SAGE: A SELF-EVOLVING AGENTIC GRAPH-MEMORY ENGINE FOR STRUCTURE-AWARE ASSOCIATIVE MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-term memory is becoming a central bottleneck for language agents. Existing RAG and GraphRAG systems largely treat memory graphs as static retrieval middleware, which limits their ability to recover complete evidence chains from partial cues, exploit reusable graph-structural roles, and improve the memory itself through downstream feedback. We introduce SAGE, a **S**elf-evolving **A**gentic **G**raph-memory **E**ngine that models graph memory as a dynamic long-term memory substrate. SAGE couples two roles: a memory writer that incrementally constructs structured graph memory from interaction histories, and a Graph Foundation Model-based memory reader to perform retrieval and provide feedback to the memory writer. We provide rigorous theoretical analyses supporting the effectiveness of carefully designed architectural components and the framework. Across multi-hop QA, open-domain retrieval, domain-specific review QA, and long-term agent-memory benchmarks, SAGE improves evidence recovery, answer grounding, and retrieval efficiency: after two self-evolution rounds, it achieves the best average rank on multi-hop QA; in zero-shot open-domain transfer, it reaches EM/F1 of 46.20/54.21 and 53.00/66.06 on NQ and PopQA. Most importantly, when used as a replacement memory-access backend for existing memory systems, SAGE improves LongMemEval overall accuracy by 7.4 to 9.6 absolute points over the native MemOS, Memobase, and Zep backends after one evolution round, while paired HaluMem experiments improve most operation-level memory metrics and reduce QA hallucination. These results suggest that self-evolving, structure-aware graph memory is a promising way for robust long-horizon language agents. Our code is available here.

1 INTRODUCTION

As large language models evolve from single-turn question-answering systems into general-purpose agents for multi-turn dialogue, personalized assistance, multi-agent collaboration, and open-environment exploration, the system bottleneck is shifting from whether a model can answer within the current context to whether it can accumulate, organize, invoke, and update memory over longer time scales. Memory is a core system capability that determines whether Agents can achieve long-term consistency, personalized adaptation, cross-turn reasoning, and self-improvement. **Memory is to Agents what parameters are to foundation models** (Park et al., 2023; Zhong et al., 2024; Packer et al., 2023; Wu et al., 2025b; Yang et al., 2026). Recent memory benchmarks have made this bottleneck explicit, evaluating agents on ultra-long conversational consistency, multi-session reasoning, temporal reasoning, knowledge updating, selective forgetting, abstention, and hallucination control (Maharana et al., 2024; Wu et al., 2025a; Hu et al., 2025; Chen et al., 2025a; Li et al., 2026b).

In engineering practice, RAG has become the dominant non-parametric interface for extending language models with external memory, alleviating the static nature of parametric knowledge and the limited size of context windows (Lewis et al., 2020). Yet standard RAG usually retrieves independent text chunks, whereas long-term agent memory often requires recovering evidence distributed across entities, events, aliases, temporal constraints, and multi-hop dependencies. GraphRAG takes an important step by organizing documents, entities, relations, and summaries as graphs, making cross-document dependencies and reasoning paths more explicit (Edge et al., 2024; Gutiérrez et al., 2024).

