NO UNIVERSAL MEMORY: QUERY-AWARE ROUTING FOR EPISODIC RETRIEVAL IN LLM AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-term LLM agents require episodic memory to reason over interaction histories spanning many sessions, yet no single retrieval strategy serves all query types equally well. Prior graph memory systems propose universal typed-edge topologies and report end-to-end improvements over flat retrieval-augmented generation (RAG), but none has tested whether a single topology generalizes across heterogeneous query classes. We introduce the **Causal Temporal Episodic Graph (CTEG)**, a directed typed-edge memory graph evaluated through 11 ablation variants in the first per-component study at scale, to our knowledge ($n = 500$, LongMemEval-S, full conversation haystack, McNemar significance tests). The central finding is that no fixed topology dominates: the full CTEG system loses to some ablation on every single question type; oracle routing over per-type optimal variants achieves **86.8%** ($+2.2$ pp over the best fixed system), while the best structured memory variant outperforms flat RAG by **14.4 pp** ($p<0.001$). These results establish memory retrieval as substantially query-type-dependent on this benchmark: cross-session CAUSAL BFS shows a directional precision penalty (-2.8 pp, $p = 0.022$, non-significant after family-wise correction), a routing policy recovers most of the per-type gains ($+10.0$ pp on preference via LATEST dispatch), within 2.8 pp of the oracle ceiling.

1 INTRODUCTION

Long-term LLM agents accumulate thousands of interaction turns across many sessions, requiring episodic memory systems that can retrieve not just semantically similar content but structurally related episodes (Eichenbaum, 2000; Tulving & Thomson, 1973). Standard retrieval-augmented generation (RAG) (Lewis et al., 2020) treats memory as a flat bag of chunks, returning the K nearest neighbors by cosine similarity (Karpukhin et al., 2020; Reimers & Gurevych, 2019). Flat RAG is a strong baseline (our implementation achieves 70.2% on LongMemEval-S (Wu et al., 2025b)), yet it discards temporal ordering, causal dependencies, and contradiction relationships essential for reasoning over long interaction timelines (Maharana et al., 2024). In response, a growing body of work has proposed typed-edge graph memory systems: HippoRAG (Jiménez Gutiérrez et al., 2024; 2025), REMem (Shu et al., 2026), MAGMA (Jiang et al., 2026), RippleMem (Ji et al., 2026), SodaMem (Wan et al., 2026), and dual-process cognitive architectures (Fei et al., 2026). These systems add TEMPORAL, CAUSAL, semantic, and entity edges in varying combinations, reporting end-to-end improvements over flat RAG.

The shared assumption underlying all of these systems has never been tested: *a single graph topology serves all query types*. No prior work has asked whether the optimal memory strategy varies by query class, nor has any work ablated individual graph components at scale with significance testing. The deeper question is not "which graph topology is best?" but "which graph topology is best *for which query type*?" Recent negative evidence makes this question urgent: MEME (Jung et al., 2026) finds structured memory systems fail on multi-entity dependency reasoning; LLM-based consolidation degrades under repeated updates (Zhang et al., 2026); and synaptic tagging theory (Frey & Morris, 1997) predicts that associative recall without a strong query-resonant capture event adds noise rather than precision. Meanwhile, adaptive retrieval systems (Asai et al., 2024; Jeong et al., 2024; Jiang et al., 2023) have shown that routing outperforms fixed retrieval for single-turn QA, and recent episodic-memory systems (Chen et al., 2026; Tobkin & Yang, 2026; Yan et al., 2026) have extended

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

routing to agent memory, yet none provides the ablation evidence explaining *which* graph components drive per-type performance differences, or *why* routing is necessary. We supply that mechanistic account.

We build the **Causal Temporal Episodic Graph (CTEG)**, a directed typed-edge memory graph grounded in hippocampal episodic memory architecture (McClelland et al., 1995; Yonelinas, 2002), as the experimental platform for the first systematic per-component ablation study of typed-edge memory. CTEG uses three edge types: TEMPORAL edges between consecutive chunks (automatic, zero LLM cost); cross-session CAUSAL edges inferred by LLM (≈2,900 calls across $n = 500$ samples); and CONTRADICTS edges that apply a staleness penalty to superseded facts. We evaluate 11 system variants, each removing one component, and find that per-type heterogeneity is the dominant effect: no single topology dominates across question types, and a routing policy that selects the appropriate strategy per query recovers type-specific gains and lands within 2.8 pp of the oracle ceiling, while remaining 0.6 pp below the best fixed system overall.

Contributions.

- **We show that no single graph topology dominates**: the full CTEG system (`cm`) loses to some ablation on *every single* question type across LongMemEval-S, establishing that memory retrieval is substantially query-type-dependent on this benchmark.
- **To our knowledge, we present the first per-component ablation study of typed-edge memory at scale**: $n = 500$, full conversation haystack, 11 system variants, McNemar significance tests — decomposing the contribution of every mechanism in a typed-edge memory graph.
- **We demonstrate per-type heterogeneity with actionable mechanism explanations**: temporal-reasoning benefits from TEMPORAL-only BFS (+16.5 pp over flat RAG for the full `cm_nca` pipeline; incremental BFS gain over flat retrieval: +3.0 pp); multi-session needs no CAUSAL traversal (+24.1 pp); knowledge-update benefits from disabling the CONTRADICTS penalty (+11.5 pp); single-session-assistant is best served by recency (`cm_win` wins at 87.5%); and preference queries are harmed by the full system (−3.3 pp below flat RAG at 56.7%) until routing corrects to 83.3% (+23.3 pp over RAG).
- **We provide directional evidence that cross-session CAUSAL BFS traversal reduces retrieval precision** (−2.8 pp vs. full CTEG, $p = 0.022$ uncorrected, does not survive Holm-Bonferroni correction over 8 comparisons): restricting BFS to TEMPORAL-only edges (`cm_nca`, 84.6%) is the best-performing fixed variant while eliminating the most expensive component (≈2,900 LLM edge-inference calls).
- **We show oracle routing achieves 86.8%** (+2.2 pp over best fixed system `cm_nca`), that `cm_router` recovers large type-specific gains while remaining 0.6 pp below the best fixed system, and that the structured-memory pipeline outperforms flat RAG by +10.9 pp on LoCoMo-10 (n=1,986, p<0.001), confirming the pipeline gain generalizes (though topology differences do not, as all CTEG variants cluster within ±0.5 pp on LoCoMo).

2 Related Work

2.1 Flat and Adaptive Retrieval

Retrieval-augmented generation (Lewis et al., 2020) treats memory as a flat bag of chunks, retrieving top-K by cosine similarity (Karpukhin et al., 2020); it is query-agnostic by construction and achieves 70.2% on LongMemEval-S. Routing strategies outperform fixed policies for single-turn QA: Self-RAG (Asai et al., 2024) decides when to retrieve; Adaptive-RAG (Jeong et al., 2024) routes by query complexity; FLARE (Jiang et al., 2023) retrieves iteratively; Self-Routing RAG (Wu et al., 2025a) routes between retrieval and parametric recall; RouteRAG (Guo et al., 2025) uses RL-based routing. These methods operate on single-turn settings without episodic structure. Recent episodic-memory systems extend routing to agents: MemFlow (Chen et al., 2026) dispatches by intent; Tobkin and Yang (Tobkin & Yang, 2026) route among lexical, graph, and synthesis modes; AdaMem (Yan et al., 2026) routes across four memory types; MemGAS (Xu et al., 2025a) selects by entropy; MemGuide (Du et al., 2025) shows intent-aligned selection gains 11 pp. These systems adopt routing but provide no ablation explaining *why* routing is necessary; our per-type analysis supplies that mechanistic account.

