Same Ranking, Different Winner: How Scoring Targets Shape LLM Memory Benchmarks

Anonymous ACL submission

Abstract

Conversational-memory systems increasingly transform dialogue history into facts, summaries, timelines, and other source-linked descendants, so a single source turn can coexist with several derived memories in the same retrieval index. This raises an under-specified evaluation question: which stored form should receive retrieval credit? We show that this *scoring-target* choice is often left implicit and can materially change benchmark conclusions. We present TIAP, a fixed-output audit that rescores saved ranked outputs under three targets—Raw, Source, and Canonical—without rerunning retrieval. On LoCoMo and LongMemEval-S, switching only the credited target changes nDCG on 83.4%–94.0% of shared queries, flips target orderings on Mem0 and MemoryOS transfer runs, and reverses parser-density recommendations. A 1,902-case semantic audit further shows that relaxed source-linked credit is fully justified only 29.2% of the time, despite high rubric reliability in a validation subset. These results reveal *target noninvariance*: conclusions about memory architectures can silently flip with a single benchmark-design choice. Conversational-memory papers should therefore define and report the scoring target explicitly.

1 Introduction

Long term conversational AI systems are increasingly built around *memory construction*. Rather than passing raw dialogue history directly to the response model, modern chat assistants extract facts, consolidate updates, build timelines, populate profile memories, and often retain the original dialogue turn alongside its transformed descendants in the same retrieval index (Chhikara et al., 2025; Kang et al., 2025; Rasmussen et al., 2025; Ong et al., 2025; Tan et al., 2025). This storage pattern is now central to memory-augmented agents (Kim et al., 2025; Ong et al., 2025; Wu et al., 2025), but it also creates an under-specified evaluation problem. A benchmark query is typically anchored to one source turn, while the evaluated store may contain several retrieval units derived from that same evidence. Which of these stored forms should receive retrieval credit?

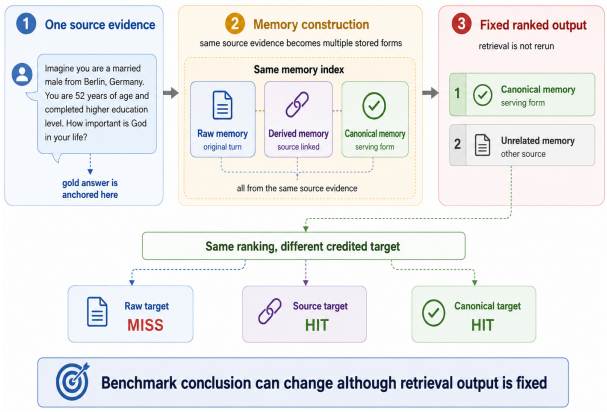

Figure 1: Overview of target noninvariance in conversational memory evaluation. A single source evidence turn can produce multiple stored memory forms in the same index. For a fixed ranked output, changing only the credited target can turn the same retrieval result into a miss or a hit.

Figure 1 summarizes the core failure mode. A single source evidence turn may be stored as a raw memory, a source-linked derived memory, and a canonical serving memory. When the ranked output is fixed, the same retrieved item can be scored as a miss under a Raw target but as a hit under Source or Canonical targets. Thus, the benchmark

conclusion can change even though retrieval behavior has not changed. The ambiguity has both semantic and numerical consequences. Semantically, a stored descendant may share the correct source anchor while omitting the answer, dropping an important condition, or preserving only weakly related context. In our 1,902 case audit of relaxed credit, only 29.2% of credited source-linked descendants fully support the query under a five-model majority vote. Thus, a more permissive target is not automatically a more correct one, and answer support can only be judged after the benchmark decides which stored evidence units are admissible (Es et al., 2024; Saad-Falcon et al., 2024).

Numerically, target choice is not a bookkeeping artifact. If relaxing the target merely raised recall-like scores, the broadest target would dominate. Instead, target choice interacts with benchmark structure and memory architecture. In transfer runs, target orderings are non-monotonic for both Mem0 and MemoryOS, depending on the benchmark. Thus, the same fixed ranked outputs can recommend different memory-store designs depending only on which stored representation is credited. Because conversational-memory benchmarks increasingly guide retriever, storage-policy, and serving-representation choices, comparisons are only meaningful if the benchmark defines what counts as a correct retrieval. In transformed-memory stores that is exactly the contested object, yet it is typically left implicit. This is a benchmark-design problem rather than a system-tuning one: unlike hyperparameter sensitivity analysis, which varies a system knob while holding evaluation fixed, our audit varies the evaluation definition while holding system output fixed. It is also distinct from classical IR evaluation-sensitivity analysis, which varies relevance judgments or changes the retrieval unit globally rather than crediting system-generated descendants of a single source turn; we develop this contrast in Section 2.

We study this setting on LoCoMo and LongMemEval-S across four native retrievers. Switching only the scoring target changes Raw versus Canonical nDCG on 83.4% to 94.0% of shared queries. Transfer runs with Mem0 and MemoryOS remain target sensitive but show different target orderings, confirming that the effect is a benchmark validity problem on fixed ranked outputs rather than an artifact of target size alone.

This paper makes three contributions.

- We identify *target noninvariance*, a benchmark validity problem in transformed conversational memory: when ranked outputs are fixed, changing only the admissible target set can change scores, system rankings, and store design recommendations.
- We show across two primary benchmarks, one supplementary stress test, and two external memory architectures that the effect is large, affecting 83.4% to 94.0% of shared queries, surviving transfer to Mem0 and MemoryOS, and not being removed by post-hoc aggregation.
- We provide TIAP, a fixed-output audit procedure, and MTEL-Mem, a reproducibility layer that packages target mappings, source fixtures, and rescoring utilities for auditing saved traces.

We do not argue that any one target is universally correct. Rather, when a benchmark can credit different stored forms of the same evidence, scoring target choice is part of the evaluation design and should be defined and reported explicitly.

2 Related Work

Conversational-memory benchmarks such as LoCoMo, LongMemEval-S, and BEAM have established retrieval- and QA-based evaluation for long-horizon dialogue systems (Maharana et al., 2024; Wu et al., 2025; Tavakoli et al., 2026). They differ in answer-level scoring: LoCoMo uses lexical metrics such as F1 and BLEU-1 alongside retrieval recall over annotated evidence turns, while LongMemEval-S uses a GPT-4o judge for flexible answer correctness and reports Recall and nDCG when retrieval traces expose answer-location labels. Our audit is upstream of these choices. It asks which stored memory IDs are eligible to satisfy the annotated evidence before either a lexical scorer or a judge evaluates the final answer.

Memory-system papers (Park et al., 2023; Zhong et al., 2023; Packer et al., 2023; Chhikara et al., 2025; Kang et al., 2025; Rasmussen et al., 2025; Ong et al., 2025; Tan et al., 2025; Chen et al., 2023; Lee et al., 2024a) focus on architecture, storage policy, and memory management. Many systems extract, summarize, or canonicalize conversational evidence before serving it. We study the evaluation target induced by these storage choices: when raw turns and derived memories are linked to the same evidence, which stored IDs should receive credit?