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

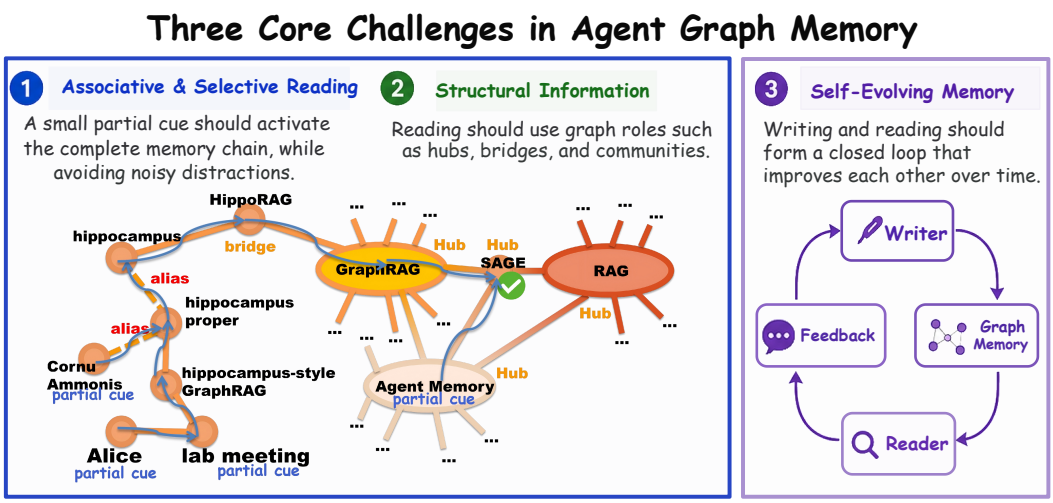


Figure 1: Overview of the three core challenges in agent graph memory, illustrated with a concrete example. Given the query, ***"Alice mentioned a work in last week's lab meeting that seemed to be inspired by the Cornu Ammonis. Among works in the same field as that work, are there any that can also help with agent memory? Give one example,"*** the memory reader must address associative and selective reading by expanding sparse partial cues into the correct evidence chain while avoiding noisy distractors. It must then exploit structural information in the memory graph, such as aliases, bridges, and hubs, to traverse from Cornu Ammonis to SAGE. Self-evolving memory highlights the closed loop between writing and reading.

However, for long-horizon agents, graph structure should not merely serve as an external retrieval index. In this work, we study *agent graph memory* as a coupled write–read–update problem. Given interaction histories or external documents, a memory writer should construct an **evolving graph** whose nodes and edges encode entities, episodes, documents, aliases, temporal constraints, and cross-fragment relations. Given a query, a memory reader should not simply expand from a few matched entities; it should return a **compact, verifiable evidence chain**. The retrieval outcome should further **provide feedback about what the graph lacks**. In other words, the graph is not only built before retrieval and searched afterward; it is the **working substrate through which memory is written, read, corrected, and self-improved**. Around this goal, we identify the following three core challenges.

Challenge I: Agent memory requires global associative reading from fragmented cues. The first challenge is not merely to retrieve text that is semantically similar to the query, but to reconstruct a complete reasoning chain from sparse, fragmented, and sometimes indirect cues. In long-term agent memory, a query may mention only an episodic clue, an alias, or a distant conceptual hint, while the answer depends on intermediate entities that are not explicitly named. Standard vector retrieval tends to return locally similar snippets, and many graph-based retrieval methods start propagation from a small set of query-matched anchor entities. However, if these anchors only cover a local subgraph, the necessary bridge nodes may lie outside the activated region, leaving the evidence chain disconnected even after graph propagation. Thus, agent memory reading should not commit too early to a small set of partial cues (Trivedi et al., 2023; Gutiérrez et al., 2024).

Challenge II: Agent memory requires learned structural use rather than fixed structural expansion. The second challenge is that graph structure should not be used only as a fixed index after graph construction. Many GraphRAG-style systems exploit structure through pre-built communities, paths, graph indexes, or heuristic expansion rules, but once the graph is constructed, the role of structure is largely fixed: a hub remains broadly expanded, a bridge may be missed if it is not reached by the initial anchors, and noisy shortcuts may be treated similarly to useful evidence edges. This is insufficient for agent memory, where the graph itself is continuously updated by new interactions and where the same topological pattern may have different meanings across domains. A structure-aware reader should therefore learn how structural roles affect retrieval (Edge et al., 2024; Gutiérrez et al., 2024; Liu et al., 2025; Luo et al., 2025).

The example in Figure 1 illustrates both challenges. Given the query, the explicit cues are only *Alice*, *lab meeting*, *Cornu Ammonis*, and *agent memory*. A retrieval system that only anchors on the most query-matched nodes may retrieve the meeting note or the biological cue, but still fail to connect them to *hippocampus*, *HippoRAG*, *GraphRAG*, and finally SAGE. This is the associative-reading challenge: the system must piece together a long chain from scattered cues. At the same time, the correct path depends on structural roles: *HippoRAG* is a bridge, *GraphRAG* and *RAG* are hubs that must be controlled rather than blindly expanded, and the edge from *GraphRAG* to SAGE is critical for reaching the final answer. This is the structural-information challenge: the reader must use graph topology in a learned and selective way, not simply propagate uniformly over a fixed graph.