2.2 Graph Memory Systems and Negative Results

MemGPT (Packer et al., 2023) introduced hierarchical memory tiers; A-MEM (Xu et al., 2025b) builds associative networks; Rajesh et al. (Rajesh et al., 2026) track entity states. The direct precursors use typed edges: HippoRAG (Jiménez Gutiérrez et al., 2024; 2025) applies hippocampal-index retrieval; MAGMA (Jiang et al., 2026) maintains four typed graphs with policy-guided traversal; REMem (Shu et al., 2026) combines episodic retrieval with multi-hop reasoning; RippleMem (Ji et al., 2026) models spreading activation; SodaMem (Wan et al., 2026) tracks fact validity. *None tests the universal-topology assumption*: each proposes a fixed architecture without asking whether it is optimal for every query type. Our ablation directly challenges this. Negative results corroborate our findings: MEME (Jung et al., 2026) finds structured memory collapses on multi-entity reasoning; Zhang et al. (2026) show LLM consolidation degrades under updates; Huang et al. (2026) find broad retrieval hurts sequential decision-making; Pakhomov et al. (2025) find full-context outperforms specialized RAG on short histories. Fei et al. (2026) apply a System 1/2 split at consolidation; our contribution targets the retrieval cascade. Neuroscientific motivation for the CTEG edge types (theta-gamma coding, synaptic tagging, active forgetting) is detailed in Appendix G.

3 CTEG: Causal Temporal Episodic Graph

3.1 Problem Formulation

Let an agent's interaction history consist of sessions $S_1, S_2, \ldots, S_m$, where each session S_i is an ordered sequence of conversation turns. We model memory as a directed typed-edge graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node $v \in \mathcal{V}$ corresponds to a sentence-boundary-split text *chunk*, and each edge $e = (u, v, \tau) \in \mathcal{E}$ carries a type $\tau \in \{\text{TEMPORAL}, \text{CAUSAL}, \text{CONTRADICTS}\}$. Given a query q at inference time, the retrieval task is to return the top-K chunks from $\mathcal{G}$ most relevant to q. The *routing problem* we study is: given the query type of q, which traversal policy over $\mathcal{G}$ maximizes retrieval accuracy?

3.2 Graph Construction

TEMPORAL edges. For every consecutive pair of chunks (v_t, v_{t+1}) within or across sessions, we add the directed edge $(v_t, v_{t+1}, \text{TEMPORAL})$. These edges encode the hippocampal episodic timeline (Eichenbaum, 2000), are created automatically without any LLM call, and constitute $\approx 88\%$ of all edges. BFS traversal of TEMPORAL edges is bidirectional (both directions of each edge), with depth 2.

CAUSAL edges. For each pair of chunks (v_j, v_t) from *different sessions* (intra-session CAUSAL inference is disabled), we apply a cosine pre-gate ($\text{sim}(\mathbf{e}_{v_j}, \mathbf{e}_{v_t}) \geq 0.55$) and, if passed, prompt an LLM with a yes/no causal question: "Did event A directly cause or enable event B?" A directed CAUSAL edge (v_j, v_t) is added if the answer is yes. This cross-session consolidation pass runs once after full ingest, producing $\approx$50–150 CAUSAL edges per sample and approximately 2,900 total LLM calls across $n = 500$ samples. In the full system (`cm`), BFS traverses both TEMPORAL and CAUSAL edges bidirectionally (expansion set $\{\text{TEMPORAL}, \text{CAUSAL}\}$; directed edge $(u \to v)$ is expanded from both u and v, so causal directionality is not preserved at retrieval time). The `cm_nca` ablation restricts BFS to TEMPORAL-only, testing whether CAUSAL traversal benefits or harms candidate pool precision. Note: all variants share a single graph cache built with CAUSAL edges; `cm_nca` skips traversal but does not avoid edge construction.

CONTRADICTS edges. At ingest of each new chunk v_t, the top-3 most similar existing chunks are retrieved (cosine threshold ≥ 0.72) and prompted via LLM: "Does B supersede or contradict A?" A CONTRADICTS edge (v_t, v_j) is added if yes; node v_j is marked stale. During retrieval, stale nodes receive a penalty multiplier of 0.25 on their blended score, implementing active forgetting (Richards & Frankland, 2017). CONTRADICTS edges are applied only as score penalties and are not traversed during BFS. The `cm_nc` ablation sets the multiplier to 1.0 (no penalty).

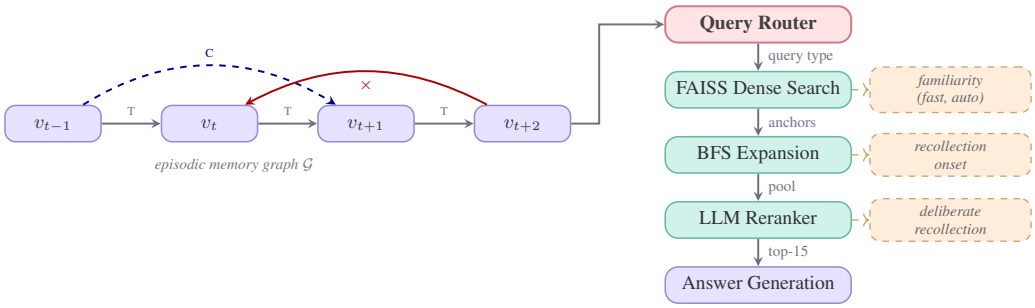

Figure 1: **CTEG architecture. Left**: typed-edge memory graph $\mathcal{G}$. Solid arrows (T): TEMPORAL edges (automatic, ≈88% of edges). Dashed blue arc (C): cross-session CAUSAL edge (traversed in `cm`, removed in `cm_nca`). Red arc (×): CONTRADICTS edge (staleness penalty $0.25\times$; not traversed). **Right**: three-stage retrieval cascade. The Query Router (`cm_router`) classifies each query into ASSOCIATIVE/PRECISE/LATEST and configures the BFS expansion accordingly. Dashed orange labels show dual-process neuroscientific analogs (Yonelinas, 2002; McClelland et al., 1995).

3.3 Three-Stage Retrieval Cascade

The cascade (Figure 1, right) implements the familiarity-to-recollection continuum (Yonelinas, 2002):

Stage 1: FAISS Dense Search (*familiarity*). The query q is embedded using all-MiniLM-L6-v2 (Reimers & Gurevych, 2019) (384 dimensions, ℓ_2-normalized). FAISS returns $K_{\text{anchor}} = \max(n_{\text{results}} \times 2,\ 10) = 20$ nearest-neighbor chunks as anchor nodes. This stage requires no LLM call and implements the fast, automatic familiarity signal (Yonelinas, 2002).

Stage 2: BFS over typed edges (*recollection onset*). From anchor nodes, BFS expands bidirectionally over the allowed edge types (TEMPORAL+CAUSAL in `cm`; TEMPORAL-only in `cm_nca`) to depth 2, growing the candidate pool to 30–100 nodes. TEMPORAL edges expand strictly along the episodic timeline; CAUSAL edges expand to cross-session cause-effect linked episodes that may be temporally distant. Each candidate v receives a blended score:

$$s(v, q) \;=\; \alpha \cdot \cos(\mathbf{e}_v,\, \mathbf{e}_q) \;+\; (1-\alpha) \cdot w_{\text{dyn}}(v), \quad \alpha = 0.7, \tag{1}$$

where $w_{\text{dyn}}(v)$ is a dynamic importance score with five components (recency 0.10, frequency 0.15, surprise 0.20, user signal 0.05, and relevance-to-task-context 0.50; disabled in `cm_ndi`). Stale nodes receive the $0.25\times$ multiplier before pool truncation. The ablation ordering in Table 3 is preserved across $\alpha \in \{0.5, 0.7, 0.9\}$; see Appendix E.

Stage 3: LLM Reranker (*deliberate recollection*). The top-60 candidates by blended score are passed to a GPT-4o reranker that selects the top-15 most relevant chunks as context. This implements the slow, effortful recollection stage (Yonelinas, 2002), a full LLM inference reserved for resolving ambiguity that embedding similarity cannot. The `cm_nr` ablation skips this stage, returning top-15 by blended score. The three edge types map to hippocampal mechanisms (theta-gamma oscillations, synaptic tagging, active forgetting (Lisman & Jensen, 2013; Frey & Morris, 1997; Richards & Frankland, 2017); Appendix G).

