CUED RECALL: MAKING LONG-TERM MEMORY ACCESSIBLE WITH RELATIONALLY CONDITIONED CUES

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-term memory systems for large language model (LLM) agents typically match a query against the event's own content. This practice creates a representation asymmetry between query formulation and memory storage. Anticipated-question generation can narrow this gap when stored information is expressed in query-like forms, but content-only generation remains tied to the event's own wording and may miss questions whose answers depend on its relational context. We propose **Cued Recall**, a memory framework that grounds anticipated-question generation in local relational structure and aligns memory representations with the diverse ways future queries may access them. Specifically, Cued Recall first constructs a relation-labeled event graph. It links events through semantic, keyword, temporal, participant, and causal relations. Then, for each event, an LLM generates anticipated questions conditioned on its local relational neighborhood. Each question is grounded in information supported by the event and its context. At query time, the cues are combined with event content to identify seed events, and relation-aware graph activation retrieves supporting evidence around those seeds. On LoCoMo, the removal of generated cues causes a 12.20% relative drop in F1. A controlled comparison shows that cue generation grounded in relational context, rather than in the event's content alone, improves F1 by 6.89 points overall and by 10.89 points on multi-hop questions. Cued Recall further achieves an average F1 of 69.37 on HotPotQA and an average accuracy of 62.31% on PersonaMem.

1 INTRODUCTION

Large language model (LLM) agents increasingly interact with the same users across weeks or months (Wang et al., 2024; Zhang et al., 2025), and sustaining coherent, personalized behavior over that span requires retaining information from earlier interactions (Tan et al., 2025; Hu et al., 2025). Because interaction histories quickly exceed the context window, agents instead rely on external *memory systems* that transform past interactions into persistent representations and retrieve relevant evidence when needed (Lewis et al., 2020; Guo et al., 2025).

Long-term memory systems typically separate construction from retrieval. During construction, interactions are segmented and transformed into summaries, embeddings, or structured memory representations (Zhong et al., 2024; Park et al., 2023; Packer et al., 2023); during retrieval, a new query is matched against these representations through similarity-based search (Karpukhin et al., 2020; Gao et al., 2023) or relational search over memory graphs (Gutiérrez et al., 2024; Edge et al., 2024). In either case, retrieval must bridge the gap between how a future query is formulated and how the relevant event was represented during construction. Yet conventional memory construction does not explicitly account for how an event may later be queried, leaving this alignment largely to the retrieval stage.

This asymmetry between how information is later asked about and how it was originally stored parallels a classical distinction in human memory between *availability* and *accessibility*: information can be retained yet remain difficult to retrieve when an appropriate cue is missing (Tulving & Pearlstone, 1966; Raaijmakers & Shiffrin, 1981). Agent memory faces a similar gap. A memory graph may preserve an event's causal, temporal, or participant relations, so the information is available in the stored representation, yet a future query may refer to that event through one of these relations

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.

rather than through its original wording. The question is therefore not only whether an event is stored, but whether its representation provides routes through which future queries can reach it.

We refer to this property as *retrieval accessibility*: the extent to which stored events can be found by plausible future queries. In conversational memory, an event's accessibility often hinges on its relational role, for instance which participant, temporal relation, or connected event offers an alternative route to the same information. This motivates *Cued Recall*: constructing memory together with cues that anticipate how stored information may later be queried (Tulving & Pearlstone, 1966; Tulving, 1983).

We instantiate this idea as **Cued Recall**, a memory framework built around **Generative Cued Retrieval**. We first construct a relation-labeled event graph from the interaction history. For each event, an LLM then considers the event together with its local graph neighborhood and generates future-oriented questions that the event can answer, retaining only questions supported by the event and its relational context as retrieval cues. At query time, explicit person entities can first project retrieval onto a relevant candidate subgraph; cue and content matching within that candidate space identifies seed evidence, and relation-aware graph activation then gathers supporting evidence around those seeds. The graph is thus used twice: to condition cue generation at construction time, and to compose evidence at retrieval time. Across LoCoMo (Maharana et al., 2024), HotPotQA (Yang et al., 2018), and PersonaMem (Jiang et al., 2025), a controlled comparison confirms that this gain comes specifically from grounding cue generation in relational context rather than from generating cues at all, and the resulting system outperforms state-of-the-art memory systems across all three benchmarks. We summarize our contributions as follows:

- We formulate long-term memory retrieval through the lens of *availability* and *accessibility*, highlighting construction-time retrieval cues as a mechanism for improving the latter.
- We propose **Generative Cued Retrieval**, which constructs future-oriented retrieval cues from event content conditioned on local relational context.
- Through controlled cue-generation comparisons, ablations, and mechanistic analysis, we isolate the contribution of relational conditioning and show that graph expansion complements cue- and content-based seeding, particularly on multi-hop questions; on LoCoMo, this conditioning improves F1 by 6.89 points over text-only cue generation and by 10.89 points on multi-hop questions.

2 Related Work

Memory Systems and Retrieval for LLM Agents. Long-term memory systems extend LLM agents beyond stateless interactions by transforming past interactions into persistent representations and retrieving relevant evidence at query time. Research in this area has advanced along two complementary fronts. On the construction side, systems store past interactions as persistent profiles, summaries, or hierarchical short- and long-term memory (Park et al., 2023; Zhong et al., 2024; Packer et al., 2023; Wang et al., 2024; Xu et al., 2026; Chhikara et al., 2025; Tan et al., 2025; Kang et al., 2025; Hu et al., 2025; Fang et al., 2026; Liu et al., 2026), with recent work further incorporating temporal, episodic, or graph-based organization for evolving agent memory (Rasmussen et al., 2025; Shu et al., 2026; Zhang et al., 2025). On the retrieval side, RAG-style query-document matching (Lewis et al., 2020; Karpukhin et al., 2020) has been extended through query reformulation, document expansion, and hierarchical search (Nogueira et al., 2019; Wang et al., 2023; Gao et al., 2023; Jiang et al., 2023; Jang et al., 2024; Sarthi et al., 2024; Wu et al., 2024), and through graph-based methods that exploit relational structure for multi-hop and associative retrieval (Edge et al., 2024; Guo et al., 2025; Li et al., 2024; Gutiérrez et al., 2024; 2025; Anokhin et al., 2024; Zhang et al., 2026). Across both fronts, accessibility remains an outcome of query-time search rather than a property shaped during construction itself. Cued Recall instead treats accessibility as an explicit construction objective, using an event's relational neighborhood to condition cue generation before any query arrives, and reusing the same relational structure to compose evidence once a query does.