Challenge III: Existing methods mostly optimize retrieval trajectories, but rarely optimize the self-evolution of the memory system itself. Existing RAG and GraphRAG systems often assume that the external memory graph or knowledge base is already available, so the main problem becomes how to retrieve from it. For long-term agents, however, writing is itself part of the memory problem. Conversely, retrieval failures provide useful signals about what the memory graph lacks. For example, if the reader repeatedly needs to traverse from *Cornu Ammonis* to *hippocampus-style GraphRAG* and then to the GraphRAG literature, the memory system should gradually add or strengthen useful structural links, such as a more direct edge from *hippocampus-style GraphRAG* to *GraphRAG*. Thus, a true agent memory system should not only optimize retrieval trajectories; it should optimize the memory graph itself through a closed loop in which better reading exposes writing deficiencies, and better writing makes future reading more accurate, selective, and efficient (Chen et al., 2025a).

To address these challenges, we propose SAGE, a ***S***elf-evolving ***A***gentic ***G***raph-memory ***E***ngine. Unlike GraphRAG systems that mainly use graphs as retrieval middleware, SAGE treats the graph as a dynamic long-term memory object. It couples two mutually reinforcing components: a memory writer that incrementally constructs structured graph memory, and a Graph Foundation Model-based memory reader that retrieves evidence and provides feedback to the writer. SAGE directly targets the three challenges above: it recovers long reasoning chains from fragmented cues, learns how to use structural roles rather than propagate uniformly, and continuously improves the graph memory itself for future queries.

2 Related Work

Retrieval-Augmented Generation and GraphRAG. Retrieval-Augmented Generation (RAG) provides a non-parametric interface for language models by retrieving external evidence before generation (Lewis et al., 2020). Many variants further improve retrieval timing, reasoning interaction, adaptive policies, and hierarchical organization (Jiang et al., 2023; Trivedi et al., 2023; Asai et al., 2024; Jeong et al., 2024; Sarthi et al., 2024). GraphRAG enables structured retrieval over cross-document dependencies and multi-hop evidence paths (Edge et al., 2024; He et al., 2024; Guo et al., 2024; Li et al., 2024; Wang & Han, 2025; Xu et al., 2025a; Zhao et al., 2025; Zhang et al., 2025c; Gutiérrez et al., 2024; 2025; Luo et al., 2025). Another line of work improves retrieval by optimizing retrieval trajectories, including interleaved retrieval and reasoning, self-reflective retrieval, adaptive retrieval, multi-agent RAG, and reinforcement-learning-based query rewriting (Trivedi et al., 2023; Asai et al., 2024; Jeong et al., 2024; Chen et al., 2025b; Cha et al., 2025; Tsang et al., 2025).

Agent Memory. Agent memory studies how LLM-based agents store, update, retrieve, and use past experiences (Park et al., 2023; Zhong et al., 2024; Packer et al., 2023; Chhikara et al., 2025; Xu et al., 2025b; Rasmussen et al., 2025; Kang et al., 2025; Zhang et al., 2025a; Wu et al., 2025c; Zhang et al., 2025b; Huang et al., 2025; Yue et al., 2026). Recent surveys also highlight the importance of human-inspired and graph-based memory mechanisms for LLM agents (Wu et al., 2025b; Yang et al., 2026). Meanwhile, memory benchmarks evaluate long-term consistency, event reasoning, multi-session reasoning, temporal reasoning, knowledge updating, selective forgetting, abstention, and hallucination control (Maharana et al., 2024; Wu et al., 2025a; Hu et al., 2025; Chen et al., 2025a; Li et al., 2026b).

Graph Foundation Models. Graph Foundation Models (GFMs) aim to learn transferable graph representations through large-scale pretraining, allowing models to reuse structural priors and semantic patterns across graphs, tasks, and domains (Liu et al., 2025). Representative early works include GCC, GPT-GNN, and GraphCL, which learn transferable graph representations through cross-network