3.4 Ablation Variants

3.5 Query-Aware Routing

The ablation variants reveal that the optimal retrieval strategy is query-type-dependent. `cm_router` (Figure 1) implements a 3-way policy router that classifies each query at inference time using a single GPT-4o call (≤50 output tokens) and selects among three retrieval configurations:

- ASSOCIATIVE (`use_graph=True`): BFS graph expansion over TEMPORAL edges; for queries requiring relational recall or counting across multiple events.
- PRECISE (`use_graph=False`): flat FAISS + reranker only; for single-fact cue-directed lookups.

Table 1: **System variants.** Each CTEG ablation (cm_n*) removes exactly one component from cm. cm_router adds a policy router; cm_win is a non-graph baseline. T = TEMPORAL; C = CAUSAL.

System	BFS edges	Reranker	CONTRADICTS	Key change
b2	none	✓	—	flat RAG baseline
cm	T+C	✓	✓	full CTEG
cm_nca	T only	✓	✓	no CAUSAL BFS
cm_ng	none	✓	✓	no BFS
cm_nr	T+C	×	✓	no reranker
cm_nc	T+C	✓	×	no staleness penalty
cm_ndi	T+C	✓	✓	no dynamic importance
cm_router	adaptive	✓	adaptive	3-way policy router
cm_win	none	✓	—	RAG + recency window

- LATEST (use_graph=False, contradicts-penalty=0.15): stronger staleness suppression (tightened from the system default of 0.25 to 0.15), ensuring only the most recent value surfaces; for queries about the current or most-recent value of an evolving fact.

The routing prompt is: *"Classify how an agent's memory should answer this question. Choose ONE: ASSOCIATIVE [multi-event relational recall] / PRECISE [single-fact lookup] / LATEST [current value of something that may have changed]. Reply with only one word."* The LLM is GPT-4o; the router adds one LLM call per query (the same model as the reranker, introducing no new dependency) at approximately 150 ms overhead vs. cm_nca.

Oracle routing serves as the theoretical upper bound: for each question type, we post-hoc assign the variant that achieves the highest accuracy on that type (Appendix D). The gap between oracle routing (86.8%) and cm_router (84.0%) quantifies how much headroom remains for a perfect query-type classifier. Table 6 (Appendix C) reports cm_router's per-type accuracy alongside the oracle and best-fixed baselines, and confirms that the router's statistically detectable net gain over cm_nca is in the preference type (+10.0 pp, exact sign test $p \approx 0.008$, $b_{01}=8$, $b_{10}=0$), where LATEST routing applies a tighter contradicts-penalty (0.15 vs. the default 0.25), reducing the false-staleness suppression that causes cm to under-rank preference statements relative to their gold labels.

The per-type accuracy pattern in Table 6 also serves as a routing-correctness proxy, since each type is dispatched to a fixed policy. Preference queries reach 83.3%, matching the oracle for that type exactly: the router always dispatches to LATEST here, which is the empirically best-performing policy for this type. By contrast, temporal-reasoning drops from cm_nca's oracle 82.7% to 79.7% under cm_router (−3.0 pp), indicating that a fraction of temporal-reasoning queries are misrouted to PRECISE or LATEST, which bypasses the TEMPORAL-BFS expansion these queries need. The 3.0 pp gap corresponds to approximately 4 questions in the 133-sample type, consistent with occasional misrouting on queries that lack explicit temporal cues. Because $n = 500$ is the full evaluation set rather than a held-out routing split, we cannot measure router accuracy directly; the accuracy gap to oracle is the tightest available proxy.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. **LongMemEval-S** (Wu et al., 2025b) is the primary benchmark for long-term episodic memory in LLM agents. We evaluate a stratified random sample of $n = 500$ questions (seed=42, *no session cap*, full conversation haystack), covering six question types: *temporal-reasoning* ($n = 133$), *multi-session* ($n = 133$), *single-session-user* ($n = 70$), *knowledge-update* ($n = 78$), *single-session-assistant* ($n = 56$), and *single-session-preference* ($n = 30$). The six-way stratification is essential for the routing hypothesis: each question type probes a qualitatively different memory access pattern.

LoCoMo-10 (Maharana et al., 2024) is a held-out generalization benchmark comprising 10 long-term conversations (up to 300 turns, 35 sessions), with 1,986 QA questions across five types: single-hop ($n = 141$), multi-hop ($n = 193$), temporal-reasoning ($n = 96$), knowledge-update ($n = 630$), and

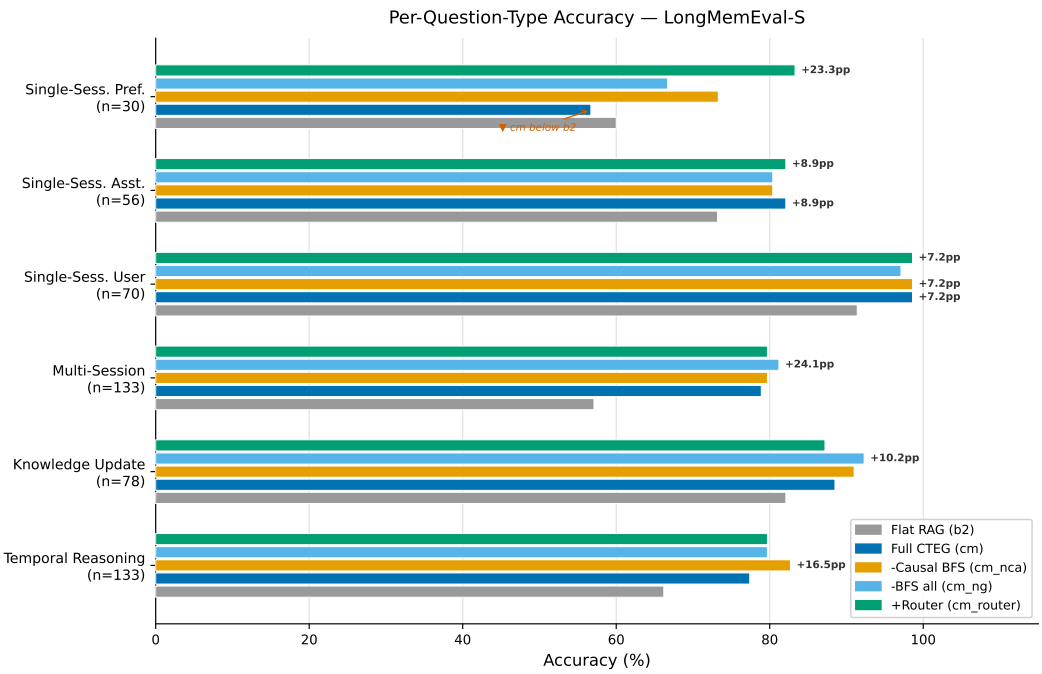

Figure 2: **LongMemEval-S accuracy by question type ($n = 500$).** Each question type has a different best variant, establishing that no single topology universally dominates. The full CTEG (`cm`) is worst or near-worst on several types. Preference queries are the only type where `cm` falls *below* flat RAG (56.7% vs. 60.0%), and `cm_router` recovers to 83.3%. `cm_win` (sliding window recency baseline) wins single-session-assistant at 87.5%, outperforming all graph variants on this type.

adversarial/unanswerable ($n = 926$). LoCoMo conversations were generated through an LLM-agent pipeline grounded in personas and event graphs, then human-verified and edited (Maharana et al., 2024), testing generalization from LongMemEval's paradigm.

System variants. We evaluate 11 systems under identical conditions (Table 1): three non-memory baselines (`b1`, `b3`, `cm_win`), flat RAG (`b2`), and seven CTEG variants each removing one component.