Generated Retrieval Representations and Cognitive Memory. Generated retrieval representations have been explored through document-to-query expansion, query-to-document generation, and retrieval-guided reformulation (Nogueira et al., 2019; Wang et al., 2023; Gao et al., 2023; Jiang et al., 2023; Jang et al., 2024), typically generating from an event's own content or from the in-

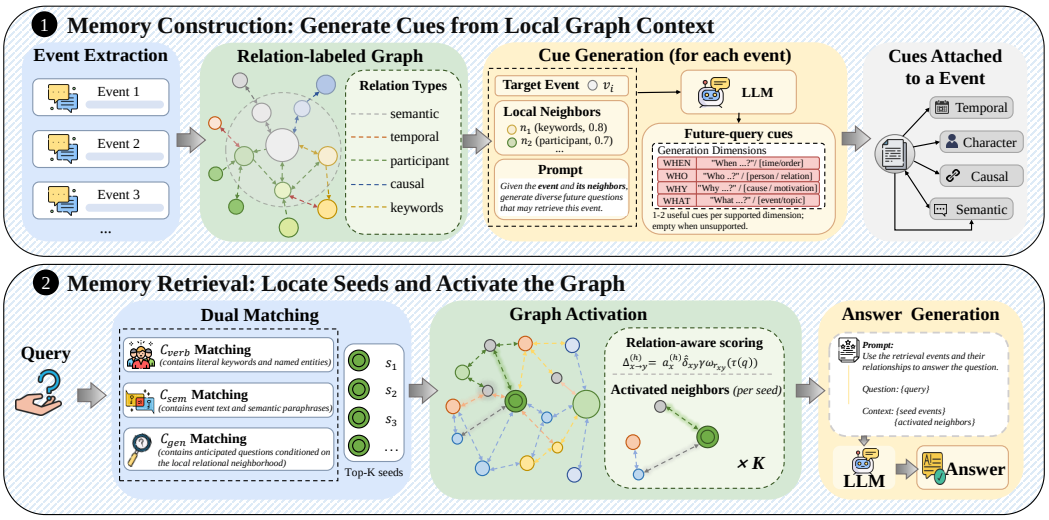

Figure 1: **Overview of Cued Recall.** During **construction**, events are organized into a relation-labeled graph, and local relational context is used to generate future-query cues for each event. During **retrieval**, the query is matched against both event content and these cues to identify candidate seeds, followed by relation-aware graph activation to retrieve supporting evidence.

coming query. Classical memory research offers a complementary account of why this may be insufficient, distinguishing information availability from accessibility and emphasizing retrieval-relevant context through encoding specificity (Tulving & Pearlstone, 1966; Tulving & Thomson, 1973), while associative-memory models characterize retrieval as search over interconnected representations (Raaijmakers & Shiffrin, 1981; Anderson, 1983). Cued Recall connects these two lines: it generates future-oriented questions in the spirit of prior generated-representation methods, but conditions them on an event's local relational neighborhood rather than on content alone, grounding generated retrieval cues in structured long-term memory.

3 METHOD

Figure 1 illustrates this framework.

3.1 DESIGN PRINCIPLE

Our framework follows a simple principle: **accessibility should be constructed alongside availability**. In human memory, availability and accessibility are distinct: information may be retained while remaining difficult to retrieve without an effective cue (Tulving & Pearlstone, 1966). In a memory system, storing an event provides availability, whereas accessibility additionally depends on whether a future query can reach that event through its retrieval representation. Thus, increasing the amount of stored information or relational structure does not necessarily provide a corresponding improvement in query accessibility.

This distinction also motivates the use of relational context during encoding. The encoding specificity principle (Tulving & Thomson, 1973) suggests that retrieval benefits when cues reflect information associated with the encoded memory. For conversational events, such information is not limited to the event's own wording: participants, temporal relations, shared topics, and causal dependencies may determine what future questions the event can answer. We therefore use the event's local graph neighborhood as context for generating future-oriented retrieval cues. The generation effect (Slamecka & Graf, 1978) further motivates constructing these retrieval-oriented representations through generation, rather than relying only on fixed textual forms. We directly test the resulting hypothesis, that relational conditioning improves accessibility beyond cues generated from event text alone, in Section 5.4.

The resulting construction order is important. We first build the relation-labeled graph so that relational context is available, and then generate cues conditioned on that context. At retrieval time, these cues determine the initial access to memory, while the graph is subsequently used to compose additional evidence around the retrieved seeds.

3.2 Memory Graph Construction

We represent long-term memory as a directed, relation-labeled graph with retrieval cues, $\mathcal{M} = (V, E, C)$. Each node $v \in V$ corresponds to an event represented as

$$v = \langle \text{text}(v), \tau(v), \text{part}(v), \mathbf{e}(v) \rangle, \qquad \mathbf{e}(v) = \phi(\text{text}(v)) \in \mathbb{R}^d,$$

where $\text{text}(v)$ is the event content, $\tau(v)$ is its timestamp, $\text{part}(v)$ is its participant set, and ϕ is a shared embedding encoder.

The graph uses five relation types

$$\mathcal{R} = \{\text{sem}, \text{key}, \text{temp}, \text{part}, \text{causal}\}.$$

These relations capture complementary forms of conversational association: semantic similarity, shared keywords, temporal proximity, shared participants, and directed causal dependencies. For each ordered event pair (u, v), relation-specific mechanisms first produce candidate edges and raw scores. **(i) Semantic** relations use the cosine similarity of event embeddings, $\delta_{uv}^{(\text{sem})} = \cos(\mathbf{e}(u), \mathbf{e}(v))$. **(ii) Keyword** relations use a score that increases with the overlap between the two events' extracted keyword sets, $\mathcal{K}(u) \cap \mathcal{K}(v)$, saturating once the overlap is large enough, with at least one shared keyword required for the pair to be a candidate at all. **(iii) Temporal** relations assign a higher score to events occurring on the same day than to events on adjacent days, and do not connect events further apart than this. **(iv) Participant** relations use a score that increases with the size of the shared participant set, $\text{part}(u) \cap \text{part}(v)$, saturating in the same manner as the keyword score, with at least one shared participant required for the pair to be a candidate. **(v) Causal** relations are directed edges identified and scored directly by a construction-time LLM. Different relation types thus use the signal most appropriate to their semantics, rather than forcing all connections into a single similarity space.

The resulting raw scores are not directly comparable across relation types. We therefore calibrate each relation using its empirical cumulative distribution function,

$$\hat{\delta}_{uv}^{(r)} = F_r\left(\delta_{uv}^{(r)}\right),$$

and retain the strongest calibrated relation for each ordered pair,

$$r^*(u, v) = \arg\max_{r \in \mathcal{R}} \hat{\delta}_{uv}^{(r)}.$$

This produces a single relation label and calibrated edge weight for each retained pair, while avoiding direct comparisons between heterogeneous raw score scales. Each event further keeps a bounded set of outgoing neighbors, stored in a bounded adjacency list, `event_neighbors`, whose entries record a neighbor, its calibrated edge strength, and its relation label.

Candidate edges are generated through corresponding retrieval mechanisms: semantic candidates come from strong embedding neighbors above a similarity threshold; keyword candidates are obtained from an inverted keyword index while excluding overly frequent keywords that would create spurious high-degree connections; temporal candidates are formed from same- and adjacent-day events; participant candidates are obtained from an inverted participant index; and causal candidates come from cached LLM judgments computed once during construction. If multiple mechanisms generate the same ordered pair, the calibration-and-maximum step above determines the retained relation. This bounded neighborhood serves the two roles used throughout the paper: it supplies relational context for cue construction (Section 3.3) and defines the adjacency traversed during graph activation (Section 3.4). For a fixed memory snapshot, the graph, causal judgments, and cues are cached and reused across queries; when new interactions are incorporated, construction is rerun for the updated snapshot.