The closest analogy is classical IR work on evaluation sensitivity (Bailey et al., 2008; Yang et al., 2019; Craswell et al., 2020; Voorhees, 2001; Voorhees and Harman, 2005; Manning et al., 2008; Järvelin and Kekäläinen, 2002). Prior work shows that conclusions can be fragile under incomplete judgments (Buckley and Voorhees, 2004; Zobel, 1998; Sanderson and Zobel, 2005; Carterette et al., 2006), and document-versus-passage evaluation shows that the scored unit can matter even when the evidence is related (Yang et al., 2019; Craswell et al., 2020). BEIR further shows that retrieval conclusions depend on benchmark design choices (Thakur et al., 2021). However, these settings usually vary judgments over fixed units or change the retrieval unit globally. In transformed memory stores, the system itself creates multiple source-linked descendants of the same evidence inside one evaluated index, and the benchmark must decide which descendants are admissible targets.

RAG evaluation frameworks such as RAGAs and ARES evaluate whether retrieved evidence supports an answer (Es et al., 2024; Saad-Falcon et al., 2024), but they hold the candidate evidence unit fixed. The same ambiguity can arise in chunk-based RAG, abstractive summarization indices, and knowledge-graph-augmented retrieval whenever an index stores non-identity transformations of source documents. Our audit makes this question explicit for conversational memory: before answer-level evaluation begins, which stored descendants of the same source evidence count as admissible retrieval targets?

3 Method

Figure 2 summarizes the TIAP audit framework in three stages. First, TIAP constructs Raw, Source, and Canonical scoring targets for each query and memory store. It then rescores the same saved ranked traces under these targets without rerunning retrieval. Finally, it measures whether the credited target changes scores, winners, design recommendations, or the semantic validity of contested credits.

3.1 Stage I: Constructing Scoring Targets

Notation. Let q be a benchmark query, S a memory store, $\mathcal{I}(S)$ the set of stored memory IDs in S, and $R_k^S(q)$ the top k ranked output emitted by S. Let $s(q)$ denote the source evidence that anchors the gold label. For any stored memory $m \in \mathcal{I}(S)$, let $\ell_S(m)$ be its source anchor. We use $\mathrm{raw}_S(m)$ to indicate that m is the raw source turn and $\mathrm{can}_S(m)$ to indicate that m is a transformed canonical memory intended for serving.

Scoring targets. A *scoring target* is the set of stored memory IDs eligible to receive credit for a query. We evaluate three targets: Raw, which credits only the exact raw source turn; Source, which credits any stored descendant linked to the same source evidence; and Canonical, which credits transformed non-raw descendants intended for serving. Canonical is a serving-form target set rather than a single globally canonical record, since a source may have multiple transformed descendants.

Formally, for a fixed source anchor $s(q)$ and store S:

$$T_{\mathrm{raw}}(q,S) = \{m \in \mathcal{I}(S) : \ell_S(m) = s(q), \mathrm{raw}_S(m)\}, \quad (1)$$

$$T_{\mathrm{src}}(q,S) = \{m \in \mathcal{I}(S) : \ell_S(m) = s(q)\}, \quad (2)$$

$$T_{\mathrm{can}}(q,S) = \{m \in \mathcal{I}(S) : \ell_S(m) = s(q), \mathrm{can}_S(m)\}. \quad (3)$$

By construction, $T_{\mathrm{raw}}(q,S) \subseteq T_{\mathrm{src}}(q,S)$ and $T_{\mathrm{can}}(q,S) \subseteq T_{\mathrm{src}}(q,S)$. In most transformed memory stores, $T_{\mathrm{raw}}(q,S)$ and $T_{\mathrm{can}}(q,S)$ are disjoint: the former contains the original evidence turn, while the latter contains transformed serving memories derived from it. We call the rules a store uses to assign ℓ_S, raw_S, and can_S its *lineage contract*. When this contract is shared across compared systems, we omit S for readability.

3.2 Stage II: Fixed Output Rescoring

Audit question. TIAP asks what changes when only the scoring target changes. It answers this by rescoring the same saved ranked outputs under Raw, Source, and Canonical targets. Retrieval is not rerun; the ranked traces remain fixed, and only the credited memory IDs change.

System gap under a target. For systems A and B, metric M, and target $t \in \{\mathrm{raw}, \mathrm{src}, \mathrm{can}\}$, we define

$$\Delta_t(A,B) = M_t(A) - M_t(B), \quad (4)$$

where M_t is M evaluated against target set $T_t(q,S)$. If the sign of $\Delta_t(A,B)$ changes as t varies, the benchmark conclusion is not target-invariant

This differs from standard assessor disagreement, where items are fixed and only judgments vary. Here, the credited IDs themselves change with the

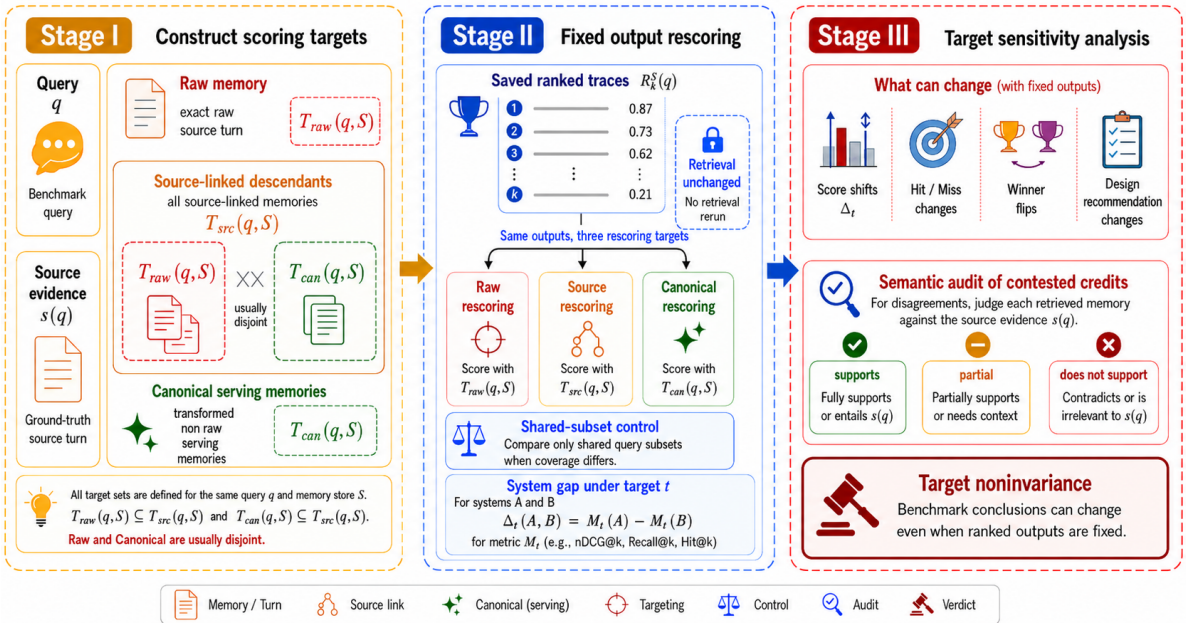

Figure 2: Overview of the TIAP audit framework. Stage I constructs Raw, Source, and Canonical scoring targets from the same query, memory store, and source evidence. Stage II performs fixed output rescoring by applying the three targets to the same saved ranked traces and controlling for shared query coverage. Stage III analyzes target sensitivity and audits contested credits for semantic support.

store: a benchmark using only Raw credit has already decided that a raw turn can succeed while a canonical descendant of the same evidence cannot. TIAP makes this representational assumption explicit and tests whether conclusions survive it.