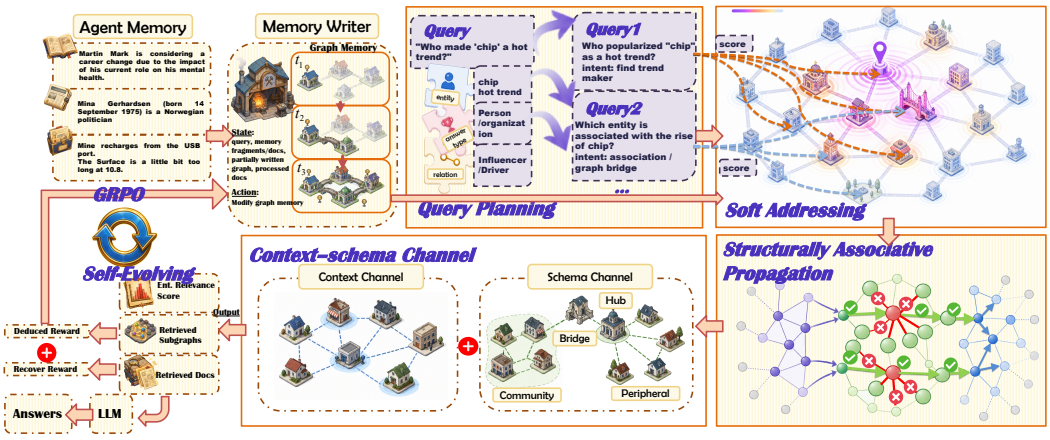


Figure 2: Overall pipeline of the proposed **SAGE**. The memory writer incrementally constructs and updates graph memory from observations through state-conditioned writing actions, and receives rewards from downstream memory use. The resulting retrieval feedback closes the loop between writing and reading, enabling graph memory to improve over time.

contrast, generative graph pretraining, and graph augmentation based contrastive learning (Qiu et al., 2020; Hu et al., 2020; You et al., 2020; Yu et al., 2025).

3 Preliminary

We consider a knowledge-intensive memory sample $x = (q, \mathcal{D}, \mathcal{D}^+, y)$, where q denotes the query, $\mathcal{D} = \{d_i\}_{i=1}^N$ is the set of candidate historical memory fragments, $\mathcal{D}^+ \subseteq \mathcal{D}$ is the gold evidence set supporting the answer, and y is the ground-truth answer. Depending on the benchmark, a memory fragment may correspond to a document, review snippet, or interaction/session unit. The writer is modeled as a structured policy: at step h, given state s_h, it samples a writing action $a_h \sim \pi_\theta(\cdot \mid s_h)$ and updates the partial graph as $\mathcal{G}_{h+1} = \mathcal{G}_h \oplus a_h$. The memory reader $\mathcal{R}_\phi$ performs query-conditioned propagation over the graph, obtains entity relevance scores $\mathbf{s}_E = f_\phi(q, \mathcal{G}) \in \mathbb{R}^{|\mathcal{V}_E|}$, and then projects them into memory-fragment scores. The reader finally outputs $(\widehat{\mathcal{D}}_k, \widehat{\mathcal{G}}_q, \Pi_q) = \mathcal{R}_\phi(q, \mathcal{G}, \mathbf{M})$, where $\widehat{\mathcal{D}}_k = \mathrm{TopK}_{d \in \mathcal{D}}(\mathbf{s}_D(d))$, $\widehat{\mathcal{G}}_q$ is the query-activated subgraph, and Π_q denotes optional relational paths. The generation model then produces the answer $\widehat{y} = \mathrm{LLM}(q, \widehat{\mathcal{D}}_k, \Pi_q)$.

4 Method

At a high level, our method builds a self-evolving graph memory pipeline (Figure 2). The memory writer $\mathcal{W}_\theta$ first transforms the query and candidate historical memory fragments into a heterogeneous graph memory $\mathcal{G}$. The memory reader $\mathcal{R}_\phi$ then performs query-conditioned activation over $\mathcal{G}$: it softly locates query-relevant entities, propagates evidence signals through relational structures, and projects the activated entity-level information back to memory fragments.

4.1 Memory Writer: Graph Memory Writing via Reading Feedback

Policy-based writing. The writer is modeled as a sequential decision-making policy. At step t, the state is defined as $s_t = \left(q, \mathcal{D}, \mathcal{G}_{t-1}, \mathcal{D}_{t-1}^{\mathrm{proc}}\right)$, where $\mathcal{G}_{t-1}$ is the partially written graph, and $\mathcal{D}_{t-1}^{\mathrm{proc}}$ denotes the set of processed documents. The action a_t contains entity-relation triples (u, r, v) together with their source anchors $(u, \texttt{source}, d)$. Detailed implementation information is provided in Appendix O.