Each event stores three complementary cue channels,

$$C(v) = C_{\text{verb}}(v) \cup C_{\text{sem}}(v) \cup C_{\text{gen}}(v).$$

The first two are derived from event content, whereas $C_{\text{gen}}(v)$ is explicitly conditioned on relational context, letting the same graph serve both roles: conditioning cue generation at construction time, and composing evidence at retrieval time.

3.3 CUE CONSTRUCTION

The purpose of cue construction is to expose an event through the different ways in which future queries may refer to it. For each event v, we generate future-oriented questions from the event together with a bounded sample of its local relational neighborhood, rather than from the event text alone.

Local relational context. Let $N(v) \subseteq \texttt{event_neighbors}(v)$ denote the neighborhood provided to the generator. To avoid overrepresenting a single relation type, when the budget b is large enough to cover every relation type, we first allocate one slot to the strongest neighbor of each relation and use the remaining slots for the strongest edges overall,

$$N(v) = \bigcup_{r \in \mathcal{R}} \{\text{Top1}_r(v)\} \cup \text{TopK}_{\text{rest}}(v; b - |\mathcal{R}|)\,.$$

Otherwise, we simply retain the top-b incident neighbors across all relation types, without the per-relation guarantee. This bounded selection exposes diverse relational signals while keeping the generation context fixed and computationally manageable.

Future-query cue generation. Conditioned on the focal event and its selected neighborhood, the cue generator π_θ produces questions organized by relation dimension,

$$C_{\text{gen}}(v) = \{c \sim \pi_\theta\,(\cdot \mid \text{text}(v), N(v))\}\,,$$

with the number of generated cues bounded both per relation dimension and in total (exact bounds are given in Section 4). The generator is instructed to retain only questions supported by the event and its relational context. The resulting cue is not intended to predict the exact wording of a future query; instead, it provides a retrieval-compatible representation of a plausible question that the event can answer. Relational conditioning is therefore used to change the space of accessible questions, rather than merely to paraphrase the event.

We maintain three retrieval channels with distinct roles. C_{verb} contains literal keywords and named entities and supports direct lexical access. C_{sem} contains event text and semantic paraphrases and supports content-level matching. C_{gen} contains future-oriented questions conditioned on the local relational neighborhood and provides an additional access route when a future query depends on information distributed across related events.

3.4 RETRIEVAL AND ANSWERING

Given a query q, retrieval proceeds through four stages: *Project*, *Seed*, *Graph Activation*, and *Answer*. The first two stages identify where the query can initially access the memory, while graph activation subsequently composes supporting evidence.

Project. A lightweight LLM extracts explicit person entities, surface keywords, temporal expressions, and a query type $\tau(q)$. The query type determines relation-specific weights $w_r(\tau(q))$ used during graph activation. When explicit persons are identified, retrieval is projected onto the union of their associated event subgraphs; if no explicit person is available, or the projected candidate space is insufficient, retrieval falls back to the full graph. Projection only narrows the candidate space; it does not alter the stored graph or its edges.

Seed: resolving accessibility. Within the projected candidate space, the query embedding $\phi(q)$ is matched independently against the three cue channels, each producing a ranked list of candidate events, which are fused using reciprocal rank fusion,

$$s_0(v) = \sum_{m \in \{\text{verb,sem,gen}\}} \frac{1}{k + \text{rank}_m(v)}.$$

The top-K events form the seed set $S_0(q)$. Because the three channels encode different access routes, their fusion allows direct lexical, content-semantic, and relationally conditioned matches to jointly determine initial event accessibility.

Graph Activation: composing availability. Seed retrieval identifies the events that are initially accessible to the query, but an answer may depend on additional events connected to those seeds. We therefore perform bounded, relation-aware activation over the pre-built graph, following the spreading-activation paradigm (Anderson, 1983). Each seed is initialized with activation $a_v^{(0)} = 1$, and propagation continues for at most H hops. At each hop, a frontier node expands to at most A neighbors. For a selected edge $x \to y$ with calibrated strength $\hat{\delta}_{xy}$ and relation r_{xy}, the propagated contribution is

$$\Delta_{x\to y}^{(h)} = a_x^{(h)} \hat{\delta}_{xy}\, \gamma\, w_{r_{xy}}(\tau(q)), \qquad a_y^{(h+1)} = \sum_{x \in \mathcal{F}_h(y)} \Delta_{x\to y}^{(h)},$$

where $\gamma \in (0, 1)$ is the per-hop attenuation factor and $\mathcal{F}_h(y)$ contains the frontier nodes reaching y at hop h. After propagation, seed nodes are excluded, and the top-B activated nodes are retained as additional evidence. The parameters A, H, and B control the expansion fan-out, propagation depth, and evidence budget, respectively (values given in Section 4). Importantly, graph activation does not replace cue-based seed retrieval: it operates only after seeds have been identified, and therefore composes relational evidence rather than determining initial accessibility.

Answer. The seed events and activated evidence are jointly passed to the answer LLM, which generates the final response to q. This final stage allows the system to combine directly accessible events with the relational evidence retrieved around them.

4 Experimental Setup

Datasets and baselines. We evaluate on three complementary benchmarks. **LoCoMo** (Maharana et al., 2024) is a long-range conversational memory benchmark. We use its 1,540 non-adversarial questions across four categories: single-hop, multi-hop, temporal, and open-domain. **HotPotQA** (Yang et al., 2018) is evaluated in a streaming retrieval setting with document pools of 400, 800, and 1,600 documents; results for Long-Context, A-Mem, Mem0, and MEMORYOS follow the protocol of Wang et al. (2026). **PersonaMem** (Jiang et al., 2025) evaluates personalization over 32K-token dialogue histories with six multiple-choice tasks. We compare with Long-Context, RAG (Lewis et al., 2020), HippoRAG (Gutiérrez et al., 2024), A-Mem (Xu et al., 2026), Mem0 (Chhikara et al., 2025), MEMORYOS (Kang et al., 2025), LightMem (Fang et al., 2026), E-mem (Wang et al., 2026) (HotPotQA only), and SimpleMem (Liu et al., 2026).

Backbones and metrics. We report LoCoMo results with both GPT-4o-mini and Qwen3-8B. GPT-4o-mini serves as the primary backbone for ablations and HotPotQA/PersonaMem experiments, while Qwen3-8B verifies the findings on an open-weight model. All construction-stage LLM calls use the same backbone as the answering model. We report F1 and BLEU-1 (Papineni et al., 2002) on LoCoMo, F1 on HotPotQA, and multiple-choice accuracy on PersonaMem.

Implementation details. Key retrieval hyperparameters are a seed budget $K = 40$, an activation fan-out $A = 10$, an activated-evidence budget $B = 20$, a maximum propagation depth $H = 2$, a per-hop decay factor $\gamma = 0.5$, and a local-neighbor budget $b = 4$. During cue construction, the LLM generates at most 3 cues per relation dimension and 12 cues per event. Baselines use their reported or official hyperparameters; backbone settings follow the corresponding evaluation protocol.