Shared query subset control. Target coverage can differ across Raw, Source, and Canonical scoring because not every source turn produces a transformed serving memory. Directly comparing targets on different query sets would confound scoring target choice with query difficulty. TIAP therefore repeats target comparisons on shared query subsets when coverage differs. A query is included only when all compared targets have at least one eligible stored ID, ensuring that score changes reflect the credited target set rather than changes in the evaluated query population.

3.3 Stage III: Target Sensitivity and Semantic Audit

Target sensitivity outputs. After rescoring, TIAP records score shifts, query-level hit status changes, winner flips, and design recommendation changes. A winner flip occurs when the sign of $\Delta_t(A, B)$ changes across scoring targets, so the preferred system depends on which stored representation is credited. A design recommendation is the memory store or configuration that appears best under a chosen metric and scoring target. If two configurations are ranked differently under Raw, Source, and Canonical while their saved ranked outputs remain fixed, then the benchmark recommendation depends on the scoring target rather than on retrieval behavior.

Semantic audit of contested credits. A broader target can award more retrieval credit, but source lineage alone does not guarantee semantic support: a transformed memory may omit the answer, remove a relevant condition, change the temporal scope, or preserve only weakly related context. We therefore audit contested credits where Raw misses but Source or Canonical receives credit. Each credited memory is judged against the source evidence $s(q)$ and the answer required by q, then assigned one of three labels: *supports* if it fully supports the answer, *partial* if it contains related evidence but requires additional context, and *does not support* if it contradicts, omits, or is irrelevant to the required evidence. This separates provenance from semantic support and measures whether relaxed credit is justified rather than merely available.

4 Experimental Setup

4.1 Datasets

We instantiate the audit on two conversational memory benchmarks, LoCoMo (Maharana et al., 2024) and LongMemEval-S (Wu et al., 2025). We also use BEAM (Tavakoli et al., 2026) as a supplementary stress test. We select these benchmarks be-

Table 1: Dataset coverage used by the audit. BEAM-1M is supplementary.

Dataset	Fixture	Eval	Native runs	Canon. cov.
LoCoMo	5,882	1,977	1,510–1,533	899–963
LongMemEval-S	231,595	470	439–470	299–332
BEAM-1M (supp.)	74,630	625	625	408

cause each annotates gold evidence at the level of source conversation turns or sessions. This makes source linked lineage recoverable and allows us to construct the Raw, Source, and Canonical target distinction. Table 1 summarizes the coverage used by the audit. **LoCoMo** (Maharana et al., 2024) evaluates long term conversational memory and annotates turn IDs for QA gold answers. It reports lexical automatic metrics for answer quality and retrieval recall over annotated evidence turns. This makes it suitable for testing whether different stored forms of the same source turn receive different retrieval credit. **LongMemEval-S** (Wu et al., 2025) benchmarks chat assistants on long term interactive memory. It annotates evidence sessions and positions within them, uses a judge based answer scorer, and reports Recall and nDCG when retrieval traces expose answer location labels. Its evidence annotations allow us to recover the source anchor for each query and construct target mappings over stored memories. **BEAM** (Tavakoli et al., 2026) provides a grounded million token slice with a `source_chat_ids` field. We use only its 1M grounded slice and treat it as a single provider robustness check rather than part of the main native run matrix. Across all benchmarks, queries, reference answers, and evidence labels are held out and used only at evaluation time. They are never provided to the parser that produces transformed memories, so they cannot leak into the memory store.

4.2 Target Mappings

For each query, we map the benchmark source evidence to Raw, Source, and Canonical targets using source linked lineage metadata. The mapping is deterministic: benchmark source fixtures are joined to the raw memory ID and to all transformed descendant IDs exported with the same source anchor. This allows every saved ranked trace to be rescored under the three target definitions without rerunning retrieval. The primary native store contains the raw source turn together with parser derived transformed memories linked back to the same source evidence. Source unions the raw and transformed IDs associated with a source fixture. Canonical keeps only non raw transformed IDs, so it is a serving form target set rather than a guarantee of one canonical item per source. The parser operates only on conversation turns. Canonical coverage can differ from Raw coverage because not every source turn produces a transformed serving memory. When a query has no Canonical target, we exclude it from Canonical comparisons. Raw and Source comparisons use their available shared query sets. This makes shared subset rescoring part of the main analysis rather than an optional check. For the parser density sweep, F_1, F_5, and F_8 denote *parser_max_facts* equal to 1, 5, and 8 transformed facts per source turn.

4.3 Retrieval and Memory Systems

We evaluate four native retrievers: lexical search, all-MiniLM-L6-v2 (Reimers and Gurevych, 2019), BGE-M3 (Chen et al., 2024), and mxbai-embed-large (Lee et al., 2024b). All main results are computed from saved top $k = 60$ ranked traces and rescored under the target sets defined in Section 3.1. We use $k = 60$ because it matches or exceeds the retrieval depth used by the memory systems in our scope and provides enough depth for source-linked descendants to appear in the ranked output. On saved traces, Recall@k plateaus after approximately $k = 40$, with less than six percentage points of additional recall from $k = 40$ to $k = 60$ across retriever and dataset combinations. Using $k = 60$ therefore reduces the risk that target sensitivity is caused by an unusually shallow retrieval cutoff, while keeping the audit focused on evidence that would plausibly be available to a deployed memory system. All native runs use extractive answer mode and random seed `1337`. The lexical run uses the same benchmark harness without dense embeddings, and embedding providers are served locally via Ollama. Shared-subset paired statistics use 3,000 paired bootstrap resamples, percentile confidence intervals, and RNG seed `1337`. The transfer analysis includes full LoCoMo and LongMemEval-S runs for both Mem0 (Chhikara et al., 2025) and MemoryOS (Kang et al., 2025). Zep (Rasmussen et al., 2025) is out of scope because its graph-native memories do not expose a stable source-turn to descendant ID mapping.

4.4 Semantic Audit Methodology

The semantic audit covers all 1,902 `raw_miss / source_hit / canonical_hit` disagreement cases extracted from the eight native runs.

Table 2: Fixed-subset rescoring on canonical-covered shared subsets for all eight native runs. Canonical exceeds Raw on every shared subset; Source exceeds Canonical on all four LoCoMo shared subsets.

Dataset	Provider	n	Raw	Source	Canonical	Canonical–Raw	Canonical–Source
LoCoMo	Lexical	899	0.1876	0.3809	0.3556	+0.168 [0.147, 0.189]	-0.025 [-0.036, -0.015]
	all MiniLM	899	0.2118	0.4198	0.3875	+0.176 [0.154, 0.198]	-0.032 [-0.043, -0.022]
	BGE M3	963	0.2548	0.4109	0.3231	+0.068 [0.045, 0.091]	-0.088 [-0.101, -0.075]
	mxbai-embed-large	899	0.2120	0.4195	0.3870	+0.175 [0.152, 0.198]	-0.033 [-0.043, -0.022]
LongMemEval-S	Lexical	332	0.1877	0.4041	0.4558	+0.268 [0.236, 0.303]	+0.052 [0.038, 0.066]
	all MiniLM	332	0.2001	0.4561	0.5309	+0.331 [0.299, 0.364]	+0.075 [0.061, 0.089]
	BGE M3	299	0.1969	0.4576	0.5326	+0.336 [0.302, 0.371]	+0.075 [0.061, 0.090]
	mxbai-embed-large	332	0.1815	0.5032	0.6015	+0.420 [0.382, 0.457]	+0.098 [0.083, 0.113]

Table 3: Inter-rater agreement on the 115-case validation subset used to confirm rubric consistency before scaling to the full 1,902-case audit. Binary labels collapse *supports* $\cup$ *partial* $\rightarrow$ *relevant*. Fleiss' κ is computed over all five raters; pairwise values are Cohen's κ.