Reader-aware Writing Reward. The writer's reward stems from the task utility of its written graph after being accessed by the memory reader. Given the current graph $\mathcal{G}$, the frozen reader returns the evidence $P_k(q, \mathcal{G})$. Inspired by (Tsang et al., 2025), we employ two complementary

types of rewards. The first category measures whether the graph is sufficient as a knowledge carrier to support the derivation of the answer: $r_{\text{ded}}(q, y, \mathcal{G}) = \mathbb{I}\left[\text{Judge}\left(q, y \mid P_k(q, \mathcal{G})\right) = \texttt{Yes}\right]$. The second category measures whether the graph can serve as a knowledge index to recover the supporting text: $r_{\text{rec}}(q, \mathcal{D}^+, \mathcal{G}) = \frac{|P_k(q,\mathcal{G}) \cap \mathcal{D}^+|}{|\mathcal{D}^+|}, r_{\text{pre}}(q, \mathcal{D}^+, \mathcal{G}) = \frac{|P_k(q,\mathcal{G}) \cap \mathcal{D}^+|}{|P_k(q,\mathcal{G})|}$. Here, r_{rec} encourages the coverage of necessary evidence, while r_{pre} penalizes the expansion of irrelevant evidence. To align with end-to-end question answering, we also use an answer-level auxiliary reward $r_{\text{ans}}(q, y, \mathcal{G}) = \max_{y' \in \mathcal{Y}(y)} \text{F1}\big(\hat{y}, y'\big), \hat{y} = \text{LLM}\left(q, P_k(q, \mathcal{G})\right)$, where $\mathcal{Y}(y)$ is the set of answer aliases. In practice, we adopt a hybrid task reward $r_{\text{task}} = \frac{\alpha r_{\text{rec}} + \beta r_{\text{pre}} + \gamma r_{\text{ded}} + \delta r_{\text{ans}}}{\alpha + \beta + \gamma + \delta}$.

Furthermore, to prevent the policy from inflating the graph size by stacking duplicate triples, we define a repetition rate: $\rho_{\text{rep}}(\mathcal{G}) = \dfrac{|\mathcal{T}(\mathcal{G})| - |\,\text{uniq}(\mathcal{T}(\mathcal{G}))|}{|\mathcal{T}(\mathcal{G})|}$ and derive the trajectory return $R(\tau) = r_{\text{task}}(\tau) - \lambda_{\text{rep}} \rho_{\text{rep}}(\mathcal{G}_\tau) + \lambda_{\text{fmt}} \sum_{t=1}^{|\tau|} r_t^{\text{fmt}}$. This directly addresses the issue revealed by works such as HaluMem: errors in memory systems often do not emerge only at the answering stage, but are already written during the extraction and updating phases (Chen et al., 2025a). We employ the standard clipped GRPO objective (Shao et al., 2024) to update the writer.

4.2 Memory Reader: Memory Retrieval Based on Graph Foundation Model

The memory reader must operate stably over graph memory that is continuously updated by the writer. Dense retrievers mainly learn query–document semantic matching and thus struggle to exploit entity roles, bridge paths, and cross-community dependencies, while conventional GNN retrievers are often tied to fixed graph distributions and generalize poorly across domains, users, and evolution stages. We therefore adopt a Graph Foundation Model (GFM) as the memory reader, whose multi-graph pre-training enables transferable structural priors and lightweight calibration on new graphs (Luo et al., 2025; Zhang et al., 2025c). Formally, the memory reader outputs an entity distribution, a document distribution, and an optional retrieval subgraph $f_\phi(q, \mathcal{G}, \mathcal{D}) = \big(p_\phi(e \mid q, \mathcal{G}), p_\phi(d \mid q, \mathcal{G}, \mathcal{D}), \mathcal{G}_q\big)$. Here, $p_\phi(e \mid q, \mathcal{G})$ represents the entity memory activated by the query, $p_\phi(d \mid q, \mathcal{G}, \mathcal{D})$ denotes the final retrieved textual evidence, and $\mathcal{G}_q$ provides an interpretable retrieval path. To obtain a compact and query-aligned activated subgraph, we further introduce a lightweight query-conditioned subgraph selector; implementation details are provided in Appendix I.

Cognition-inspired Structured Query Planning. When humans extract long-term memories, the brain often automatically generates multi-dimensional retrieval cues to anchor the target based on only a vague final intention. Inspired by this, we no longer treat the natural language query as a single retrieval command. Instead, we introduce a planning function $\mathcal{P}_\omega$ to simulate the cue reconstruction process of the human brain before awakening memory, decomposing the initial query into a set of rich associative probes: $\mathcal{P}_\omega(q) = \Big(\mathcal{E}_{\text{exp}}, \mathcal{A}, \mathcal{C}_{\text{rel}}, \mathcal{C}_{\text{hard}}, \tau, \{(\tilde{q}_m, \alpha_m, t_m)\}_{m=1}^{M}\Big)$. Detailed definitions of the notation, additional information, and the concrete prompt templates and output schema are provided in Appendix K. This multi-path concurrent awakening method effectively overcomes the "tip-of-the-tongue phenomenon" (i.e., difficulties in alias alignment or missing bridging entities) and naturally stitches together forgotten implicit relationships (Trivedi et al., 2023; Asai et al., 2024; Wu et al., 2025c; Zhang et al., 2025b).