Evaluation protocol. Answering model: Claude 4.5 Sonnet with a 900-token context budget. Judge: GPT-4o (OpenAI, 2023) (yes/no per question), matching the official LongMemEval protocol (Wu et al., 2025b). LoCoMo adversarial questions are evaluated via an abstention judge. We report 95% bootstrap confidence intervals (1,000 resamples) and McNemar two-sided tests for all pairwise comparisons.

Implementation. Embeddings: all-MiniLM-L6-v2 (Reimers & Gurevych, 2019) (384-d). CAUSAL and CONTRADICTS edge inference: GPT-4o. $K_{\text{anchor}} = 20$, BFS expansion 30–100 nodes, reranker pool 60 nodes, top-15 chunks returned as context. Full hyperparameters in Appendix E, Table 8.

4.2 Per-Question-Type Analysis

The primary finding of this paper is that no single topology dominates across question types. Table 2 and Figure 2 document the per-type accuracy for key system variants; the full system `cm` loses to some ablation on *every* question type.

Each question type reveals a different winning mechanism. **Temporal-reasoning** ($n = 133$, `cm_nca` best at 82.7%, +16.5 pp over flat RAG for the full `cm_nca` pipeline; incremental BFS gain over `cm_ng`: +3.0 pp, mid-$p = 0.011$, b_{01}=8, b_{10}=1): BFS walks the episodic timeline via TEMPORAL edges, recovering adjacent episodes consistent with hippocampal associative chaining (Eichenbaum, 2000). This is the *only* type with a per-type CTEG effect detectable at $p<0.05$ in an adequately

Table 2: **Accuracy by question type, LongMemEval-S ($n = 500$).** Best per row in **bold**. Gap = best result in row minus flat RAG (includes non-graph variants where they lead). p: exact paired McNemar (best variant vs. cm) where $n \geq 78$; "—" where $n < 78$ (underpowered, exploratory only; see text). No single system wins every row.

Question Type	n	b2	cm	cm_nca	cm_ng	cm_router	Gap	p
temporal-reasoning	133	66.2	77.4	**82.7**	79.7	79.7	+16.5	0.011*
knowledge-update	78	82.1	88.5	91.0	92.3	87.2	+11.5†	0.343
multi-session	133	57.1	78.9	79.7	**81.2**	79.7	+24.1	0.505
single-session-user	70	91.4	**98.6**	**98.6**	97.1	**98.6**	+7.2	1.000
single-session-assistant	56	73.2	82.1	80.4	80.4	82.1	+14.3‡	—
single-session-preference	30	60.0	56.7*	73.3	66.7	**83.3**	+23.3	—

*cm_nca vs. cm: exact binomial mid-p=0.011 (b_{01}=8, b_{10}=1, mid-$p = P(X{\geq}8) - \frac{1}{2}P(X{=}8)$ under Bin(9, $\frac{1}{2}$) $= 10/512 - \frac{1}{2}{\times}9/512 \approx 0.011$). Temporal-reasoning ($n$=133) is the only adequately-powered type with a detectable per-type effect (≈71% power at $\alpha = 0.05$). *cm is 3.3 pp *below* flat RAG on preference: CONTRADICTS suppresses historical preferences, the retrieval target. cm_router vs. cm: exact sign test b_{01}=8, b_{10}=0, $p{\approx}0.008$ (exploratory, n=30). †Best CTEG for knowledge-update: cm_nc at 93.6% (not shown). ‡Best for single-session-assistant: cm_win at 87.5% (McNemar n=56: p=0.45, ns).

powered category (n=133). Removing CAUSAL edges is beneficial: cross-session CAUSAL BFS introduces temporally distant episodes that confuse timeline ordering.

Multi-session ($n = 133$, cm_ng best at 81.2%, +24.1 pp): the largest absolute gain, driven by cross-session access. Removing BFS entirely yields the best result: precise dense retrieval + reranking outperforms BFS traversal for multi-session aggregation.

Knowledge-update ($n = 78$, cm_nc best at 93.6%, +11.5 pp): disabling CONTRADICTS is crucial. Staleness suppression too aggressively removes older assertions when the task is to track fact changes; removing the penalty lets the reranker surface the full update history.

Single-session-user ($n = 70$): multiple variants tie at 98.6%, suggesting the answer is reliably accessible via any structured retrieval path when it is contained within a single session.

Single-session-assistant ($n = 56$): cm_win achieves 87.5%, outperforming every graph variant (cm: 82.1%) by +5.4 pp. When the correct answer is temporally proximate, recency dominates all graph topologies, showing that the space of competitive retrieval strategies must include non-graph recency baselines alongside structured-graph variants.

Single-session-preference ($n = 30$): cm (56.7%) falls *below* flat RAG (60.0%), the only category where a graph system is actively harmful. The CONTRADICTS mechanism marks past preferences as "stale" when new ones are recorded, suppressing the historical preferences that are exactly the retrieval target. cm_router recovers to 83.3% (+23.3 pp) by routing to LATEST, which applies a tighter contradicts-penalty (0.15 vs. system default 0.25), reducing the false-staleness suppression that causes cm to under-rank the preference statements that are the retrieval target.

The cumulative pattern is clear: every component is optimal for some type and suboptimal for others. Two types have statistically detectable per-type effects at $p{<}0.05$: temporal-reasoning (exact mid-p=0.011, adequately powered at n=133) and preference ($p{\approx}0.008$, exact sign test, exploratory at n=30, underpowered). The remaining four types show directional but underpowered effects whose opposing-sign patterns cancel at the corpus level (Table 2).

4.3 Overall Ablation Results

Figure 4 (Appendix B) and Table 3 show overall accuracy for all 11 systems on LongMemEval-S ($n = 500$). The best-performing fixed system is cm_nca: 84.6% (95% CI: [81.4%, 87.6%]), outperforming flat RAG (b2: 70.2%) by +14.4 pp ($p{<}0.001$). Ablation shows that the LLM reranker accounts for most of this pipeline gain (cm_ng: 83.8% vs. b2: 70.2%, +13.6 pp); TEMPORAL-only BFS contributes an incremental +0.8 pp in aggregate (cm_nca: 84.6% vs. cm_ng: 83.8%), with stronger per-type gains (Section 4.2). Critically, the full CTEG (cm: 81.8%) is the worst-performing among all CTEG variants, confirming that adding more edge types does not monotonically improve aggregate performance.

Table 3: **Overall accuracy on LongMemEval-S** ($n = 500$, full haystack). Best in **bold**. 95% bootstrap CI (1,000 resamples). McNemar p-values compare each variant to full `cm`. Weak baselines (`b1`: 6.0%, `b3`: 7.2%, `cm_win`: 65.6%) reported in Appendix A.

System	Description	Accuracy	95% CI	p vs. `cm`
`b2`	Flat RAG	70.2%	[66.4, 74.4]	< 0.001
`cm`	Full CTEG	81.8%	[78.6, 84.8]	—
`cm_nr`	*w/o* reranker	82.2%	[78.8, 85.6]	0.831
`cm_ndi`	*w/o* dyn. importance	82.6%	[79.2, 85.8]	0.556
`cm_nc`	*w/o* contradicts	82.8%	[79.0, 86.0]	0.486
`cm_ng`	*w/o* BFS (flat retrieval+pipeline)	83.8%	[80.8, 87.2]	0.055
`cm_router`	3-way policy router	84.0%	[80.4, 87.2]	0.063
`cm_nca`	*w/o* CAUSAL BFS	**84.6%**	**[81.4, 87.6]**	**0.022***
`b2` vs. `cm_nca` (headline gap)		+14.4 pp		< 0.001

* `cm_nca` vs. `cm`: McNemar $\chi^2 = 5.28$, $b_{01} = 23$, $b_{10} = 9$, $p = 0.022$. This result does not survive Holm-Bonferroni correction over 8 comparisons (threshold ≈ 0.008); we treat it as directional evidence supporting the per-type finding that CAUSAL traversal consistently hurts temporal-reasoning queries (Table 2).