	Binary		Three-class	
Rater / pair	κ	Agree	κ	Agree
All five raters (Fleiss)				
1 human + 4 LLMs	**0.83**	n/a	0.79	n/a
Each model vs. human (Cohen)				
GPT-5.5	0.87	93.9%	0.83	88.7%
Claude Opus 4.7	0.85	93.0%	0.83	88.7%
Gemini 3.1 Pro	0.82	91.3%	0.74	82.6%
DeepSeek V4 Pro	0.77	88.7%	0.73	81.7%
Model–model pairs (Cohen, binary)				
Gemini $\leftrightarrow$ GPT	0.91	95.7%	n/a	n/a
Claude $\leftrightarrow$ GPT	0.91	95.7%	n/a	n/a
Claude $\leftrightarrow$ Gemini	0.89	94.8%	n/a	n/a
Claude $\leftrightarrow$ DeepSeek	0.77	88.7%	n/a	n/a
GPT $\leftrightarrow$ DeepSeek	0.75	87.8%	n/a	n/a
Gemini $\leftrightarrow$ DeepSeek	0.74	87.0%	n/a	n/a

These cases are selected by a deterministic extraction script over saved traces and constitute the complete set of native run cases where Raw misses while both Source and Canonical receive credit. Labels follow the three-way rubric defined in Section 3.3: `supports`, `partial`, and `does_not_support`. Auxiliary serving semantics tags are reported only as exploratory notes and are not used as primary evidence.

To validate the rubric before scaling, we first had a human annotator label a 115-case subset drawn from LongMemEval-S with all-MiniLM and LoCoMo with BGE-M3. The subset was stratified by credited rank bucket, using ranks 1 to 5, 6 to 20, and 21 to 60, and evenly spaced by query index within each bucket. A second human reviewer independently labeled a 36-case overlap subset, and disagreements were adjudicated. We then ran four frontier LLMs on the same 115 cases found at Table 13. Across the five raters, the validation subset achieved Fleiss' $\kappa = 0.83$ under a binary mapping, indicating that the rubric produced consistent classifications on this subset. After this validation, we applied the rubric to all 1,902 contested cases using five frontier LLMs via the OpenRouter API: `claude-sonnet-4`, `claude-opus-4-7`, `gemini-3.1-pro`, `gpt-5.5`, and `deepseek-v4-pro`. Each model judged each case independently at temperature 0.

5 Results

We organize the results around whether scoring targets change scores, store design recommendations, query-type behavior, relaxed-credit reliability, or transfer behavior.

5.1 Scoring Target Alters Retrieval Scores

The central result is that target choice materially changes retrieval scores even when ranked outputs are unchanged. On canonical-covered shared subsets, switching only the scoring target changes nDCG on 83.4% to 94.0% of Raw versus Canonical rescoring comparisons across LoCoMo and LongMemEval-S. Thus, target choice is not a minor implementation detail: it changes the measured relevance of the same ranked outputs for most evaluated queries.

Coverage must be controlled explicitly. Queries with a Canonical target are easier under Raw retrieval than queries without Canonical coverage: Raw nDCG is higher by a mean run-level gap of $+0.086$ with bootstrap 95% CI $[+0.049, +0.123]$, while answer F1 changes much less, with a mean gap of $+0.008$ and CI $[+0.002, +0.014]$. We therefore report Canonical versus Raw comparisons only on canonical-covered shared query subsets.

The score shifts are large and target dependent (Table 2). On LongMemEval-S, Canonical exceeds Raw for every provider. The largest Canonical minus Raw effects appear for BGE-M3, with

Table 4: Raw versus Canonical query-level nDCG instability.

Dataset	Provider	$\|Q\|$	$\Delta \neq 0$	Rate
LoCoMo	Lexical	899	758	84.3%
LoCoMo	all MiniLM	899	794	88.3%
LoCoMo	BGE M3	963	851	88.4%
LoCoMo	mxbai	899	794	88.3%
LongMemEval-S	Lexical	332	277	83.4%
LongMemEval-S	all MiniLM	332	309	93.1%
LongMemEval-S	BGE M3	299	281	94.0%
LongMemEval-S	mxbai	332	312	94.0%

Table 5: Target-dependent winner by parser density (F_1, F_5, F_8) on the fixed shared subset ($n = 453$). The winning configuration depends on whether scoring uses Raw, Source, or Canonical.

Provider	Comp.	Raw	Source	Canonical
Lexical	F_1 vs F_5	F_1	F_5	F_1
	F_1 vs F_8	F_1	F_8	F_1
	F_5 vs F_8	F_5	F_8	F_8
all-MiniLM	F_1 vs F_5	F_1	F_5	F_1
	F_1 vs F_8	F_1	F_8	F_1
	F_5 vs F_8	F_8	F_8	F_8

$+0.336$ and CI $[0.302, 0.371]$, and mxbai-embed-large, with $+0.420$ and CI $[0.382, 0.457]$. On LoCoMo, Canonical also exceeds Raw for every provider, but Source remains above Canonical for all four providers. The strongest Source versus Canonical reversal appears on BGE-M3, where Canonical minus Source is -0.088 and the confidence interval excludes zero. Table 4 reports the compact query-level instability matrix behind the headline result. The rate of nonzero nDCG change ranges from 83.4% to 94.0% across the eight native runs. The shifts are not concentrated near zero: in the full distribution, 65% of LoCoMo and 71% of LongMemEval-S query-provider pairs shift by at least 0.1 nDCG (Figure 3 in Appendix A). Thus, target choice changes both absolute retrieval scores and the relative evaluation landscape across datasets and retrievers.

5.2 Scoring Target Changes Store Design Recommendations

The score shifts also change an engineering decision. We run a parser density sweep on a fixed matched subset with $n = 453$, using `parser_max_facts` values of 1, 5, and 8. We denote these settings as F_1, F_5, and F_8. Changing only the scoring target changes which density appears best. For the lexical comparison between F_1 and F_5, Raw prefers the sparser F_1 store, Source prefers the denser F_5 store, and Canonical returns to F_1. A practitioner reading only the Raw-scored benchmark would choose `parser_max_facts` equal to 1, saving extraction cost, storage, and index size. The same saved comparison under Source scoring instead recommends `parser_max_facts` equal to 5. The ranked outputs are fixed across these rescorings. What changes is which source-linked memory IDs are allowed to receive credit.

Across the six pairwise comparisons in Table 5, Source prefers the denser setting in all six comparisons, Raw prefers the sparsest setting in four of six comparisons, and Canonical is split. Thus, target choice can affect downstream engineering decisions, not merely benchmark reporting.