5 Experiments

5.1 Main Results

Cued Recall achieves the highest overall F1 on both backbones in Table 1, reaching 57.93 with GPT-4o-mini and 46.30 with Qwen3-8B, and the improvement holds consistently across single-hop,

Table 1: LoCoMo main results. F1 and BLEU-1 in paired columns. Bold: best per column within backbone; underline: second-best.

Backbone	Method	Overall		Single-hop		Multi-hop		Temporal		Open-domain	
		F1	B1	F1	B1	F1	B1	F1	B1	F1	B1
GPT-4o-mini	Long-Context	36.87	29.07	45.92	37.13	30.18	23.65	24.72	17.58	17.94	12.80
	RAG	44.95	39.59	52.74	48.01	25.37	19.80	47.83	41.29	24.66	18.27
	HippoRAG	47.02	40.20	53.69	47.72	33.45	25.08	47.24	38.80	**27.66**	**23.41**
	A-Mem	40.53	33.74	46.23	39.43	27.90	21.37	45.17	36.21	12.14	12.00
	Mem0	44.60	36.17	47.90	39.85	38.61	27.05	47.85	40.73	22.35	15.53
	MEMORYOS	44.14	36.02	49.66	43.02	37.83	27.44	42.08	30.98	21.27	16.77
	LightMem	41.25	33.88	44.18	36.47	35.03	25.19	44.25	40.07	23.75	16.01
	SimpleMem	48.13	41.27	53.77	47.85	36.60	27.11	51.29	44.04	22.03	15.97
	Cued Recall	**57.93**	**51.55**	**63.96**	**58.81**	**45.67**	**37.83**	**62.24**	**56.24**	26.67	12.62
Qwen3-8B	Long-Context	33.06	21.87	43.05	33.87	32.18	24.90	12.55	10.00	16.65	12.22
	RAG	36.95	28.03	45.33	35.45	35.26	22.93	22.45	17.73	17.06	12.47
	HippoRAG	40.91	34.65	46.10	40.47	32.15	24.89	41.25	33.60	**20.01**	**15.83**
	A-Mem	36.96	27.23	41.99	33.14	33.71	20.80	32.68	22.45	16.71	10.39
	Mem0	39.61	29.78	44.21	34.72	36.83	23.45	37.26	28.30	15.37	10.00
	MEMORYOS	34.14	29.13	41.27	36.03	21.66	17.15	31.74	25.92	16.38	14.67
	LightMem	34.06	26.55	40.75	31.82	27.35	20.18	28.11	22.90	15.00	11.26
	SimpleMem	40.97	35.37	47.35	41.28	29.12	24.06	42.34	36.23	15.32	14.01
	Cued Recall	**46.30**	**40.44**	**52.11**	**46.81**	**38.59**	**29.90**	**46.25**	**40.87**	18.26	14.18

Table 2: HotPotQA and PersonaMem results on GPT-4o-mini. HotPotQA reports F1; PersonaMem reports multiple-choice accuracy (RF = Recall Facts, SI = Suggest Ideas, TE = Track Evolution, RR = Revisit Reasons, AR = Aligned Recommendation, GS = General Scenario). Bold: best; underline: second-best. HotPotQA results for Long-Context, A-Mem, Mem0, and MEMORYOS are taken from Wang et al. (2026) under an identical streaming protocol; "–" marks a benchmark E-mem was not evaluated on.

	HotPotQA (F1)				PersonaMem (Acc.)						
Method	400	800	1600	Avg.	RF	SI	TE	RR	AR	GS	Avg.
Long-Context	56.56	49.71	53.92	53.40	28.77	11.83	45.32	60.61	40.00	28.07	44.82
A-Mem	33.90	30.22	31.37	31.83	46.58	30.11	43.88	73.74	52.73	26.32	46.52
Mem0	32.85	31.74	27.41	30.67	**65.75**	**33.33**	46.76	83.84	61.82	61.40	58.40
MEMORYOS	26.47	23.10	24.16	24.58	64.38	31.18	48.92	84.85	56.36	40.35	55.86
E-mem	61.46	55.46	55.76	57.56	–	–	–	–	–	–	–
Cued Recall	**70.14**	**69.04**	**68.94**	**69.37**	63.70	32.26	**56.83**	**86.87**	**72.73**	**68.42**	**62.31**

multi-hop, and temporal questions. Open-domain performance is comparatively lower for every method, Cued Recall included, an effect we return to in Section 5.5 and Appendix C.4.

5.2 HotPotQA and PersonaMem

On HotPotQA, Cued Recall remains stable as the document pool grows from 400 to 1,600 documents, achieving F1 scores of 70.14, 69.04, and 68.94 with an average of 69.37. Under the same protocol, E-mem and Long-Context obtain average F1 scores of 57.56 and 53.40, respectively. On PersonaMem, Cued Recall achieves an average accuracy of 62.31 and obtains the highest score on four of the six evaluated tasks. Mem0 achieves higher accuracy on Recall Facts, with scores of 65.75 compared with 63.70 for Cued Recall, and on Suggest Ideas, with scores of 33.33 compared with 32.26. Cued Recall achieves higher scores on the remaining four tasks, including Track Evolution, Revisit Reasons, Aligned Recommendation, and General Scenario (Table 2).

Table 3: Ablation summary.

Config.	Perf. (drop ↓)	
	LoCoMo	PersonaMem
Cued Recall (full)	**57.93**	**62.31**
w/o Gen. cues	50.86 (↓ 12.20%)	58.74 (↓ 5.73%)
w/o Activation	55.24 (↓ 4.64%)	59.76 (↓ 4.09%)
Uniform expand	54.33 (↓ 6.22%)	59.42 (↓ 4.64%)

Table 4: Construction cost.

Method	Tokens (M) ↓			Calls ↓
	In	Out	Sum	
A-Mem	9.126	2.368	11.494	11,754
Mem0	10.958	1.239	12.196	9,181
MEMORYOS	**1.889**	**0.939**	**2.868**	5,534
Cued Recall	2.544	1.202	3.746	**1,829**

5.3 Ablation

Table 3 summarizes component ablations. Removing generative cues causes the largest performance drop on both benchmarks, with relative decreases of 12.20% on LoCoMo and 5.73% on PersonaMem, highlighting cue generation as a major contributor to the framework's gains. Removing graph activation further reduces F1 by 4.64% and 4.09%, while replacing relation-aware activation with uniform expansion leads to larger drops of 6.22% and 4.64%. Uniform expansion also performs worse than removing activation entirely, achieving F1 scores of 54.33 and 55.24 on LoCoMo, and 59.42 and 59.76 on PersonaMem, respectively. These results suggest that relation-specific weighting helps prioritize useful graph evidence, whereas relation-agnostic expansion may introduce less relevant connections.