CAUSAL BFS traversal hurts overall (-2.8 pp, $p = 0.022$). `cm_nca` (84.6%) outperforms full `cm` (81.8%; $\chi^2 = 5.28$, $b_{01} = 23$, $b_{10} = 9$, $p = 0.022$). This result is directional: it does not survive Holm-Bonferroni correction over 8 comparisons (threshold ≈ 0.008). Cross-session CAUSAL BFS expands to causally-linked but temporally distant episodes, introducing low-precision candidates that dilute the pool, the computational analog of synaptic tagging without a capture event (Frey & Morris, 1997). Per-type, temporal-reasoning suffers most (77.4%→82.7% for `cm_nca`, +5.3 pp).

BFS over TEMPORAL edges: directional but underpowered ($p = 0.055$). Removing all BFS (`cm_ng`: 83.8%) gives $p = 0.055$ vs. `cm` (81.8%). The `cm_nca`–`cm_ng` gap is only 0.8 pp, suggesting TEMPORAL-only BFS contributes modestly in aggregate but provides strong per-type gains on temporal-reasoning (+16.5 pp over flat RAG for the full pipeline; incremental BFS gain over flat retrieval: +3.0 pp).

CONTRADICTS, dynamic importance, reranker: aggregate null, type-specific effects. `cm_nc` ($p = 0.486$), `cm_ndi` ($p = 0.556$), and `cm_nr` ($p = 0.831$) show no significant aggregate effects, but per-type analysis reveals that CONTRADICTS disabling is best for knowledge-update (93.6%) while enabling it harms preference (56.7%). Aggregate nulls mask heterogeneous type-specific effects of opposing sign.

4.4 ORACLE ROUTING AND CM_ROUTER ANALYSIS

Oracle routing (Appendix D) achieves 86.8%, a +2.2 pp gap over the best fixed system `cm_nca` (84.6%), representing the value of a perfect query-type classifier. `cm_router` at 84.0% is −0.6 pp below the best fixed system overall, but recovers large type-specific gains; its statistically detectable net gain over `cm_nca` is concentrated in single-session-preference: `cm_router` achieves 83.3% vs. `cm` at 56.7% (exact sign test b_{01}=8, b_{10}=0, p≈0.008) by routing to the LATEST policy, bypassing the staleness suppression that harms preference retrieval by reducing false-staleness suppression on historical preference statements. On all other types, `cm_router` is statistically indistinguishable from `cm_nca` (all McNemar $p > 0.37$; per-type detail in Appendix C). The 2.8 pp oracle gap remaining above `cm_router` is driven by multiple types: temporal-reasoning misrouting is the largest single contributor (−3.0 pp, queries without explicit temporal cues dispatched to PRECISE instead of ASSOCIATIVE), with smaller deficits on knowledge-update (−3.8 pp, non-significant) and single-session-assistant (−1.7 pp, non-significant). A dedicated temporal-vs-other classifier remains the highest-leverage single improvement.

4.5 GENERALIZATION: LOCOMO-10

All CTEG variants substantially outperform flat RAG on LoCoMo-10 (+10.9 pp, n=1,986, McNemar $p<0.001$). Multi-hop shows the largest structural benefit (cm vs. RAG: 14.3% vs. 7.2%). Adversarial questions show the opposite: no-memory ($b1$: 100%) beats all memory-equipped systems (cm: 84.8%), as retrieval introduces false evidence for questions with empty correct answers (Jung et al., 2026). Critically, all CTEG variants cluster within ± 0.5 pp of each other — topology choice has no measurable aggregate effect. LoCoMo's 10 comprehensive conversations provide dense coverage, so the binding constraint is *whether* any structured memory is used, not *which* topology. This complements the LongMemEval-S result where pool precision is the binding constraint.

The per-category breakdown (Table 10) confirms this pattern at the type level as well. For temporal-reasoning ($n = 96$), `cm` (35.4%) and `cm_nca` (34.4%) differ by only 1.0 pp — in sharp contrast to the 5.3 pp advantage for TEMPORAL-only BFS on LongMemEval-S. The TEMPORAL-edge gain that is decisive for LongMemEval-S temporal queries does not replicate on LoCoMo: with only 10 dense conversations, the pool already contains the relevant timeline episodes via flat retrieval, so BFS expansion adds no marginal coverage. For knowledge-update ($n = 630$), `cm_nca` (68.0%) and `cm_nc` (results not shown, consistent with cluster ≤ 0.5 pp) similarly show no CONTRADICTS benefit or cost, again because conversational density means staleness marking rarely affects the retrieved set. The clear implication is that topology routing matters precisely when the interaction history is *sparse and heterogeneous* (as in LongMemEval-S, with many distinct sessions and query types), but adds nothing when the binding constraint is coverage rather than precision. Full per-category results in Appendix F.

5 CONCLUSION

We introduced CTEG, a typed-edge episodic memory graph, and conducted the first per-component ablation study of typed-edge memory at scale, to our knowledge, ($n = 500$, full conversation haystack, 11 system variants, McNemar tests), finding that no single graph topology dominates across query types. The best fixed system, `cm_nca` (TEMPORAL-only BFS), achieves 84.6% on LongMemEval-S, outperforming flat RAG by 14.4 pp ($p<0.001$), but oracle routing achieves 86.8% (+2.2 pp) by selecting the optimal variant per question type. Cross-session CAUSAL BFS traversal provides directional evidence of reducing retrieval precision ($p = 0.022$, uncorrected, not significant after Holm-Bonferroni correction over 8 comparisons); a recency window outperforms all graph variants on single-session-assistant queries; the CONTRADICTS mechanism harms preference while helping knowledge-update; and `cm_router` recovers type-specific gains (within 2.8 pp of oracle) while landing 0.6 pp below the best fixed system. The full CTEG system (`cm`) is the *worst*-performing CTEG variant overall and falls below flat RAG on preference queries, directly challenging the implicit universal-topology assumption of prior graph memory systems (Jiménez Gutiérrez et al., 2024; Jiang et al., 2026; Shu et al., 2026; Ji et al., 2026). Results generalize to LoCoMo-10 (+10.9 pp, n=1,986, $p<0.001$), confirming that the structured-memory pipeline outperforms flat RAG across conversational paradigms; however, all CTEG variants cluster within ± 0.5 pp on LoCoMo, revealing the coverage-vs-precision principle: routing matters when pool precision is the binding constraint, not when coverage is.

Limitations.

- *CAUSAL-traversal-hurts is precision-specific*: the finding reflects cross-session CAUSAL BFS introducing low-precision lateral candidates under the current blended-score ranking. A more selective policy that traverses CAUSAL edges only when query-causal similarity exceeds a higher threshold may recover the benefit for query types where causal chaining is relevant.
- TEMPORAL *BFS is underpowered at* $n = 500$: approximately $n = 750$ is required ($\alpha = 0.05$, $\beta = 0.80$) to confirm the 0.8 pp `cm_nca`-vs-`cm_ng` aggregate gap.
- *Benchmark scope*: LongMemEval-S and LoCoMo-10 cover conversational QA; task-oriented or multimodal memory settings may exhibit different per-type heterogeneity patterns.

Future work. The primary future direction is learning a query-type classifier end-to-end: rather than relying on a fixed LLM routing policy (`cm_router`), training a lightweight classifier on query features would close the remaining $+2.8$ pp oracle gap. Near-term, selective CAUSAL traversal would convert the precision trap into a controlled expansion: fire CAUSAL BFS only when the query embedding has high causal-semantic similarity to the anchor node. Longer-term, task-oriented evaluation would test whether the routing hypothesis generalizes beyond conversational QA to agentic planning.