5.3 Target Sensitivity Is Broad Across Query Types

Target sensitivity is not confined to one query family. Across shared queries, Raw versus Canonical rescoring changes nDCG for 83.6% to 93.2% of queries within major LoCoMo categories and for 77.1% to 100.0% of queries within LongMemEval-S categories. The effect also persists in categories with lower Canonical coverage, including LoCoMo Single-hop and Temporal questions and LongMemEval-S single-session questions.

A useful boundary appears in Source versus Canonical comparisons on LongMemEval-S. In that benchmark, instability is concentrated in multi-session and knowledge-update questions, while it is near zero for single-session assistant questions. This suggests that Source versus Canonical target choice matters less when the answer is a simple single-session fact, but becomes more consequential when a query requires aggregation, update tracking, or reasoning across sessions.

5.4 Relaxed Credit Is Only Partially Reliable

A natural concern is that relaxed targets may give credit too easily. Using the semantic audit methodology in Section 4.4, we audit all 1,902 `raw_miss / source_hit / canonical_hit` disagreement cases across the eight native runs.

Five frontier LLMs were prompted to judge every case independently (Table 6). Under the five-model majority vote, 29.2% of contested credits fully support the query, 39.6% are partially supportive, and 31.2% do not support it. The gap between datasets is substantial. On LoCoMo, 33.1% of relaxed credit is fully justified, compared with 21.3%

Table 6: Semantic audit of contested credits. Majority denotes the five-model majority-vote label.

Dataset	Judge	n	Supports	Partial	Not support
LoCoMo	Sonnet 4	1,278	384 (30.0%)	553 (43.3%)	341 (26.7%)
	Opus 4.7	1,278	409 (32.0%)	513 (40.1%)	356 (27.9%)
	Gemini 3.1	1,278	447 (35.0%)	487 (38.1%)	344 (26.9%)
	GPT-5.5	1,269	413 (32.5%)	588 (46.3%)	268 (21.1%)
	DeepSeek V4	1,264	409 (32.4%)	577 (45.6%)	278 (22.0%)
	Majority	**1,256**	**416 (33.1%)**	**514 (40.9%)**	**326 (26.0%)**
LongMemEval-S	Sonnet 4	624	120 (19.2%)	206 (33.0%)	298 (47.8%)
	Opus 4.7	624	122 (19.6%)	203 (32.5%)	299 (47.9%)
	Gemini 3.1	624	136 (21.8%)	227 (36.4%)	261 (41.8%)
	GPT-5.5	622	147 (23.6%)	238 (38.3%)	237 (38.1%)
	DeepSeek V4	622	135 (21.7%)	218 (35.0%)	269 (43.2%)
	Majority	**620**	**132 (21.3%)**	**229 (36.9%)**	**259 (41.8%)**

Table 7: Transfer validation on Mem0 and MemoryOS under Recall@60, MRR, and nDCG@60.

System	Dataset	Target	$\|Q\|$	R@60	MRR	nDCG@60
Mem0	LoCoMo	Raw	1977	0.723	0.158	0.275
		Source	1977	0.945	0.620	0.663
		Canonical	1977	0.901	0.577	0.617
	LongMemEval-S	Raw	470	0.978	0.454	0.567
		Source	470	0.987	0.687	0.717
		Canonical	470	0.758	0.412	0.435
MemoryOS	LoCoMo	Raw	1977	0.770	0.364	0.434
		Source	1977	0.994	0.481	0.579
		Canonical	1977	0.990	0.265	0.409
	LongMemEval-S	Raw	470	0.359	0.207	0.231
		Source	470	0.995	0.789	0.812
		Canonical	470	0.989	0.682	0.730

on LongMemEval-S. The non-supportive rate is 26.0% on LoCoMo and 41.8% on LongMemEval-S. The judge agreement statistics show the same qualitative pattern. Across the 1,876 cases where all five models returned valid labels, mean pairwise binary agreement is 90.9%, and mean three-class agreement is 85.9% (Appendix G Table 14). Even under the most permissive binary collapse, where `supports` and `partial` are both treated as relevant, 31.2% of relaxed credit is entirely unjustified. The partial category, which accounts for 39.6% of cases, requires additional context that may or may not appear elsewhere in the ranked list. Therefore, provenance-relaxed credit cannot be treated as automatically valid semantic evidence.

Target choice also propagates to answer-level metrics. On LoCoMo, Raw best predicts answer F1, while on LongMemEval-S, Source and Canonical align better with the judge-based scorer (Appendix E). This confirms that no single target is universally correct even at the downstream answer level.

5.5 Target Sensitivity Transfers Across Memory Architectures

To test whether the effect extends beyond the primary parser-derived store, we evaluate Mem0 (Chhikara et al., 2025) and MemoryOS (Kang et al., 2025). The transfer runs remain strongly target-sensitive across Recall@60, MRR, and nDCG@60, but the target ordering is not uniform (Table 7). On full LoCoMo runs, Mem0 follows the same ordering across all three metrics, with Source above Canonical and Canonical above Raw. MemoryOS also places Source highest on LoCoMo, but under the rank-sensitive metrics MRR and nDCG@60 it shows the nonmonotone pattern Source above Raw and Raw above Canonical. On full LongMemEval-S runs, MemoryOS is monotone across all three metrics, whereas Mem0 is nonmonotone, with Source above Raw and Raw above Canonical. Unlike the native parser runs, these transfer systems have zero targetless queries for all three targets, so full run values are also shared subset values. These runs rebut a simple inflation story. A more abstract target is not always better, because Raw and Canonical trade places depending on where the relevant detail survives and which metric is used. The effect is therefore not confined to one store design.

As a further stress test on a different retrieval unit and grounding structure, we rescore BEAM's grounded 1M slice (Tavakoli et al., 2026) under one provider. On the full 625 query run, nDCG@60 moves from 0.055 under Raw to 0.097 under Source and 0.197 under Canonical. On the shared subset with $n = 408$, the ordering persists, with $0.071 < 0.134 < 0.197$.

6 Conclusion

Conversational memory benchmarks often evaluate stores that contain both raw dialogue turns and transformed descendants of the same evidence. We show that the credited scoring target is not a minor detail: under fixed ranked outputs, changing only this target alters scores on 83% to 94% of shared queries, changes store design recommendations, and transfers across memory architectures.

The semantic audit further shows that relaxed credit is only partially reliable, since source-linked descendants do not always preserve the evidence needed to answer the query. Benchmark reports should therefore specify the credited target set, use shared-subset rescoring when coverage differs, and flag cases where target choice changes the winner. TIAP makes this sensitivity visible and reproducible.

7 Limitations

This study is about transformed conversational memory, not IR evaluation in general. The native headline matrix is concentrated in one parser-derived store construction. The Mem0 and MemoryOS runs broaden the evidence, but they are still a small architectural sample.

The semantic audit covers all 1,902 disagreement cases across the eight native runs using five frontier LLM judges whose rubric was validated on a 115-case human-annotated subset (Fleiss' $\kappa = 0.83$, 90.9% mean pairwise binary agreement on the full set). The five-model majority vote reduces single-model bias, but the human validation subset is still small (115 cases, one primary annotator with a 36-case adjudicated overlap), so prevalence estimates should be interpreted accordingly.

Canonical coverage is not random. Canonical-covered queries have higher Raw nDCG than uncovered queries in the native runs, so full-set Raw scores are not directly comparable to Canonical scores on the covered subset. This is why the main Raw–Canonical claims use matched canonical-covered query subsets.