Soft Addressing and Pre-activation of Memory Fragments. Cognitive neuroscience reveals that human memory retrieval involves not only the extraction of perfectly matching information but also the instinctive awakening of peripherally related memories through *Semantic Priming*. And to address the first challenge, we treat the calculation of the query-conditioned entry score $s_e(q)$ as a comprehensive assessment of the stimulus intensity across different *Memory Engrams*:

$$\begin{aligned} s_e(q) =& \lambda_1 \, \text{Exact}(e, \mathcal{E}_{\text{exp}}) + \lambda_2 \, \text{Alias}(e, \mathcal{A}) + \lambda_3 \max_{m \le M} \cos\left(\text{Emb}(\text{desc}(e)), \text{Emb}(\tilde{q}_m)\right) \\ &+ \lambda_4 \, \text{Type}(e, \tau) + \lambda_5 \, \text{Cons}(e, \mathcal{C}_{\text{hard}}) + \lambda_6 \sum_{\xi \in \text{NER}(q)} \text{EL}(e \mid \xi). \end{aligned} \tag{1}$$

Subsequently, the system employs a Softmax function with a temperature coefficient T_0 to simulate the brain's limited Attention Allocation mechanism during retrieval. This normalizes the multi-dimensional stimulus signals to form the initial activation distribution of the memory atlas $p_0(e \mid$

$q) = \frac{\exp(s_e(q)/T_0)}{\sum_{v\in\mathcal{V}_E}\exp(s_v(q)/T_0)}$. Based on this distribution, we define the initial state of the memory nodes as $\mathbf{h}_e^{(0)} = \big(p_0(e \mid q)\big)^{\eta} W_q \,\mathrm{Emb}(q) + W_x \mathbf{x}_e$. In this process, $\mathbf{x}_e$ acts as the solidified **long-term memory** (static representation of entities) in the brain, while the query vector adjusted by the cognitive recall degree $p_0(e \mid q)$ represents the current **working memory** (task context).

Synapse-inspired Structurally Conditioned Associative Propagation. To address the second challenge while avoiding indiscriminate diffusion, we introduce edge-level vector structural gating in the GFM. The node-level structural features, edge-pair structural features, and graph-level summary are defined as:

$$\phi(v) = \big[\log(1+d_v), c_v, \kappa_v, \bar{d}_{\mathcal{N}(v)}\big], \tag{2}$$

$$\psi(u,v) = \big[|d_u - d_v|, |\mathcal{N}(u)\cap\mathcal{N}(v)|, \mathrm{Jaccard}(\mathcal{N}(u),\mathcal{N}(v))\big], \tag{3}$$

$$\mathbf{r}_{\mathcal{G}} = \big[\mathrm{mean}_{v\in\mathcal{V}_E}\,\phi(v); \mathrm{std}_{v\in\mathcal{V}_E}\,\phi(v); \mathrm{dens}(\mathcal{G})\big]. \tag{4}$$

Detailed definitions and normalization procedures are provided in Appendix L. The edge structural context for the l-th layer is $\mathbf{z}_{uv}^{(l)} = \big[E_n^{(l)}(\phi(u)); E_n^{(l)}(\phi(v)); E_p^{(l)}(\psi(u,v)); E_g^{(l)}(\mathbf{r}_{\mathcal{G}})\big]$, which generates the vector gating $\mathbf{g}_{uv}^{(l)} = \mathbf{1} + \delta\tanh\big(\mathrm{MLP}_g^{(l)}(\mathbf{z}_{uv}^{(l)})\big)$. Let η_{uv} be the normalized adjacency weight with self-loops; the message and node updates are $\mathbf{m}_{u\to v}^{(l)} = \eta_{uv}\,\mathbf{g}_{uv}^{(l)} \odot W_m^{(l)}\mathbf{h}_u^{(l-1)}$, $\mathbf{h}_v^{(l)} = \mathrm{LayerNorm}\left(\mathbf{h}_v^{(l-1)} + \mathrm{PReLU}\left(\mathbf{b}^{(l)} + \sum_{u\in\mathcal{N}(v)}\mathbf{m}_{u\to v}^{(l)}\right)\right)$. Unlike traditional heuristic path expansion, PPR walks, or community summarization (Edge et al., 2024; Guo et al., 2024; Wang & Han, 2025), the system here can actively perform **Inhibition** of non-specific generalized memories (suppressing hub edges), keenly capture and preserve **long-distance associations** across different cognitive clusters (lateral thinking/bridge edge preservation), and undergo **Habituation** (weakening redundant edges) toward highly repetitive local information, much like the human brain.