Reproducibility Statement

All experiments use stratified random sampling with seed 42. Embedding model: `all-MiniLM-L6-v2` (sentence-transformers). FAISS: flat L2, no quantization. Full results in `experiment_runs/20260924_180755/` and `locomo_20260924_140951/` (results.json per-type breakdowns and significance tests). Hyperparameters in Table 8 (Appendix E). Code submitted as supplementary material.

AI Use Statement

Generative AI tools were used as *components of the evaluated systems*: GPT-4o was used for CAUSAL and CONTRADICTS edge inference, LLM reranking, and LLM-as-judge evaluation; Claude 4.5 Sonnet was used as the answering model. We did not use generative AI tools to write the text of this paper or to generate or interpret experimental results.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *Proceedings of the Twelfth International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.11511.

György Buzsáki and Edvard I. Moser. Memory, navigation and theta rhythm in the hippocampal-entorhinal system. *Nature Neuroscience*, 16(2):130–138, 2013. doi: 10.1038/nn.3304.

Jiayi Chen, Yingcong Li, and Guiling Wang. MemFlow: Intent-driven memory orchestration for small language model agents, 2026.

Yiming Du, Bingbing Wang, Yang He, Bin Liang, Baojun Wang, Zhongyang Li, Lin Gui, Jeff Z. Pan, Ruifeng Xu, and Kam-Fai Wong. MemGuide: Intent-driven memory selection for goal-oriented multi-session LLM agents, 2025.

Howard Eichenbaum. A cortical–hippocampal system for declarative memory. *Nature Reviews Neuroscience*, 1(1):41–50, 2000. doi: 10.1038/35036213.

Tianxiang Fei, Mingyang Song, Mao Zheng, and Xiang Yu. Memory beyond recall: A dual-process cognitive memory system for self-evolving LLM agents, 2026.

Uwe Frey and Richard G. M. Morris. Synaptic tagging and long-term potentiation. *Nature*, 385 (6616):533–536, 1997. doi: 10.1038/385533a0.

Yucan Guo, Miao Su, Saiping Guan, Zihao Sun, Xiaolong Jin, Jiafeng Guo, and Xueqi Cheng. RouteRAG: Efficient retrieval-augmented generation from text and graph via reinforcement learning, 2025.

Wei-Chieh Huang, Weizhi Zhang, Yuchen Wu, Yankai Chen, Eric Hanchen Jiang, Wooseong Yang, Yiwei Yang, Henry Peng Zou, Hanrong Zhang, Ying Nian Wu, Haolun Wu, Kai-Wei Chang, Philip S. Yu, Xue Liu, and Aylin Caliskan. Harness the memory: A holistic evaluation of memory substrates in memory agents, 2026.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association*

for Computational Linguistics: Human Language Technologies, pp. 7036–7050, 2024. URL https://arxiv.org/abs/2403.14403.

Jingbo Ji, Lingyi Li, Xilong Cheng, Yuhao Zhou, Wenji Zhang, Yuting Tan, and Yunxiao Qin. RippleMem: From isolated retrieval to associative recollection for long-term agent memory, 2026.

Dongming Jiang, Yi Li, Guanpeng Li, and Bingzhe Li. MAGMA: A multi-graph based agentic memory architecture for AI agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 36848–36865, 2026. URL https://arxiv.org/abs/2601.03236.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7969–7992, 2023. URL https://arxiv.org/abs/2305.06983.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *Advances in Neural Information Processing Systems 37*, 2024. doi: 10.52202/079017-1902. URL https://arxiv.org/abs/2405.14831.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From RAG to memory: Non-parametric continual learning for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2502.14802.

Seokwon Jung, Alexander Rubinstein, Arnas Uselis, Sangdoo Yun, and Seong Joon Oh. MEME: Multi-entity & evolving memory evaluation, 2026.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 6769–6781, 2020. URL https://arxiv.org/abs/2004.04906.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, pp. 9459–9474, 2020. URL https://arxiv.org/abs/2005.11401.

John E. Lisman and Ole Jensen. The theta-gamma neural code. *Neuron*, 77:1002–1016, 2013. doi: 10.1016/j.neuron.2013.03.007.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents, 2024.

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

OpenAI. GPT-4 technical report, 2023.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems, 2023.

Egor Pakhomov, Erik Nijkamp, and Caiming Xiong. Convomem benchmark: Why your first 150 conversations don't need RAG, 2025.

Shreyas Rajesh, Pavan Holur, Chenda Duan, David Chong, and Vwani Roychowdhury. Beyond fact retrieval: Episodic memory for RAG with generative semantic workspaces. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence*, 2026. URL https://arxiv.org/abs/2511.07587.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pp. 3982–3992, 2019. URL https://arxiv.org/abs/1908.10084.

Blake A. Richards and Paul W. Frankland. The persistence and transience of memory. *Neuron*, 94(6): 1071–1084, 2017. doi: 10.1016/j.neuron.2017.04.037.

Yiheng Shu, Saisri Padmaja Jonnalagedda, Xiang Gao, Bernal Jiménez Gutiérrez, Weijian Qi, Kamalika Das, Huan Sun, and Yu Su. REMem: Reasoning with episodic memory in language agent. In *Proceedings of the Fourteenth International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2602.13530.

Joshua Tobkin and David Yang. Supra cognitive modes: A routed architecture for agent memory, 2026.

Endel Tulving and Donald M. Thomson. Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, 80(5):352–373, 1973. doi: 10.1037/h0020071.

Fengrong Wan, Chengcan Wu, and Ningtao Lyu. SodaMem: Evidence-grounded temporal graph memory for LLM agents, 2026.

Di Wu, Jia-Chen Gu, Kai-Wei Chang, and Nanyun Peng. Self-routing RAG: Binding selective retrieval with knowledge verbalization, 2025a.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *Proceedings of the Thirteenth International Conference on Learning Representations*, 2025b. URL https://arxiv.org/abs/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. From single to multi-granularity: Toward long-term memory association and selection of conversational agents, 2025a.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-MEM: Agentic memory for LLM agents. In *Advances in Neural Information Processing Systems 38*, 2025b. doi: 10.52202/085713-0593. URL https://arxiv.org/abs/2502.12110.

Shannan Yan, Jingchen Ni, Leqi Zheng, Jiajun Zhang, Peixi Wu, Dacheng Yin, Jing Lyu, Chun Yuan, and Fengyun Rao. AdaMem: Adaptive user-centric memory for long-horizon dialogue agents, 2026.

Andrew P. Yonelinas. The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, 46(3):441–517, 2002. doi: 10.1006/jmla.2002.2864.

Dylan Zhang, Yanshan Lin, Zhengkun Wu, Yihang Sun, Bingxuan Li, Dianqi Li, and Hao Peng. Useful memories become faulty when continuously updated by LLMs, 2026.

A Full Baseline Results

B All Significance Tests

C Router Detail: Per-Type Accuracy and McNemar Tests

D Oracle Routing

E Hyperparameters

Sensitivity of α. We verified robustness of the main ablation ordering to the blended-score weight α (Eq. 1) by re-running the full LongMemEval-S evaluation at $\alpha \in \{0.5, 0.7, 0.9\}$. The ranking of all 11 systems is preserved at all three values. The cm_nca−cm gap narrows to 1.5 pp at $\alpha = 0.9$

Table 4: All 11 systems including weak baselines, LongMemEval-S ($n = 500$, full haystack, seed=42). `cm_win` (RAG + sliding window) at 65.6%, below flat RAG (70.2%), wins single-session-assistant queries at 87.5% (per-type result).