TIAP currently treats target membership as binary once a retrieved item falls inside a target set. It does not assign graded credit to partially sufficient evidence or compose support across multiple retrieved facts. This hurts most for queries whose answer is distributed across evidence units: multi-hop questions, multi-session preference questions, temporal comparisons, counts, totals, and other aggregation queries. In those cases, a retrieved item can be provenance-correct but insufficient, or one of several necessary facts. Our semantic audit surfaces exactly this boundary, so binary target membership should be read as an audit baseline rather than a complete model of evidential sufficiency.

The Raw / Source / Canonical basis used here is a minimal audit basis for the source-linked transformed-memory systems we study. It is useful because these systems expose recoverable lineage from transformed memories back to benchmark source evidence. Other memory architectures may expose different stored units and may require different target sets or audit constructions. We therefore treat this basis as an instantiation for this setting, not as a complete taxonomy of conversational memory. The audit also assumes stable source anchors $s(q)$, recoverable lineage from retrieved IDs back to those anchors, enumerable target sets over the descendants, and exported fixed ranked traces. Parser- or extraction-based stores satisfy this directly, which is why our current instantiation covers the native store, Mem0, and MemoryOS. Graph-native memories may require node-, edge-, or path-level target definitions beyond the Raw/Source/Canonical basis used here.

More broadly, TIAP and MTEL-Mem are meant to make target sensitivity auditable, not to prescribe a universal target set or a single pass/fail threshold. Extending the audit to graph-native memories and other retrieval settings remains future work.

8 Availability Statement

An anonymized implementation of MTEL-Mem, including target mappings, rescoring code, example ranked traces, and semantic-audit materials sufficient to reproduce the main analyses in this paper, is available at https://anonymous.4open.science/r/MTEL-Mem-1F71.

9 Ethical Considerations

All experiments use publicly released benchmark datasets (LoCoMo, LongMemEval) and system-generated retrieval traces; no private user data is involved. The primary ethical concern is evaluation transparency: reporting a single scoring-target choice without disclosure can overstate or understate system quality, potentially misleading downstream adoption decisions. Our protocol is designed to surface this ambiguity rather than resolve it in favor of any particular system, and we advocate explicit target reporting to reduce the risk of misleading benchmark claims. More broadly, benchmark-guided memory design choices influence which information deployed conversational systems retain, surface, or suppress for end users, so transparent evaluation directly supports informed system design.

Acknowledgments

The authors used Claude (Sonnet 4.5) and ChatGPT (5.4) to assist with proofreading and improving the clarity of the manuscript text. All scientific content, analysis, and conclusions are entirely the authors' own.

References

Peter Bailey, Nick Craswell, Ian Soboroff, Paul Thomas, Arjen P. de Vries, and Emine Yilmaz. 2008. Relevance assessment: Are judges exchangeable and

does it matter? In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 667–674.

Chris Buckley and Ellen M. Voorhees. 2004. Retrieval evaluation with incomplete information. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 25–32.

Ben Carterette, James Allan, and Ramesh Sitaraman. 2006. Minimal test collections for retrieval evaluation. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 268–275.

Howard Chen, Ramakanth Pasunuru, Jason Weston, and Asli Celikyilmaz. 2023. Walking down the memory maze: Beyond context limit through interactive reading. *arXiv preprint arXiv:2310.05029*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multilinguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. 2025. Mem0: Building production-ready AI agents with scalable long-term memory.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. 2020. Overview of the TREC 2019 deep learning track. *arXiv preprint arXiv:2003.07820*.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. 2025. Memory OS of AI agent.

Sangyeop Kim, Yohan Lee, Sanghwa Kim, Hyunjong Kim, and Sungzoon Cho. 2025. Pre-storage reasoning for episodic memory: Shifting inference burden to memory for personalized dialogue. In *Findings of the Association for Computational Linguistics: EMNLP 2025*.

Kuang-Huei Lee, Xinyun Chen, Hiroki Furuta, John Canny, and Ian Fischer. 2024a. A human-inspired reading agent with gist memory of very long contexts. *arXiv preprint arXiv:2402.09727*.

Sean Lee, Aamir Shakir, Darius Koenig, and Julius Lipp. 2024b. Open source strikes bread – new fluffy embedding model. https://www.mixedbread.ai/blog/mxbai-embed-large-v1.

Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. 2025. Hello again! LLM-powered personalized agent for long-term dialogue. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*.

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 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Kai Tzu-iunn Ong, Namyoung Kim, Minju Gwak, Hyungjoo Chae, Taeyoon Kwon, Yohan Jo, Seungwon Hwang, Dongha Lee, and Jinyoung Yeo. 2025. Towards lifelong dialogue agents via timeline-based memory management. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8631–8661.

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 *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. 2025. Zep: A temporal knowledge graph architecture for agent memory.

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 (EMNLP-IJCNLP)*, pages 3982–3992.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 338–354.

Mark Sanderson and Justin Zobel. 2005. Information retrieval system evaluation: Effort, sensitivity, and reliability. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 162–169.

Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, Anand Rajan Iyer, Tianlong Chen, Huan Liu, Chen-Yu Lee, and Tomas Pfister. 2025. In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8416–8439.

Mohammad Tavakoli, Alireza Salemi, Carrie Ye, Mohamed Abdalla, Hamed Zamani, and J. Ross Mitchell. 2026. Beyond a million tokens: Benchmarking and enhancing long-term memory in LLMs. In *The Fourteenth International Conference on Learning Representations*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

Ellen M. Voorhees. 2001. The philosophy of information retrieval evaluation. In *Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001*, pages 355–370.

Ellen M. Voorhees and Donna K. Harman, editors. 2005. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press.

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

Wei Yang, Kuang Lu, Peilin Yang, and Jimmy Lin. 2019. Critically examining the "neural hype": Weak baselines and the additivity of effectiveness gains from neural ranking models. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1129–1132.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2023. MemoryBank: Enhancing large language models with long-term memory. *arXiv preprint arXiv:2305.10250*.

Justin Zobel. 1998. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 307–314.

Supplementary Material

A Query-Level Instability Details

Figure 3 shows the per-query magnitude of nDCG change when switching from Raw to Canonical targets. The distribution is not concentrated near zero: 65% of LoCoMo and 71% of LongMemEval-S query-provider pairs shift by at least 0.1 nDCG, with mean $|\Delta|$ of 0.25 and 0.33, respectively. Table 16 expands the compact instability summary in Table 4 to the full set of pairwise target comparisons. For each of the eight provider–dataset pairs, it reports the number of shared queries, the count of binary hit flips at k=60, top-1 rank flips, the number of queries whose nDCG changed, and the overall change rate. Raw versus Canonical comparisons show the highest change rates (83.4%–94.0%), while Raw versus Source rates are lower but still substantial (31.3%–66.2%). Source versus Canonical rates cluster between 44.2% and 55.2%, confirming that the effect is not driven solely by the Raw–Canonical gap.

