recall	Δ vs fixed	MPNet recall	Δ vs fixed
Fixed, $N = 3$	.394 [.336,.450]	reference	.408 [.351,.460]	reference
Random	.360 [.304,.414]	-.033 [-.068, -.004]	.358 [.303,.408]	-.051 [-.084, -.019]
Execution	.360 [.306,.417]	-.034 [-.066, -.004]	.356 [.303,.411]	-.052 [-.086, -.019]
Cohesion	.329 [.278,.384]	-.065 [-.102, -.027]	.322 [.271,.377]	-.086 [-.124, -.048]
LLM boundary	.328 [.270,.387]	-.066 [-.111, -.021]	.323 [.265,.377]	-.086 [-.133, -.040]
Fixed at LLM count	.304 [.247,.360]	-.089 [-.130, -.052]	.299 [.249,.351]	-.109 [-.152, -.069]

Table 1: AMA-Bench test evidence recall at an exact 8K-character budget, with 95% trajectory bootstrap intervals. Deltas are paired at the question level and clustered by trajectory ($n = 1,117$ questions, 139 trajectories).

Tier	MiniLM	MPNet	n /clusters
128K	.564 [.513,.617]	.576 [.510,.636]	264/15
500K	.554 [.502,.604]	.525 [.474,.574]	475/27
1M	.458 [.425,.499]	.447 [.417,.482]	420/24
10M	.331 [.276,.394]	.308 [.262,.360]	130/8

Table 2: Flat evidence recall at an exact 8K-character budget. Intervals resample conversations.

fixed-block deficit by 0.029, but its own 0.056 loss relative to flat still has an interval below zero. Under MPNet, it halves the fixed-block point-estimate deficit and is not separated from flat. Neighbor expansion spends scarce context on adjacent messages and performs worst.

Doubling the budget improves retrieval at every tier. The MiniLM deltas are +0.037, +0.039, +0.036, and +0.034 from 128K through 10M; all paired intervals are above zero. The MPNet deltas are +0.054, +0.055, +0.033, and +0.063, again with every interval above zero. More budget helps access, but its 10M gain is small relative to the 0.23 to 0.27 scale deficit.

5.3 Small-cluster sensitivity

All eight MiniLM leave-one-conversation-out estimates remain positive for the 16K budget arm, ranging from +0.024 to +0.039. The exact sign-flip p value is 0.0156. The fixed-block hierarchy remains negative in every leave-one-out estimate, from -0.107 to -0.064 , but its exact p value is 0.0547. The cohesion hierarchy also remains negative in every omission and has $p = 0.0469$. The multi-vector hierarchy remains negative in every omission, from -0.067 to -0.044 , with $p = 0.0312$. With MPNet, all leave-one-out signs are stable, but the exact p values are 0.0469 for the budget gain, 0.1641 for fixed blocks, 0.1250 for cohesion blocks, and 0.2969 for multi-vector blocks. These discrete tests support the direction of the effects while showing the limited inferential resolution of eight conversations.

6 Corrected 10M QA Check

We reconstruct the exact contexts for flat 8K, flat 16K, fixed-block hierarchy, and multi-vector hierarchy on the 130 test questions. Gemini 2.5 Flash answers with temperature zero and thinking disabled. Gemini 2.5 Flash-Lite judges each answer three times with a lenient semantic-equivalence rubric. Majority vote defines correctness. All 520 arm-level records completed without an API error. The recorded model cost is USD 0.5325.

Table 4 shows that the single-vector hierarchy failure transfers to answer accuracy. Every leave-one-conversation-out effect is negative, from -0.130 to -0.104 , and the exact sign-flip p value is 0.0156. The multi-vector arm is not separated from flat: its paired delta is -0.015 [-0.061 , $+0.032$], its leave-one-out range is [-0.035 , 0.000], and $p = 0.750$. The retrieval gain from 16K has no detected transfer: its QA interval crosses zero, its leave-one-out range reaches zero, and its exact p value is 0.625. This is an important boundary on the budget result. More context contains more annotated evidence, but we detect no corresponding improvement in the tested reader’s accuracy. Figure 2c places the retrieval and QA deltas together so that this separation is visible rather than inferred across tables.