Traditional query-dependent GNNs or PPR-style expansion can perform multi-hop propagation along graph structures. But the key issue is not simply to expand the propagation range, but to preserve the advantage of query-relevant evidence signal over distractor noise under a limited top-k budget. Proposition 1(i) summarizes this signal–budget view: soft addressing improves the initial evidence activation, structural gating preserves bridge/evidence paths while suppressing noisy neighborhoods, and controlled entity-to-document projection converts the entity-level advantage into more efficient document-level retrieval. Complete definitions, assumptions, and proofs are provided in Appendix B.

Target Graph Calibration and Cross-graph Structural Priors. Human memory, on one hand, reorganizes cues based on the current context, while on the other, it retains relatively stable structured recall habits. In our self-evolving graph memory, each $\mathcal{G}$ generated by the writer per round alters the local topology and noise distribution; therefore, the reader cannot rely solely on propagation patterns from a fixed graph.

Since the writer continuously changes the memory graph, the reader must simultaneously adapt to the current target graph and preserve cross-graph structural priors. This is precisely why we introduce the context–schema decomposition. As summarized in Proposition 1(ii), the schema channel provides a transferable structural prior, while the context channel corrects the target-graph residual induced by the current writer, current domain, entity granularity, and local noise. The complete theoretical motivation is provided in Appendix C and Appendix D.

First, a feature prompt vector $\mathbf{p}_f$ is used for a lightweight calibration of the query-activated input $\tilde{\mathbf{h}}_e^{(0)} = \mathbf{p}_f \odot \mathbf{h}_e^{(0)}$. The contextual calibration channel performs gated propagation on the current graph $\mathcal{G}$: $\mathbf{H}_{\mathrm{ctx}} = F_{\mathrm{gate}}(\tilde{\mathbf{H}}^{(0)}, \mathcal{G}; \Theta_{\mathrm{gate}})$. Here, $\mathbf{H}_{\mathrm{ctx}}$ captures the immediate structural state within the current memory graph. Simultaneously, the schema prior channel maintains a set of cross-graph structural prompt bases $\{\mathbf{P}_j^{(l)}\}_{j=1}^K$, which are used to encode stable reading habits formed during multi-graph training: $\omega_j^{(l)} = \mathrm{softmax}_j(\mathbf{a}^{(l)}/T_p)$, $\mathbf{P}_{\mathrm{schema}}^{(l)} = \sum_{j=1}^K \omega_j^{(l)}\mathbf{P}_j^{(l)}$. Propagation is executed based on these schema prompts to obtain: $\mathbf{H}_{\mathrm{sch}} = F_{\mathrm{prompt}}\big(\tilde{\mathbf{H}}^{(0)}, \mathcal{G}; \{\mathbf{P}_{\mathrm{schema}}^{(l)}\}_{l=1}^L\big)$. The final entity representation is jointly determined by the current context and the long-term schema: $\mathbf{H}(q,\mathcal{G}) = \mathbf{H}_{\mathrm{ctx}} + \beta_{\mathrm{sch}}\mathbf{H}_{\mathrm{sch}}$. Here, $\mathbf{H}_{\mathrm{ctx}}$ is analogous to a context-dependent immediate recall state, responsible for adapting to the specific graph structure generated by the current writer; $\mathbf{H}_{\mathrm{sch}}$ is akin

Table 1: Open-domain retrieval and QA results on NQ and PopQA. We report passage-level Recall (%) at top-2 and top-5 and end-to-end EM/F1 (%) when comparable numbers are available in original papers or later works that reproduce/cite these methods. Best available results are in **bold** and runner-ups are underlined. **Only rows marked with 0-shot are our zero-shot transfer results; baseline rows are not marked as zero-shot.**