The category breakdown in Table 7 sharpens this picture. Uniform expansion has its largest effect on multi-hop questions, where F1 falls from 45.7 to 40.1, the setting in which knowing which relation to follow matters most. Removing generative cues produces the largest drop in every LoCoMo category, including 9.0 points on single-hop and 5.2 points on multi-hop questions, and on PersonaMem the effect is most pronounced on Aligned Recommendation ; Revisit Reasons is the one task left unchanged, since its answer is typically stated directly in the dialogue rather than requiring a routing cue.

5.4 Isolating the Source of the Gain: Graph-Conditioned vs. Text-Only Cues

The ablation above shows that generative cues matter, but not whether relational conditioning specifically drives the gain, as opposed to cue generation in general. To isolate this effect, we compare three cue settings on a fixed LoCoMo subset with the same downstream retrieval pipeline: **No Cue**, **Text-only Cue**, which conditions the same generator only on the focal event, and **Graph-conditioned Cue (ours)**, which additionally provides the local relational neighborhood.

Text-only Cue does not improve over No Cue, achieving an F1 of 41.32 compared with 41.95 for No Cue, showing that generating additional questions from event text alone is not sufficient to explain the gain. Graph-conditioned Cue improves over Text-only Cue by 6.89 F1 points and 6.03 BLEU-1 points, and over No Cue by 6.27 F1 points, isolating relational conditioning as the key source of improvement. This gain is largest on multi-hop questions, where F1 improves by 10.89 points over Text-only Cue (Appendix B).

Table 5: Novelty isolation on LoCoMo subset (Qwen3-8B). Bold: best.

Condition	F1	BLEU-1
No Cue	41.95	35.93
Text-only	41.32	35.18
Graph-cond. (ours)	**48.22**	**41.21**

5.5 Mechanistic Analysis

To understand where the gain originates, we evaluate retrieval directly against annotated gold evidence, before answer generation enters the picture. Cue-based retrieval alone reaches 69.79 R@10 and 47.46 MRR, well above the 52.54 and 32.16 obtained from content-only retrieval, and combining the two channels further raises R@10 to 71.29 and MRR to 48.64 (Table 6).

Gold-evidence coverage (distinct from the query-level R@10 above) measures the fraction of annotated gold evidence contained in the retrieved seed events. By this measure, content and cue seeding individually cover 50.89% and 25.28% of gold evidence, while their union reaches 67.50%, confirming complementary rather than overlapping coverage. Graph activation's contribution is con-

Table 6: Cue–question alignment and seed-evidence coverage on LoCoMo (Qwen3-8B). (a) Overall retriever-level metrics. (b) By LoCoMo category (R@10); ΔR@10 denotes the gain from adding the cue channel, i.e., Union R@10 − Content R@10. (c) Seed-evidence coverage by LoCoMo category. All values are percentages. "New Gold" is the share of questions where activation supplies a gold node absent from the seed.

(a) Overall retriever-level metrics

Retriever	R@1	R@5	R@10	MRR
Content only	21.94	42.97	52.54	32.16
Cue only	35.81	61.52	69.79	47.46
Content+Cue	**36.65**	**62.83**	**71.29**	**48.64**

(b) By LoCoMo category (R@10)

Category	Content	Cue	Union	ΔR@10
Single-hop	47.16	73.05	**74.47**	↑27.31
Temporal	52.96	76.64	**77.26**	↑24.30
Open-domain	30.43	47.83	**50.00**	↑19.57
Multi-hop	56.60	68.49	**70.27**	↑13.67

(c) Seed-evidence coverage by LoCoMo category

Metric	Overall	S-Hop	Temp.	Open.	M-Hop
Content Seed	50.89	44.53	70.43	35.62	52.97
Cue Seed	25.28	21.61	13.62	28.77	49.32
Union Seed	67.50	64.09	78.99	52.05	69.86
New Gold	↑1.54	↑0.44	↑1.56	↑2.74	↑4.57

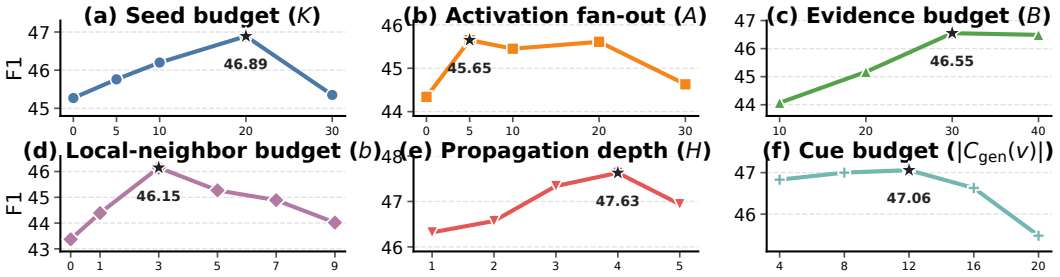

Figure 2: Hyperparameter sensitivity on LoCoMo with Qwen3-8B. Each panel is an independent one-factor sweep, varying one hyperparameter while the others remain at their defaults. Stars mark the best-performing value within each individual sweep.

centrated on multi-hop questions, where it supplies new gold evidence, absent from the seed set, for 4.57% of questions. Appendix C provides representative retrieval traces.

5.6 Efficiency and Sensitivity

Construction cost. Table 4 reports one-time construction costs. Cued Recall requires 3.746M LLM tokens across 1,829 calls, substantially reducing construction overhead compared with other memory systems. The graph, edges, and cues are cached and reused across subsequent queries.

Hyperparameter sensitivity. Figure 2 shows one-factor sweeps with other parameters fixed. Performance remains stable across moderate ranges: larger seed and activation budgets provide limited gains, while excessive propagation depth hurts performance. The local-neighbor budget b benefits from a small relational context, and increasing $|C_{\text{gen}}(v)|$ beyond 12 cues yields diminishing returns.

6 Conclusion and Future Work

Cued Recall treats retrieval accessibility as a construction-time property of long-term memory by generating future-oriented cues from events and their relational context to guide graph-based retrieval. Experiments on LoCoMo, HotPotQA, and PersonaMem show consistent improvements, with analyses highlighting the role of relationally conditioned cues and complementary retrieval channels. Limitations include single-backbone trace analysis, fixed graph-activation weights, and upfront LLM construction costs. Future work will explore learned relation weighting and more efficient incremental construction. Overall, our results suggest that accessibility can be explicitly constructed alongside availability rather than left solely to query-time search.

AI USE STATEMENT

In this work, we used generative AI tools for manuscript editing, including improving readability, clarity, and language expression. We have not used generative AI tools for developing the research methodology, designing experiments, implementing the proposed methods, generating synthetic datasets, performing data analysis, or drawing conclusions from experimental results. The code used in our experiments was implemented, verified, and tested by the authors. We have reviewed all AI-assisted content and take responsibility for the final manuscript, including all text, claims, interpretations, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies memory systems for LLM agents that store and retrieve information from user interactions. Such systems introduce privacy and reliability considerations, as conversational histories may contain sensitive personal information, and retrieved memories may retain or amplify biases, errors, or misinformation from the original interactions. We use only publicly available research datasets (LoCoMo, HotPotQA, and PersonaMem) and collect no new human-subject data. Any deployment of memory-augmented agents should incorporate appropriate safeguards, including user consent, secure data management, and mechanisms for correcting or mitigating erroneous or biased memories.