7 Discussion

Boundary placement and grain are different variables. The post-plan trajectory result is more stable after correction. Random and execution-based boundaries are not tied with fixed windows; they are modestly worse with both encoders. Cohesion and LLM boundaries are worse again. A system that removes segmentation by replacing many small units with one large unit does not isolate boundary placement. It changes candidate granularity and makes each retrieval decision consume more budget. Matched segment counts are therefore a necessary control for claims about episode

10M arm	MiniLM recall	Δ vs flat	MPNet recall	Δ vs flat
Flat, 8K	.331 [.276,.394]	reference	.308 [.262,.360]	reference
Flat, 16K	.365 [.308,.425]	+.034 [.014,.058]	.371 [.303,.438]	+.063 [.021,.108]
Fixed, single-vector	.246 [.181,.309]	-.085 [-.148, -.023]	.262 [.221,.304]	-.046 [-.100, .010]
Cohesion, single-vector	.227 [.174,.274]	-.104 [-.169, -.042]	.249 [.202,.300]	-.059 [-.117, -.005]
Fixed, multi-vector	.275 [.228,.326]	-.056 [-.097, -.020]	.279 [.244,.309]	-.029 [-.074, .012]
Neighbor expansion	.120 [.084,.159]	-.211 [-.243, -.180]	.102 [.064,.146]	-.207 [-.242, -.169]

Table 3: BEAM 10M evidence recall on 130 questions from eight test conversations. Every arm except flat 16K uses an exact 8K-character budget. Paired intervals resample conversations.

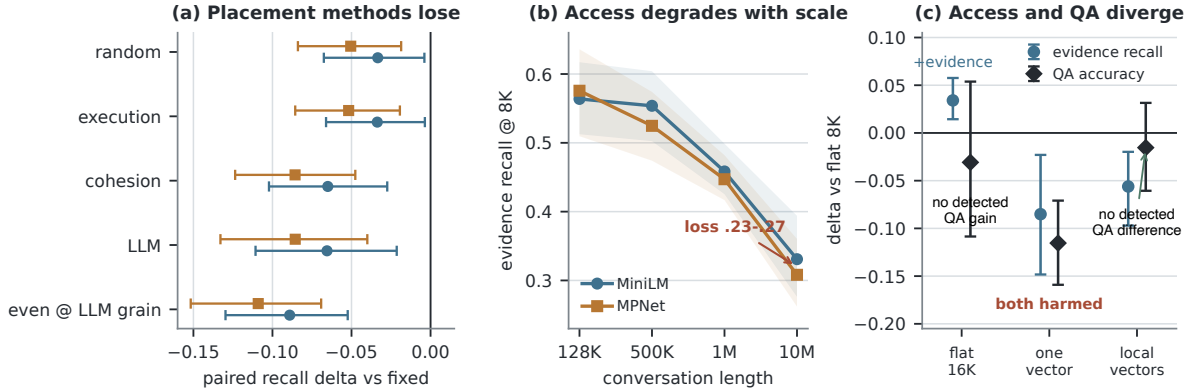


Figure 2: The empirical story across both studies. (a) Every trajectory boundary arm loses to tuned fixed windows under both encoders at matched grain. (b) flat retrieval loses 0.23–0.27 exact evidence recall from 128K to 10M. (c) at 10M, doubling the budget improves evidence recall without a detectable QA gain; a single coarse vector harms both retrieval and QA, while local block vectors recover most of the QA loss. Error bars and bands are 95% stream-clustered bootstrap intervals.

Arm	Acc. [95% CI]	Δ [95% CI]
Flat 8K	.338 [.268,.406]	reference
Flat 16K	.308 [.237,.377]	-.031 [-.109, .054]
Single-vector	.223 [.148,.299]	-.115 [-.159, -.071]
Multi-vector	.323 [.264,.388]	-.015 [-.061, .032]

Table 4: Corrected QA accuracy on 130 BEAM 10M questions. Intervals resample eight conversations.

boundaries.