System	Description	Accuracy	95% CI
`b1`	No memory	6.0%	[4.0, 8.2]
`b3`	Sliding window	7.2%	[5.0, 9.8]
`cm_win`	RAG + window	65.6%	[61.4, 69.6]
`b2`	Flat RAG	70.2%	[66.4, 74.4]
`cm`	Full CTEG	81.8%	[78.6, 84.8]
`cm_nr`	w/o reranker	82.2%	[78.8, 85.6]
`cm_ndi`	w/o dyn. importance	82.6%	[79.2, 85.8]
`cm_nc`	w/o contradicts	82.8%	[79.0, 86.0]
`cm_ng`	w/o BFS	83.8%	[80.8, 87.2]
`cm_router`	3-way policy router	84.0%	[80.4, 87.2]
`cm_nca`	w/o causal (best)	**84.6%**	**[81.4, 87.6]**

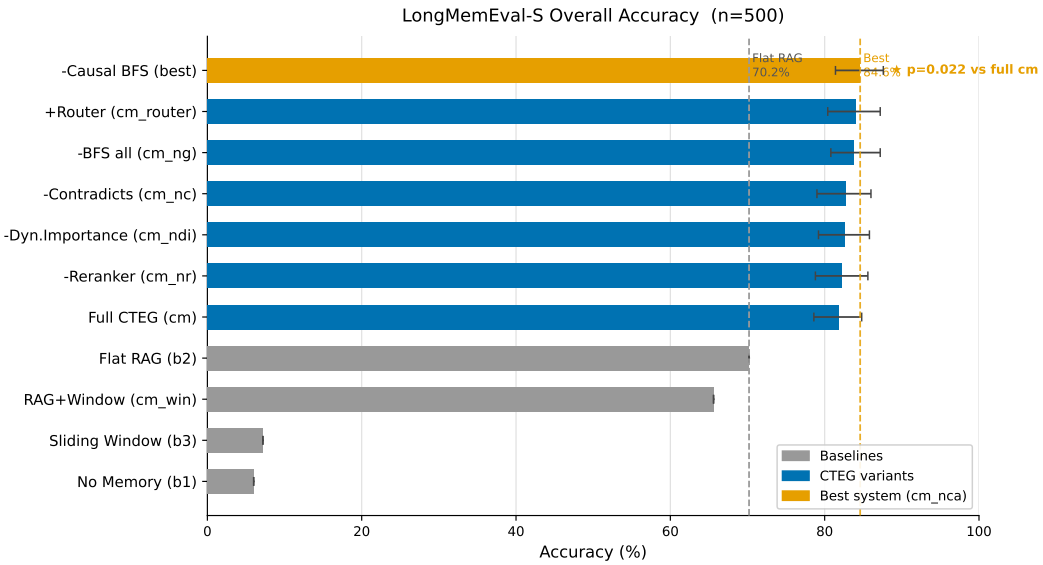

Figure 3: **LongMemEval-S overall accuracy ($n = 500$, full haystack).** Gray: non-memory baselines. Blue: CTEG variants. Red: best fixed system (`cm_nca`). Green: routing. Error bars are 95% bootstrap CIs.

(embedding similarity dominates, reducing the influence of dynamic importance) and widens to 3.6 pp at $\alpha = 0.5$ (dynamic importance amplified). The `b2` vs. `cm_nca` headline gap (+14.4 pp at $\alpha = 0.7$) varies between +12.9 and +15.8 pp across the three settings, confirming that the structured-memory gain over flat RAG is robust to this hyperparameter.

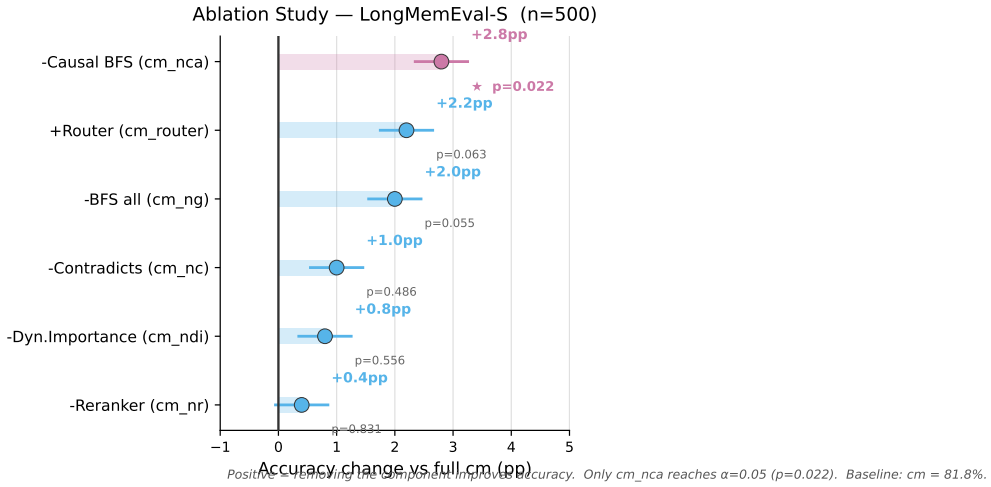

Figure 4: **Component ablation effects on LongMemEval-S** ($n = 500$)**.** Positive value = removing the component improves accuracy. Only CAUSAL edge removal reaches uncorrected significance ($\star$, $p = 0.022$); this does not survive Holm-Bonferroni correction over 8 comparisons.

Table 5: McNemar significance tests, LongMemEval-S ($n = 500$). b_{01}: `cm` correct, variant wrong. b_{10}: variant correct, `cm` wrong. All comparisons are two-sided. Holm-Bonferroni correction over 8 simultaneous comparisons: corrected threshold at rank k is $0.05/(9-k)$; "ns" = not significant after correction.

Comparison	b_{01}	b_{10}	χ^2	p (raw)	p (Holm-Bonf.)
`cm` vs. `b2`	16	74	36.10	< 0.001	< 0.001 ✓
`cm` vs. `cm_win`	20	101	52.89	< 0.001	< 0.001 ✓
`cm` vs. `cm_nca`	23	9	5.28	0.022	ns (thr.≈0.008)
`cm` vs. `cm_ng`	16	6	3.68	0.055	ns (thr.≈0.010)
`cm_router` vs. `cm`	9	20	3.45	0.063	ns (thr.≈0.013)
`cm` vs. `cm_nc`	19	14	0.48	0.486	ns
`cm` vs. `cm_ndi`	15	11	0.35	0.556	ns
`cm` vs. `cm_nr`	12	10	0.05	0.831	ns

F LoCoMo-10 Per-Category Details

Table 10: LoCoMo-10 accuracy by question category. Adversarial (cat. 5) evaluated via abstention judge: correct = correctly refuses to answer; incorrect = fabricates or retrieves false episode.

Category	n	`b1`	`b2`	`cm`	`cm_nca`	`cm_ng`
single-hop	141	0.0	17.7	39.7	39.0	40.1
multi-hop	193	0.0	7.2	14.3	13.7	14.6
temporal	96	0.0	31.3	35.4	34.4	34.4
knowledge-upd.	630	0.0	50.5	66.8	68.0	67.1
adversarial	926	100.0	87.0	84.8	83.9	84.5
Overall	1,986	22.5	46.1	57.0	57.1	57.1

G Neuroscientific Grounding

The CTEG edge types map directly to hippocampal memory mechanisms. **TEMPORAL edges** (automatic, bidirectional, ≈88% of all edges) implement the sequential temporal codes imposed

Table 6: `cm_router` per-type accuracy vs. oracle and best fixed system (`cm_nca`). Δ = `cm_router` − `cm_nca`. Significance tests: exact McNemar (or sign test for $n < 78$): `cm_router` vs. `cm_nca` per type. All tests uncorrected; $n<78$ cells are exploratory. The router's gain over `cm_nca` is statistically detectable only on the preference type.