REFERENCES

John R Anderson. A spreading activation theory of memory. *Journal of verbal learning and verbal behavior*, 22(3):261–295, 1983.

Petr Anokhin, Nikita Semenov, Artyom Sorokin, Dmitry Evseev, Andrey Kravchenko, Mikhail Burtsev, and Evgeny Burnaev. Arigraph: Learning knowledge graph world models with episodic memory for llm agents. *arXiv preprint arXiv:2407.04363*, 2024.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, et al. Lightmem: Lightweight and efficient memory-augmented generation. In *International Conference on Learning Representations*, volume 2026, pp. 98706–98729, 2026.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1762–1777, 2023.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tian Ao, and Chao Huang. Lightrag: Simple and fast retrieval-augmented generation. In *EMNLP (Findings)*, pp. 10746–10761, 2025.

Bernal J Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. Hipporag: Neurobiologically inspired long-term memory for large language models. *Advances in neural information processing systems*, 37:59532–59569, 2024.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From rag to memory: Non-parametric continual learning for large language models. *arXiv preprint arXiv:2502.14802*, 2025.

Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, Fangyi Zhu, Jiahang Lin, Honglin Guo, Shihan Dou, Zhiheng Xi, et al. Memory in the age of ai agents. *arXiv preprint arXiv:2512.13564*, 2025.

Yunah Jang, Kang-il Lee, Hyunkyung Bae, Hwanhee Lee, and Kyomin Jung. Itercqr: Iterative conversational query reformulation with retrieval guidance. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8121–8138, 2024.

Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J Taylor, and Dan Roth. Know me, respond to me: Benchmarking llms for dynamic user profiling and personalized responses at scale. *arXiv preprint arXiv:2504.14225*, 2025.

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.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. Memory os of ai agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25972–25981, 2025.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 6769–6781, 2020.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Shilong Li, Yancheng He, Hangyu Guo, Xingyuan Bu, Ge Bai, Jie Liu, Jiaheng Liu, Xingwei Qu, Yangguang Li, Wanli Ouyang, et al. Graphreader: Building graph-based agent to enhance long-context abilities of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 12758–12786, 2024.

Jiaqi Liu, Yaofeng Su, Peng Xia, Siwei Han, Zeyu Zheng, Cihang Xie, Mingyu Ding, and Huaxiu Yao. Simplemem: Efficient lifelong memory for llm agents. *arXiv preprint arXiv:2601.02553*, 2026.

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

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375*, 2019.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G Patil, Ion Stoica, and Joseph E Gonzalez. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Jeroen G Raaijmakers and Richard M Shiffrin. Search of associative memory. *Psychological review*, 88(2):93, 1981.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: a temporal knowledge graph architecture for agent memory. *arXiv preprint arXiv:2501.13956*, 2025.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher Manning. Raptor: Recursive abstractive processing for tree-organized retrieval. In *International Conference on Learning Representations*, volume 2024, pp. 32628–32649, 2024.

Yiheng Shu, Padmaja Jonnalagedda, Xiang Gao, Bernal Jimenez Gutierrez, Weijian Qi, Kamalika Das, Huan Sun, and Yu Su. Remem: Reasoning with episodic memory in language agent. In *International Conference on Learning Representations*, volume 2026, pp. 55958–55987, 2026.

Norman J Slamecka and Peter Graf. The generation effect: Delineation of a phenomenon. *Journal of experimental Psychology: Human learning and Memory*, 4(6):592, 1978.

Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, et al. 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)*, pp. 8416–8439, 2025.

Endel Tulving. Elements of episodic memory. 1983.

Endel Tulving and Zena Pearlstone. Availability versus accessibility of information in memory for words. *Journal of verbal learning and verbal behavior*, 5(4):381–391, 1966.

Endel Tulving and Donald M Thomson. Encoding specificity and retrieval processes in episodic memory. *Psychological review*, 80(5):352, 1973.

Kaixiang Wang, Yidan Lin, Jiong Lou, Zhaojiacheng Zhou, Bunyod Suvonov, and Jie Li. E-mem: Multi-agent based episodic context reconstruction for llm agent memory. *arXiv preprint arXiv:2601.21714*, 2026.

Liang Wang, Nan Yang, and Furu Wei. Query2doc: Query expansion with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9414–9423, 2023.

Yu Wang, Chi Han, Tongtong Wu, Xiaoxin He, Wangchunshu Zhou, Nafis Sadeq, Xiusi Chen, Zexue He, Wei Wang, Gholamreza Haffari, et al. Towards lifespan cognitive systems. *arXiv preprint arXiv:2409.13265*, 2024.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. *arXiv preprint arXiv:2410.10813*, 2024.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. *Advances in Neural Information Processing Systems*, 38:17577–17604, 2026.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2369–2380, 2018.

Kai Zhang, Xinyuan Zhang, Ejaz Ahmed, Hongda Jiang, Caleb Kumar, Kai Sun, Zhaojiang Lin, Sanat Sharma, Shereen Oraby, Aaron Colak, et al. Assomem: Scalable memory qa with multi-signal associative retrieval. In *International Conference on Learning Representations*, volume 2026, pp. 103385–103405, 2026.

Zeyu Zhang, Quanyu Dai, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model-based agents. *ACM Transactions on Information Systems*, 43(6):1–47, 2025.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 19724–19731, 2024.

APPENDIX

A. Construction and Retrieval Algorithm
Operational details of memory construction, cue generation, and retrieval.

B. Supplementary Mechanism Analysis
Fine-grained ablations and analyses of graph-conditioned cues and activation.

B.1 Fine-Grained Ablation
B.2 Novelty Isolation by Category

C. Case Studies
Retrieval traces illustrating accessibility, relational conditioning, and evidence composition.

C.1 Accessibility: Full vs. No Cue
C.2 Relational Conditioning: Full vs. Text-only Cue
C.3 Evidence Composition: Full vs. w/o Activation
C.4 Open-Domain Boundary Cases

Algorithm 1 Cued Recall: Memory Construction and Retrieval

Input: Conversation history $\mathcal{D}$ and query q.

Memory construction:

1. Extract event nodes V from $\mathcal{D}$. Each event is represented as
$$v = \langle \text{text}(v), \tau(v), \text{part}(v), \mathbf{e}(v) \rangle,$$
where $\mathbf{e}(v)$ is produced by the shared embedding encoder.
2. Construct the relation-labeled memory graph $\mathcal{M} = (V, E, C)$ using semantic, keyword, temporal, participant, and causal relations.
3. For each relation type, compute relation-specific edge scores, calibrate them into comparable strengths, and retain the strongest relation for each ordered event pair. Store the bounded adjacency list `event_neighbors`.
4. For each event v, retrieve a bounded neighborhood $N(v) \subseteq \texttt{event_neighbors}(v)$ and generate future-oriented cues:
$$C_{\text{gen}}(v) = \text{LLM}(\text{text}(v), N(v)).$$
5. Store the complete cue representation
$$C(v) = C_{\text{verb}}(v) \cup C_{\text{sem}}(v) \cup C_{\text{gen}}(v),$$
and cache the resulting memory snapshot.