B Aggregation and Mitigation Checks

A natural follow-up is whether a single aggregate score can absorb target sensitivity. We test three aggregation functions over the per-query nDCG triplet $(\mathrm{nDCG}_{Raw}, \mathrm{nDCG}_{Source}, \mathrm{nDCG}_{Canonical})$: arithmetic mean, geometric mean, and per-query minimum. Table 8 reports per-provider nDCG@60

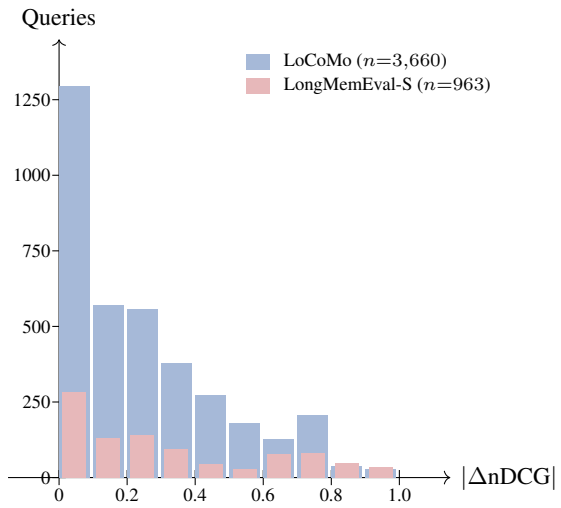

Figure 3: Per-query $|\Delta\mathrm{nDCG}|$ when switching from Raw to Canonical targets, pooled across native providers on shared canonical-covered subsets. 65% of LoCoMo and 71% of LongMemEval-S query–provider pairs shift by ≥ 0.1 nDCG (mean $|\Delta|$: 0.25 and 0.33, respectively).

Table 8: Provider nDCG@60 under individual targets and three ensemble functions across all eight native runs. No ensemble produces a novel consensus ranking. τ_d: Kendall-tau distance from the Raw provider ordering.

Dataset	Method	lexical	allminilm	bge_m3	mxbai	τ_d
LoCoMo	Raw	.176	.198	.240	.197	n/a
	Source	.290	.320	.339	.318	0
	Canonical	.275	.301	.283	.299	2
	Arith. mean	.247	.273	.288	.272	0
	Geom. mean	.202	.228	.240	.227	0
	Min	.148	.166	.175	.165	0
LongMem-Eval-S	Raw	.154	.157	.155	.134	n/a
	Source	.307	.338	.333	.361	3
	Canonical	.344	.391	.384	.430	3
	Arith. mean	.268	.295	.291	.309	3
	Geom. mean	.215	.228	.226	.187	0
	Min	.147	.154	.153	.133	0

Table 9: Supplementary BEAM-1M lexical stress test. The ordering remains Raw < Source < Canonical both on the full run and on the canonical-covered shared subset.

Slice	Q	Raw	Source	Canonical
Full run	625	0.0554	0.0969	0.1969
Shared subset	408	0.0706	0.1342	0.1969

under the three aggregation functions. No aggregation produces a ranking distinct from an existing individual target. On LoCoMo, all three aggregations reproduce the Raw or Source provider ordering. On LongMemEval-S, the aggregation function determines which ordering emerges: geometric mean and minimum reproduce Raw, while arithmetic mean reproduces Source or Canonical. Aggregation therefore selects one target-dependent ordering rather than removing the target choice. Agreement filtering also fails. Restricting evaluation to queries where all three targets agree on binary hit status retains 72% to 85% of queries, but nDCG values still diverge, with mean per-query spreads between 0.16 and 0.30. Provider rankings on LongMemEval-S continue to flip between targets, showing that agreement on binary retrieval does not imply agreement on graded relevance.

C Supplementary BEAM Runs

The main paper uses BEAM as a supplementary stress test. This appendix provides the supporting BEAM details, which are not shown in the main tables. On the full 625-query run, nDCG@60 moves from 0.055 under Raw to 0.097 under Source and 0.197 under Canonical. On the canonical-covered shared subset with $n = 408$, the same ordering persists. Table 10 reports supporting BEAM-100K runs. Both the lexical and mxbai settings preserve

Table 10: Supporting BEAM-100K runs with consistent Raw < Source < Canonical ordering.

Scale	Setting	Q	Raw	Source	Canonical
100K	Lexical full	355	0.1105	0.2064	0.3489
100K	Lexical shared subset	246	0.1430	0.2815	0.3489
100K	mxbai full	355	0.1233	0.2281	0.3898
100K	mxbai shared subset	245	0.1617	0.3135	0.3898

Table 11: Representative LongMemEval-S all-MiniLM disagreement cases. Scores are nDCG.

Query	Raw	Src.	Can.
What degree did I graduate with?	0.000	0.613	1.000
How long is my daily commute to work?	0.173	1.000	1.000
What play did I attend at the local community theater?	0.000	0.500	0.500
How long did I wait for the decision on my asylum application?	—	0.171	0.000

the same Raw < Source < Canonical ordering on the full run and on the canonical-covered shared subset.

D Qualitative Disagreement Cases

Table 11 lists representative disagreement cases from LongMemEval-S with all-MiniLM. These examples illustrate how the same saved ranked output can receive different scores depending on whether credit is assigned to the raw source turn, a source-linked descendant, or a canonical serving memory.

E Auxiliary Answer-Level Alignment

This appendix provides additional detail for the answer-alignment analysis referenced in Section 5.4. Table 12 reports the disagreement-cell counts and mean answer F1. We join target-specific hit labels to answer F1 on the same saved traces. On LoCoMo, Raw has the strongest average answer-alignment signal across the eight native runs. On LongMemEval-S, Source and Canonical align better, though on a thinner base of four native runs.

This difference is consistent with the benchmarks' answer scorers: LoCoMo's lexical scoring rewards exact answer wording more directly, whereas LongMemEval-S's judge-based scoring is more tolerant of paraphrased or transformed evidence. This analysis does not replace the paper's main result. Instead, it provides a downstream check showing that answer-level alignment does not identify one universally preferred target set.

F Semantic Audit Agreement

To validate the semantic audit rubric before scaling, we ran four frontier LLMs with extended reasoning on a 115-case validation subset: `gpt-5.5`, `claude-opus-4.7`, `gemini-3.1-pro`, and `deepseek-v4-pro`. These labels were compared with the human validation labels. Each model received the same rubric and case material and returned one of {*supports*, *partial*, *does_not_support*}. The goal is to test rubric reproducibility rather than model accuracy.

Table 12: Secondary query-level answer-alignment check on shared Raw-versus-Canonical eligible queries. 'Raw only' and 'Canonical only' denote disagreement cells where hit@60 changes under identical ranked outputs. Means are averaged over native full runs within each dataset; counts are summed across those runs.

Dataset	Raw-only n	Canon-only n	Raw-only F1	Canon-only F1	Raw-only strong	Canon-only strong
LoCoMo	507	1142	0.123	0.046	0.063	0.013
LongMemEval-S	32	312	0.030	0.065	0.000	0.004

Table 13: Inter-rater agreement on the 115-case validation subset used to confirm rubric consistency before scaling to the full 1,902-case audit. Binary labels collapse *supports* ∪ *partial* → *relevant*. Fleiss' κ is computed over all five raters; pairwise values are Cohen's κ.