A hierarchy needs a selective coarse representation. The BEAM hierarchy fails for a clear mechanism. A mean-pooled coarse block acts as a gate. If that single mean does not align with the query, all fine messages inside the block become unretrievable. The multi-vector arm directly tests this explanation. Local subblock vectors recover part of the retrieval loss and almost all of the QA loss, even though the serialized reader context budget and fine retriever are unchanged. Cohesion-based block placement does not repair the gate. Representation, not the existence of two retrieval stages by itself, is therefore the main bottleneck in these arms.

This result does not rule out query-aware organization, learned routers, extracted facts, summaries, or graph traversal. Those methods can create a better coarse representation or use query-time computation that our arms avoid. The present result gives them a minimum comparison: they should beat tuned single-message retrieval and a larger-budget control under the same serialized context limit.

Retrieval headroom is not answer headroom.

The scale curve establishes a retrieval deficit, and the 16K arm recovers a small part of it. The QA check detects no corresponding accuracy improvement. Added evidence can be redundant, weakly ordered, or hard to integrate. A memory paper should therefore report both evidence access and answer use when making an end-to-end claim. Retrieval recall diagnoses the memory stage; it does not substitute for reader accuracy.

Practical applicability. These results define a deployment test for agent memory. Before adding episode boundaries or coarse routing, match grain, encoder, and realized prompt budget against flat retrieval. Log coarse-route exclusions because one

false negative can hide a block’s fine evidence. If access improves, test the same questions at the reader layer; our 16K arm shows that more evidence may not improve answers. Query-aware alternatives remain motivated, but must earn their complexity against matched flat and larger-budget controls.

8 Conclusion

We tested inexpensive query-agnostic memory organization under matched granularity, encoder, and serialized-context budgets. In the post-plan AMA-Bench analysis, fixed three-turn windows outperformed four alternative boundary methods. On BEAM, flat retrieval recall declined by 0.23–0.27 from 128K to 10M tokens. Single-vector hierarchies showed further losses, while local block vectors recovered part of the deficit without surpassing flat retrieval. The QA evaluation complements these retrieval findings: the tested single-vector hierarchy reduced answer accuracy, whereas the multi-vector arm was not statistically distinguishable from flat retrieval, and additional retrieved evidence produced no detected accuracy gain. Together, these findings suggest that query-agnostic structure earns its budget when its representations preserve access to evidence that the reader can effectively use.

9 Limitations

The trajectory analysis uses a post-plan step-reference proxy. It is explicit and covers a large sample, but the decision to elevate it was not preregistered. The registered answer-string proxy failed its coverage requirement and cannot provide a comparably powered confirmation.

Matching segment counts does not equalize segment-length distributions. Differences in unit length may affect mean-pooled representations and budget packing, so these comparisons do not isolate boundary placement alone.

AMA-Bench and BEAM are benchmark distributions, not a sample of deployed agent histories. Their questions and evidence annotations inherit any benchmark-generation and validation errors. Four 1M BEAM conversations are excluded because duplicate message identifiers prevent unique evidence resolution. The 10M analysis contains only eight test conversations. Cluster bootstrap intervals, leave-one-out estimates, and exact sign-flip tests cannot create more independent conversa-

tions.

Character budgets are exact and provider-independent, but they are not token budgets. The 10M label describes source-stream scale, not the 8K or 16K character read limits. We test two English sentence encoders, single-vector means, and one fixed local multi-vector construction. We do not test trained routers, cross-encoders, LLM-generated summaries, temporal graphs, query-aware write policies, AMA-Agent, or LIGHT. The study accounts for serialized reader context and recorded hosted-model cost, but does not compare retrieval cost or latency, indexing latency, embedding storage, or end-to-end serving cost. The QA check uses one reader family and one judge family, with no human calibration. Results may change with other memory methods, encoders, read budgets, readers, rubrics, or future provider model versions.

The study was corrected after the initial selector failed validation. The step-reference analysis was also chosen after the registered proxy proved too sparse. We disclose both events, retain source hashes and the failed-proxy result, and do not adjust intervals for the number of contrasts. These choices make the evidence auditable but do not turn the analysis into a preregistered confirmatory study.

10 Ethical Considerations

The experiments use existing benchmarks and no recruited participants. Hosted QA sends benchmark excerpts to a model provider; reproductions must follow dataset terms and provider policy. Credentials and caches are excluded. AMA-Bench is MIT; BEAM and BEAM-10M are CC BY-SA 4.0, so derived records preserve attribution and share-alike terms.