Query retrieval:

6. Extract query entities, keywords, temporal expressions, and query type $\tau(q)$. When explicit person entities appear, first filter events associated with these entities and restrict retrieval to the resulting candidate subgraph.
7. Match the query against event content and all cue channels: C_{verb}, C_{sem}, and C_{gen}. Fuse the retrieval rankings and select the top-K events as seed nodes:
$$S_0(q).$$
8. Initialize activation scores:
$$a_v^{(0)} = 1, \qquad v \in S_0(q).$$
For each hop, propagate activation through the cached graph. The contribution from node x to neighbor y is computed as:
$$\Delta x \to y^{(h)} = a_x^{(h)} \hat{\delta}_{xy} \gamma w_{r_{xy}}(\tau(q)),$$
where $\hat{\delta}_{xy}$ denotes calibrated edge strength, γ is the hop decay factor, and $w_{r_{xy}}(\tau(q))$ represents query-dependent relation importance.
Only the highest-scoring activated neighbors are expanded.
9. Combine the seed events and activated nodes:
$$S(q) = S_0(q) \cup S_{\text{act}}(q),$$
and provide the resulting evidence set to the answer LLM to generate the final response y.

Output: Answer y.

A Cued Recall Construction and Retrieval Algorithm

B Supplementary Mechanism Analysis

This appendix provides additional analyses supporting the mechanism claims in Section 5. The analyses focus on three aspects: the contribution of generative cues, the effect of relational conditioning during cue construction, and the role of graph activation in evidence composition.

B.1 Fine-Grained Ablation

Table 7 provides category-level results behind the aggregate ablation study in Table 3. Removing generative cues decreases performance across most categories, with larger drops on questions requiring indirect access to relevant events. Removing activation mainly affects categories requiring evidence composition, while uniform expansion performs worse than relation-aware activation, especially on multi-hop questions.

On PersonaMem, the effect varies across tasks. Tasks requiring preference or behavior evolution show larger degradation after removing generative cues, whereas tasks whose evidence is explicitly stated in the dialogue are less affected.

Table 7: Fine-grained ablation on LoCoMo categories (F1) and PersonaMem tasks (accuracy), using GPT-4o-mini. Bold denotes the full model.

	LoCoMo (F1)				PersonaMem (Acc.)					
Config.	S-Hop	M-Hop	Temp.	Open.	RF	SI	TE	RR	AR	GS
Full (Ours)	**64.0**	**45.7**	**62.2**	**26.7**	**63.70**	**32.26**	**56.83**	**86.87**	**72.73**	**68.42**
w/o Gen. cues	55.0	40.5	61.1	22.3	61.64	26.88	54.68	86.87	60.00	61.40
w/o Activation	59.0	42.7	60.2	20.6	58.22	30.11	56.12	84.85	70.91	66.67
Uniform expand	61.0	40.1	62.0	21.3	60.27	29.03	56.83	85.86	62.27	61.40

B.2 Novelty Isolation by Category

Table 8 decomposes the comparison in Section 5.4 by LoCoMo category. Graph-conditioned cues obtain larger improvements on categories requiring relational information, particularly multi-hop and single-hop questions. The gains are smaller on temporal and open-domain questions, where the required information can be more directly expressed or less dependent on specific relations.

Table 8: Novelty isolation on a 384-question LoCoMo subset using Qwen3-8B. Bold denotes the best result within each category and metric.

	Multi-hop		Temporal		Open-domain		Single-hop	
Method	F1	BLEU-1	F1	BLEU-1	F1	BLEU-1	F1	BLEU-1
No Cue	35.78	28.63	53.13	45.75	18.37	8.95	41.67	37.04
Text-only Cue	30.16	23.10	51.58	44.59	26.99	**17.06**	42.35	37.33
Graph-conditioned Cue	**41.05**	**31.91**	**55.12**	**47.15**	**27.45**	16.42	**49.95**	**44.60**

C CASE STUDIES

We trace representative LoCoMo questions under four configurations: **Full** uses graph-conditioned cues and graph activation; **Text-only Cue** generates cues from event content without relational context; **No Cue** removes the cue channels; and **w/o Activation** retains cue-based seeding but removes graph activation. All predictions and per-question F1 scores are taken from the project evaluator on questions for which the compared conditions produced a valid prediction. Green marks indicate correct or partially correct answers, while red marks indicate incorrect or missing answers. Gold answers and evidence are taken from the dataset annotations.

The case studies are organized around the distinction between *accessibility* and *composition*: the first two sections examine how cues route a query to the appropriate event, the third examines how graph activation composes additional evidence after a seed has been found, and the fourth examines cases at the boundary of the mechanism, where the difficulty lies beyond relational retrieval altogether.

C.1 ACCESSIBILITY: FULL VS. NO CUE

These questions illustrate cases in which the relevant event is present in the memory graph but is not sufficiently accessible through content-based matching alone. Full uses graph-conditioned cues to provide an additional retrieval route. The comparison with w/o Activation isolates this routing effect from subsequent evidence composition.

Table 9: Accessibility cases comparing Full with No Cue. Full recovers the gold answer in all five cases, while No Cue fails to reach the relevant evidence. The matching results of Full and w/o Activation indicate that these examples primarily reflect cue-based routing rather than graph-based composition.

Full	**Text-only**	**No Cue**	**w/o Activation**
Q1. Which city have both Jean and John visited? *(Multi-hop; Gold: Rome; D2:5, D15:1)*			
Rome (1.00)	Paris (0.00)	cannot be determined (0.00)	Rome (1.00)
Q2. How long has Caroline had her current group of friends for? *(Temporal; Gold: 4 years; D3:13)*			
4 years (1.00)	22 October 2023 (0.00)	2023 (0.00)	4 years (1.00)
Q3. What country is Caroline's grandma from? *(Single-hop; Gold: Sweden; D4:3)*			
Sweden (1.00)	No information provided (0.00)	Caroline's grandma is from Poland (0.00)	Sweden (1.00)
Q4. What was the focus of John's recent research and writing on his blog? *(Open-domain; Gold: education reform and infrastructure development; D12:5)*			
education reform and infrastructure development (1.00)	politics and the government (0.00)	politics and the government (0.00)	education reform and infrastructure development (1.00)
Q5. What was discussed in the LGBTQ+ counseling workshop? *(Open-domain; Gold: therapeutic methods and how to best work with trans people; D4:13)*			
different therapeutic methods and how to best work with trans people (0.95)	LGBTQ support/self-acceptance/mental health/community (0.00)	transgender journey and LGBTQ community (0.00)	different therapeutic methods and how to best work with trans people (0.95)

Analysis. The failures of No Cue are not near matches: "Paris," "2023," "Poland," "politics and the government," and "transgender journey" refer to different events from those supporting the gold answers. In all five cases, the relevant evidence is present in the memory graph but is not reached by content-based retrieval alone. Since w/o Activation matches Full in every row, these examples isolate the accessibility benefit of cue-based seeding.