	Binary		Three-class	
Rater / pair	κ	Agree	κ	Agree
All five raters (Fleiss)				
1 human + 4 LLMs	**0.83**	n/a	0.79	n/a
Each model vs. human (Cohen)				
GPT-5.5	0.87	93.9%	0.83	88.7%
Claude Opus 4.7	0.85	93.0%	0.83	88.7%
Gemini 3.1 Pro	0.82	91.3%	0.74	82.6%
DeepSeek V4 Pro	0.77	88.7%	0.73	81.7%
Model–model pairs (Cohen, binary)				
Gemini ↔ GPT	0.91	95.7%	n/a	n/a
Claude ↔ GPT	0.91	95.7%	n/a	n/a
Claude ↔ Gemini	0.89	94.8%	n/a	n/a
Claude ↔ DeepSeek	0.77	88.7%	n/a	n/a
GPT ↔ DeepSeek	0.75	87.8%	n/a	n/a
Gemini ↔ DeepSeek	0.74	87.0%	n/a	n/a

Table 13 reports Fleiss' κ across the five raters and pairwise Cohen's κ for both the original three-class and collapsed binary schemes. All four models reach substantial to almost perfect consistency with the human labels.

G Full-Audit Multi-Model Agreement

After validating the rubric, we ran all five LLMs on the complete 1,902-case audit. Of these cases, 1,876 received valid labels from all five models; 26 cases returned empty responses from GPT-5.5 or DeepSeek V4 Pro after retries. Table 14 reports pairwise agreement on the 1,876 common-valid cases.

Mean binary agreement across all ten model pairs is 90.9%, and mean three-class agreement

Table 14: Pairwise agreement among five LLM judges on the full 1,902-case semantic audit ($n = 1{,}876$ cases where all five returned valid labels). Binary collapses *supports* $\cup$ *partial* $\rightarrow$ *relevant*.

Model pair	3-class agree	Binary agree
Sonnet 4 $\leftrightarrow$ Opus 4.7	85.9%	90.0%
Sonnet 4 $\leftrightarrow$ Gemini 3.1	85.8%	91.3%
Sonnet 4 $\leftrightarrow$ GPT-5.5	84.7%	89.4%
Sonnet 4 $\leftrightarrow$ DeepSeek V4	83.9%	88.4%
Opus 4.7 $\leftrightarrow$ Gemini 3.1	89.9%	94.2%
Opus 4.7 $\leftrightarrow$ GPT-5.5	86.1%	91.3%
Opus 4.7 $\leftrightarrow$ DeepSeek V4	83.4%	88.9%
Gemini 3.1 $\leftrightarrow$ GPT-5.5	89.9%	94.1%
Gemini 3.1 $\leftrightarrow$ DeepSeek V4	83.6%	90.0%
GPT-5.5 $\leftrightarrow$ DeepSeek V4	86.4%	90.9%
Mean (all 10 pairs)	85.9%	90.9%

is 85.9%. The five-model majority vote matches the original Sonnet 4 labels on 90.1% of cases, reducing concern that the semantic audit is driven by a single model. The majority-vote distribution is 29.2% supports, 39.6% partial, and 31.2% does not support. The dataset-level gap, approximately 33% fully supportive on LoCoMo versus 21% on LongMemEval-S, is reproduced by all five judges.

H Reporting-Practice Audit

Table 15 audits whether 11 recent conversational-memory papers publicly specify the components needed for a scoring-target ontology. We inspect only what each paper states or releases; we do not claim the authors lacked internal conventions.

Scope. We selected papers cited in our study that either propose a conversational-memory benchmark or propose a memory system evaluated on retrieval or QA tasks. For each paper, we checked four questions:

1. **Transforms memory?** Does the system or benchmark produce non-raw representations, such as extracted facts, summaries, or knowledge-graph entries, from raw conversation turns?

2. **Reports retrieval unit?** Is the scored unit formally defined?

3. **Reports target mapping?** Is there a documented mapping from scored items to source conversation evidence?

4. **Specifies descendant credit?** When one source turn produces multiple stored items, does the paper specify which descendants receive credit?

Table 15: Reporting-practice audit of target specification in 11 conversational-memory papers. All transform memory; none fully specifies which stored descendants receive scoring credit. ● = explicit, ○ = partial, × = not specified.

Paper	Trans.	Unit	Map	Credit
Benchmarks				
LoCoMo (Maharana et al., 2024)	●	●	○	×
LongMemEval-S (Wu et al., 2025)	●	○	○	×
BEAM (Tavakoli et al., 2026)	●	○	○	×
Memory systems				
Mem0 (Chhikara et al., 2025)	●	○	×	×
MemoryOS (Kang et al., 2025)	●	○	×	×
Zep (Rasmussen et al., 2025)	●	○	×	×
Theanine (Ong et al., 2025)	●	○	×	×
RMM (Tan et al., 2025)	●	○	×	×
MemGPT (Packer et al., 2023)	●	○	×	×
MemoryBank (Zhong et al., 2023)	●	○	×	×
LD-Agent (Li et al., 2025)	●	○	×	×

Findings. All 11 papers create transformed memory representations. Retrieval units are described at the architectural level in most papers, but only LoCoMo explicitly labels and separately evaluates distinct unit types. No paper fully specifies a target mapping that traces credit through the transformation layer to source evidence. No paper addresses credit assignment when a single source turn produces multiple stored descendants.

Three benchmarks provide partial provenance. LoCoMo annotates turn IDs on QA gold answers (Maharana et al., 2024). LongMemEval-S annotates evidence sessions and positions within them (Wu et al., 2025). BEAM includes a `source_chat_ids` field in its question schema, although some published examples leave this field empty (Tavakoli et al., 2026). Two system papers touch on the gap indirectly. Ong et al. (2025) observe that existing long-term conversation datasets lack gold mappings between dialogue contexts and memories. Rasmussen et al. (2025) maintain episodic edges linking extracted facts to source episodes but state that these connections are not examined in their experiments. This audit motivates our recommendation: transformed-memory evaluations should state the scored target set and release target mappings alongside retrieval traces.

Table 16: Query-level instability between target definitions across all eight main-matrix provider-dataset pairs.

Dataset	Provider	Pair	Shared queries	Hit flips	Top-1 flips	nDCG changed	Change rate
LoCoMo	Lexical	Raw vs Source	1,533	181	173	526	34.3%
		Raw vs Canonical	899	244	193	758	84.3%
		Source vs Canonical	899	63	20	397	44.2%
	all-MiniLM	Raw vs Source	1,533	162	198	541	35.3%
		Raw vs Canonical	899	226	217	794	88.3%
		Source vs Canonical	899	64	19	425	47.3%
	BGE-M3	Raw vs Source	1,510	134	167	473	31.3%
		Raw vs Canonical	963	259	214	851	88.4%
		Source vs Canonical	963	125	47	532	55.2%
	mxbai-embed-large	Raw vs Source	1,533	162	190	541	35.3%
		Raw vs Canonical	899	226	208	794	88.3%
		Source vs Canonical	899	64	18	425	47.3%
LongMemEval-S	Lexical	Raw vs Source	470	50	113	259	55.1%
		Raw vs Canonical	332	68	113	277	83.4%
		Source vs Canonical	332	18	0	158	47.6%
	all-MiniLM	Raw vs Source	470	70	129	301	64.0%
		Raw vs Canonical	332	78	129	309	93.1%
		Source vs Canonical	332	8	0	158	47.6%
	BGE-M3	Raw vs Source	439	62	112	276	62.9%
		Raw vs Canonical	299	67	112	281	94.0%
		Source vs Canonical	299	5	0	144	48.2%
	mxbai-embed-large	Raw vs Source	470	130	152	311	66.2%
		Raw vs Canonical	332	131	152	312	94.0%
		Source vs Canonical	332	1	0	154	46.4%