AI Assistance Disclosure

AI assistants were used for coding support, literature search, manuscript editing, and automated cross-checking. The authors verified cited sources, executed and validated the experiments, inspected the raw artifacts, and take responsibility for every claim and sentence. Model-generated boundaries, reader outputs, and judge votes used as research objects are separately identified in the experimental protocol.

Reproducibility Statement

The release records hashes, splits, tuning, encoders, evidence indices, selected context, predictions, votes, costs, and errors. Validators reconstruct aggregates, including MPNet results, and reject budget, pairing, provenance, or vote-count violations. The release bundle contains the paper, code, figures, summaries, and raw records, but excludes dataset text and embedding caches.

References

- Marti A. Hearst. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1).
- Dongming Jiang, Yi Li, Songtao Wei, Jinxin Yang, Ayushi Kishore, Alysa Zhao, Dingyi Kang, Xu Hu, Feng Chen, Qiannan Li, and Bingzhe Li. 2026. Anatomy of agentic memory: Taxonomy and empirical analysis of evaluation and system limitations. *arXiv preprint arXiv:2602.19320*.
- Wenquan Ma, Jiayan Nan, Wenlong Wu, and Yize Chen. 2025. What deserves memory: Adaptive memory distillation for LLM agents. *arXiv preprint arXiv:2508.03341*.
- Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. [Evaluating very long-term conversational memory of LLM agents](#). In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 13851–13870.
- Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2023. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*.
- Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *ACM Symposium on User Interface Software and Technology (UIST)*.
- Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. 2025. Zep: A temporal knowledge graph architecture for agent memory. *arXiv preprint arXiv:2501.13956*.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3982–3992. Association for Computational Linguistics.
- Jiaqi Shao, Yiyi Lu, Yunzhen Zhang, and Bing Luo. 2026. When stored evidence stops being usable: Scale-conditioned evaluation of agent memory. *arXiv preprint arXiv:2605.07313*.
- Mohammad Tavakoli, Alireza Salemi, Carrie Ye, Mohamed Abdalla, Hamed Zamani, and J. Ross Mitchell. 2025. Beyond a million tokens: Benchmarking and enhancing long-term memory in LLMs. *arXiv preprint arXiv:2510.27246*.
- Kuan Wang. 2026. MemDelta: Controlled baselines and hidden confounds in agent memory evaluation. *arXiv preprint arXiv:2606.29914*.
- Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2024. LongMemEval: Benchmarking chat assistants on long-term interactive memory. *arXiv preprint arXiv:2410.10813*.
- Derong Xu, Yi Wen, Pengyue Jia, Yingyi Zhang, Wenlin Zhang, Yichao Wang, Huifeng Guo, Ruiming Tang, Xiangyu Zhao, Enhong Chen, and Tong Xu. 2025a. Towards multi-granularity memory association and selection for long-term conversational agents. *arXiv preprint arXiv:2505.19549*.
- Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2025b. A-MEM: Agentic memory for LLM agents. *arXiv preprint arXiv:2502.12110*.
- Xiucheng Xu, Bingbing Xu, Xueyun Tian, Zihe Huang, Rongxin Chen, Yunfan Li, and Huawei Shen. 2026. [Chain-of-memory: Lightweight memory construction with dynamic evolution for LLM agents](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11618–11631. Association for Computational Linguistics.
- Wei Yang, Bryce Kan, Shixuan Li, Li Li, Yuehan Qin, Jiate Li, Paul Bogdan, and Jesse Thomason. 2026. [RaMem: Contextual reinstatement for long-term agentic memory](#). *arXiv preprint arXiv:2606.22844*.
- Yujie Zhao, Boqin Yuan, Junbo Huang, Haocheng Yuan, Zhongming Yu, Haozhou Xu, Lanxiang Hu, Abhilash Shankarampeta, Zimeng Huang, Wentao Ni, Yundong Tian, and Jishen Zhao. 2026. AMA-Bench: Evaluating long-horizon memory for agentic applications. *arXiv preprint arXiv:2602.22769*.