C.2 Relational Conditioning: Full vs. Text-only Cue

These cases isolate the contribution of relational conditioning during cue generation. The generator, prompt, and cue budget are held fixed between Full and Text-only Cue; only the conditioning context differs. The comparison therefore distinguishes graph-conditioned cue generation from generative cue expansion based on event text alone.

Table 10: Relational-conditioning cases comparing Full with Text-only Cue. Text-only Cue performs no better than No Cue in these examples, whereas Full is correct or substantially closer to correct in all four cases.

Full	**Text-only**	**No Cue**	**w/o Activation**
Q1. What are Melanie's pets' names? *(Multi-hop; Gold: Oliver, Luna, Bailey; D13:4, D7:18)*			
Luna and Oliver (0.80)	No pets mentioned (0.00)	No pets mentioned (0.00)	Luna and Oliver (0.80)
Q2. When did Maria start volunteering at the homeless shelter? *(Temporal; Gold: around August 2022; D27:4)*			
23 August 2022 (0.67)	3 August 2023 (0.33)	3 August 2023 (0.33)	23 August 2022 (0.67)
Q3. Who is Melanie a fan of in terms of modern music? *(Single-hop; Gold: Ed Sheeran; D15:28)*			
Ed Sheeran (1.00)	Caroline (0.00)	Caroline (0.00)	Ed Sheeran (1.00)
Q4. Which activity has John done apart from yoga at the studio? *(Single-hop; Gold: weight training; D25:17)*			
weight training (1.00)	kickboxing (0.00)	kickboxing (0.00)	weight training (1.00)

Analysis. Q1 requires combining pet names distributed across multiple events. Q2 requires distinguishing the volunteering date in August 2022 from a nearby event in August 2023. Q3 and Q4 depend on relation-specific attributes—a person's modern-music preference and an activity performed in addition to yoga—that are not directly expressed in the focal event's wording. These examples show that the benefit of generative cues depends on the relational context used during their construction.

C.3 Evidence Composition: Full vs. w/o Activation

These cases illustrate situations in which cue-based seeding reaches a relevant event, but the seed alone does not contain all information needed for the answer. Removing graph activation preserves the initial seed while removing the additional relational evidence used for completion, disambiguation, or attribute binding.

Analysis. Three composition patterns emerge. First, *entity completion* is illustrated by Q1 and Q2: the seed reaches one relevant event, whereas activation retrieves additional events needed to complete the set of entities or causes. Second, *event disambiguation* appears in Q3 and Q4, where nearby events can lead to an incorrect date without relational context. Third, *attribute binding* appears in Q5–Q7, where the retrieved event must be connected to a specific observation, emotion, or location. These failures occur after a relevant seed has been found and therefore concern evidence composition rather than initial accessibility.

Table 11: Evidence-composition cases involving entity completion and event disambiguation. Graph activation supplies additional events needed to complete the answer or distinguish among temporally related events.

Full	Text-only	No Cue	w/o Activation
Q1. What are Maria's dogs' names? *(Multi-hop; Gold: Coco, Shadow; D30:1, D31:4)*			
Shadow and Coco (1.00)	Coco (0.67)	Shadow and Coco (1.00)	Coco (0.67)
Q2. What causes has John done events for? *(Multi-hop; Gold: toy drive, community food drive, veterans, domestic violence)*			
Homeless shelter, toy drive for kids in need, community event for a good cause, veterans' rights, and domestic abuse victims (0.38)	Community Support, Volunteering, Other Causes (0.35)	Community Support, Volunteering, Other Causes (0.35)	Homeless shelter, toy drive for kids in need, community event for a good cause, marching event for veterans' rights, and events for veterans' rights (0.21)
Q3. When did Maria's grandmother pass away? *(Temporal; Gold: the week before 6 March 2023; D8:1)*			
March 2023 (0.57)	the week of 23 August 2023 (0.40)	27 February 2023 (0.25)	10 April 2023 (0.25)
Q4. When did Melanie's family go on a roadtrip? *(Temporal; Gold: the weekend before 20 October 2023)*			
22 October 2023 (0.50)	the week of 23 October 2023 (0.40)	19 October 2023 (0.50)	the week of 23 August 2023 (0.20)

Table 12: Evidence-composition cases involving attribute binding. The seed locates the relevant event, while activation supplies the associated observation, emotion, or location needed for the final answer.

Full	Text-only	No Cue	w/o Activation
Q5. What did Melanie and her family see during their camping trip last year? *(Open-domain; Gold: Perseid meteor shower; D10:14)*			
the Perseid meteor shower (1.00)	explored nature, roasted marshmallows around the campfire, went on a hike, and saw the view from the top (0.00)	explored nature, roasted marshmallows around the campfire, went on a hike, and saw the view from the top (0.00)	roasted marshmallows around the campfire, explored nature, went on a hike, and shared stories (0.00)
Q6. How did Melanie feel while watching the meteor shower? *(Open-domain; Gold: in awe of the universe; D10:18)*			
awe-inspiring and in awe of the universe (0.89)	amazing, at one with the universe, awe-inspiring (0.20)	awe-inspiring (0.00)	awe-inspiring (0.00)
Q7. Where did Oliver hide his bone once? *(Single-hop; Gold: in Melanie's slipper)*			
in my slipper (0.67)	No evidence (0.00)	under the bed (0.00)	your slipper (0.40)

C.4 Open-Domain Boundary Cases

The mechanism is not expected to resolve every open-domain question. These cases characterize situations in which the answer requires broad abstraction, attribute aggregation, or answer calibration rather than retrieval of a specific relational dependency.

Table 13: Open-domain boundary cases. Q1 remains imperfect under all configurations, while Q2 is correctly answered by all configurations that retrieve the relevant event.

Full	**Text-only**	**No Cue**	**w/o Activation**
Q1. What attributes describe John? *(Gold: selfless, family-oriented, passionate, rational; D26:6, D2:14, D3:5, D4:6)*			
John is resilient, resourceful, and passionate about community and family. He also experiences personal struggles and seeks purpose. (0.11)	John is passionate, supportive, family-oriented, and community-focused. He finds purpose in helping others and is motivated by his family and community work. (0.17)	John's attributes include his background, motivations, family life, personal struggles, hobbies, work experience, and military experience. (0.00)	John is resilient, resourceful, and passionate about community and family. He shows resilience by staying calm during unexpected incidents and seeking assistance. (0.03)
Q2. Around which US holiday did Maria get into a car accident? *(Gold: Independence Day; D21:3)*			
July 4th Independence Day (0.67)	July 4th Independence Day (0.67)	July 4th (0.00)	July 4th Independence Day (0.67)

Analysis. Q1 requires aggregating and abstracting personality attributes across multiple events; all configurations produce incomplete or differently formulated descriptions, and Text-only Cue obtains the highest score in this case, though the difference does not establish a general advantage since all answers remain imperfect. Q2 shows that some apparent retrieval failures are lexical rather than relational: No Cue produces "July 4th," whereas the other configurations recover the equivalent holiday name. These cases define a boundary of the mechanism rather than contradicting its intended role in improving access to and composition of relational evidence.