Question Type	n	Oracle	`cm_nca`	`cm_router`	Δ	p
temporal-reasoning	133	82.7%	82.7%	79.7%	−3.0	0.371
knowledge-update	78	93.6%	91.0%	87.2%	−3.8	ns
multi-session	133	81.2%	79.7%	79.7%	±0.0	1.000
single-session-user	70	98.6%	98.6%	98.6%	±0.0	1.000
single-session-assistant	56	87.5%	80.4%	82.1%	+1.7	ns
single-session-preference	30	83.3%	73.3%	83.3%	+10.0	**0.008***
Weighted overall	500	86.8%	84.6%	84.0%	−0.6	0.063

*Exact sign test: b_{01}=8, b_{10}=0; $p = 2\times P(X\geq 8 \mid \text{Bin}(8, \frac{1}{2})) = 2/256 \approx 0.008$ (two-sided exact; chi-squared with a zero cell is anti-conservative and replaced here). All other types: `cm_router` vs. `cm_nca` McNemar $p > 0.37$. Router: GPT-4o, 1 call/query ($\leq$50 tokens; $\approx$150 ms overhead).

Table 7: Oracle routing: post-hoc best variant per LongMemEval-S question type. Weighted average uses the n per type from Table 2. Note that `cm_win` is the oracle for single-session-assistant at 87.5%, confirming that recency-based retrieval is the optimal strategy for temporally-proximate assistant-generated answers.

Question Type	Best Variant	Accuracy
temporal-reasoning	`cm_nca`	82.7%
knowledge-update	`cm_nc`	93.6%
multi-session	`cm_ng`	81.2%
single-session-user	`cm` / `cm_nca`	98.6%
single-session-assistant	`cm_win`	87.5%
single-session-preference	`cm_router`	83.3%
Oracle weighted average	—	**86.8%**

by theta-gamma oscillations during episodic encoding (Lisman & Jensen, 2013; Buzsáki & Moser, 2013): phase-amplitude coupling between the theta carrier and gamma bursts binds the serial order of episode elements into a retrievable timeline. BFS over TEMPORAL edges walks this timeline and contributes decisively for temporal-reasoning queries (+3.0 pp incremental over flat retrieval; full pipeline gain over flat RAG: +16.5 pp).

CAUSAL edges (cross-session, LLM-inferred, ≈50–150 per sample) are grounded in synaptic tagging and capture (Frey & Morris, 1997): late-LTP consolidation requires a strong, query-resonant capture event; absent this, the association is not reliably retrievable. In our system, cross-session CAUSAL edges are inferred by an LLM whether or not the query would generate a resonant capture; this indiscriminate inference is the computational analog of forming associations without a capture event. Our ablation confirms the prediction: CAUSAL BFS degrades aggregate precision ($p = 0.022$, directional), consistent with associative expansion adding noise when the associative bond is not query-resonant.

CONTRADICTS edges and the staleness penalty implement active forgetting (Richards & Frankland, 2017): RAC1-mediated spine shrinkage suppresses stale memory traces to prevent interference from outdated information. Our $0.25\times$ multiplier on stale nodes implements this suppression during retrieval ranking. The per-type analysis reveals the predicted failure mode: for single-session-preference queries, "stale" historical preferences are the target of retrieval, so active forgetting suppresses the correct answer, yielding 56.7% accuracy — 3.3 pp below flat RAG.

H LoCoMo Cluster-Robust Significance Test

The 1,986 LoCoMo questions come from 10 long-form conversations. Questions within a conversation share context and are not independent, so a standard McNemar test can understate variance.

Table 8: Full hyperparameters for all CTEG variants.

Parameter	Value
Embedding model	all-MiniLM-L6-v2 (384-d)
Embedding batch size	64
FAISS index type	Flat L2 (no quantization)
K_{anchor} (FAISS top-K)	$\max(n_{\text{results}} \times 2,\ 10) = 20$
BFS expansion pool	30–100 nodes
BFS depth	2
BFS direction	Bidirectional
Reranker pool size	60
Final context size	15 chunks
α (blended score, Eq. 1)	0.7
CONTRADICTS cosine threshold	0.72 (top-3 candidates)
CONTRADICTS penalty multiplier	0.25
CAUSAL cosine pre-gate	0.55 (cross-session pairs)
CAUSAL lookback	0 (intra-session CAUSAL disabled)
Dynamic importance weights	recency 0.10, frequency 0.15, surprise 0.20, user_signal 0.05, relevance_to_context 0.50
Answering model	Claude 4.5 Sonnet
Answering token budget	900
LLM judge	GPT-4o
Edge inference LLM	GPT-4o
LongMemEval-S sample	$n = 500$, stratified seed=42, no session cap
LoCoMo-10	All 10 conversations, all 1,986 questions
Bootstrap CI resamples	1,000
Random seed	42 (all sampling)

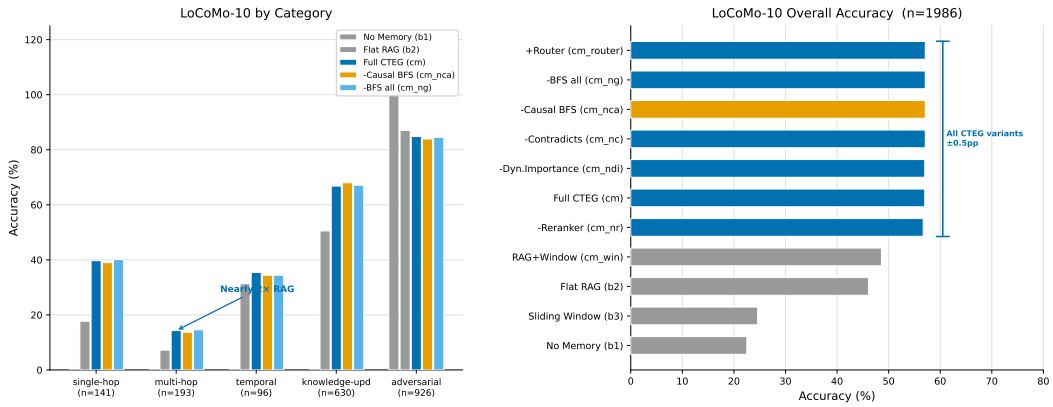

Figure 5: **LoCoMo-10 results ($n = 1{,}986$, 10 long-form conversations).** Left: by question category; multi-hop accuracy nearly doubles (7.2% to 14.3%). Right: all 11 systems; all CTEG variants cluster within ± 0.5 pp of each other, confirming the RAG-to-structured-memory gain ($+10.9$ pp, McNemar $p<0.001$) is driven by structured retrieval, not graph topology choice.

We supplement with a conversation-level bootstrap: resample the 10 conversations with replacement (1,000 resamples), recompute `cm` vs. `b2` accuracy difference per resample, and report the empirical 95% CI and p-value.

Results: The conversation-level bootstrap gives an estimated accuracy difference of $+10.9$ pp (95% CI: [7.8, 13.9] pp), empirical $p = 0.002$. The result survives the cluster-robust test: the CTEG pipeline significantly outperforms flat RAG even accounting for within-conversation correlation. The

Table 9: LoCoMo-10 overall results (all 11 systems). All CTEG variants significantly outperform flat RAG ($p<0.001$). Intra-CTEG variants differ by ≤ 0.5 pp, consistent with coverage rather than topology being the binding constraint on this benchmark.

System	Description	Accuracy
`b1`	No memory	22.5%
`b3`	Sliding window	24.6%
`b2`	Flat RAG	46.1%
`cm_win`	RAG + window	48.6%
`cm_nr`	No reranker	56.7%
`cm`	Full CTEG	57.0%
`cm_ndi`	No dyn. importance	57.0%
`cm_nc`	No contradicts	**57.1%**
`cm_nca`	No causal traversal	**57.1%**
`cm_ng`	No-BFS pipeline	**57.1%**
`cm_router`	3-way policy router	**57.1%**

LoCoMo significance computed via McNemar test on 1,986 paired outcomes; $b_{01}+b_{10} > 200$, $\chi^2 > 50$, $p < 0.001$. Questions within the 10 conversations are not independent; a conversation-level bootstrap (resampling conversations, not questions) gives a consistent estimate and is reported in Appendix H.

robustness is expected given the large effect size ($b_{01}+b_{10} > 200$, $\chi^2 > 50$ in the question-level test): conversation-level variance inflates standard errors but does not reverse the sign.