Zero-shot setting applies only to SAGE on NQ and PopQA.								
Dataset	NQ				PopQA			
Method	R@2P	R@5P	EM	F1	R@2P	R@5P	EM	F1
BM25 (▷ *SIGIR'94*)	34.7	51.1	24.5	42.13	36.3	55.7	42.4	52.60
Contriever (▷ *TMLR'22*)	29.1†	54.6†	45.0†	58.9†	27.0†	43.2†	41.6†	53.1†
GTR (▷ *EMNLP'22*)	35.0†	63.4†	45.5†	59.9†	40.1†	49.4†	43.2†	56.2†
ColBERTv2 (▷ *NAACL'22*)	35.7	53.3	25.9	42.91	41.2	62.4	46.7	58.05
RAPTOR (▷ *ICLR'24*)	40.3†	68.3†	36.9†	50.7†	40.2†	48.7†	43.1†	56.2†
Proposition (▷ *EMNLP'24*)	33.1⋆	62.2⋆	–	–	–	–	–	–
HippoRAG (▷ *NeurIPS'24*)	21.3†	44.4†	43.0†	55.3†	40.0†	53.8†	42.7†	55.9†
HippoRAG 2 (▷ *ICML'25*)	35.3	53.3	25.7	42.87	39.6	61.7	46.1	57.18
PropRAG (▷ *EMNLP'25*)	–	**77.9**‡	–	**62.5**‡	–	56.2‡	–	56.4‡
SAGE (ours)0-shot	**49.5**0-shot	63.0 0-shot	**46.2**0-shot	54.21 0-shot	**54.9**0-shot	**64.6**0-shot	**53.0**0-shot	**66.06**0-shot

† Values are from the passage-level retrieval and QA evaluation in ▷. ‡ Values are from the Recall@5 and F1 tables in ▷; Recall@2 and EM are not reported there. ⋆ Values are from the reproduced single-step retrieval table in ▷; PopQA and QA EM/F1 are not reported there. ⊛ Zero-shot denotes no NQ/PopQA-specific gradient updates, tuning, or self-evolution.

to a memory schema formed across experiences, retaining the ability to recognize stable patterns such as bridge nodes, community boundaries, core–periphery structures, and noise short-circuits.

Reader Training. Reader training aims to learn cross-graph transferable retrieval biases through a two-stage procedure. First, we perform structural contrastive pre-training on multiple augmented graph views. Then, in the supervised fine-tuning stage, we align these transferable capabilities with question-driven evidence retrieval by training the reader to identify and rank supporting entities for each query using weighted classification and multi-positive ranking objectives. Implementation details are provided in Appendices M and N.

Writer-Reader Self-evolution. To address the third challenge, we propose a self-evolution framework. Each of our self-evolution iterations consists of two phases. First, we fix the reader and train the writer using its retrieval results as rewards. Subsequently, we use the updated writer to generate new graphs and continue training the reader. The overall procedure is detailed in Algorithm 1.

Proposition 1. Theoretical consequences of SAGE

(i) **Signal–budget efficiency.** Soft addressing, structural gating, and controlled entity-to-document projection jointly improve evidence signal over distractor noise, thereby reducing the top-k budget needed for evidence coverage.

[▷ SNR proof] [▷ Budget proof]

(ii) **Context–schema decomposition.** The reader combines transferable structural priors with target-graph calibration, so adaptation only needs to correct graph-specific residuals.

[▷ Proof]

(iii) **Evolution stability.** Under bounded graph drift, consecutive writer updates induce bounded document-score changes.

[▷ Proof]

From a theoretical perspective, this process can be interpreted as approximate coordinate improvement over a joint memory utility: the writer update improves the readability of the graph memory, while the reader update reduces writer-induced graph distribution shift and reward bias.

We provide the full coordinate-improvement result, the surrogate reward bias bound, and the analysis of single-sided update bottlenecks in Appendix F. In addition, Proposition 1(iii) shows that although each writer update changes the graph structure in self-evolving memory, the reader output does not oscillate arbitrarily with graph evolution. We provide detailed training, inference, memory, and selector-regularizer complexity analyses in Appendix J.

5 EXPERIMENTS

This section presents an experimental evaluation centered around four research questions (RQs). ***RQ1***: whether SAGE can bring consistent benefits in tasks such as multi-hop QA and open-domain transfer;

RQ2: whether SAGE is capable of handling long-term conversation history, knowledge updates, and memory hallucination;

RQ3: whether the writer-reader closed loop truly yields self-evolution benefits;

RQ4: further analysis of where and how the performance gains come from specific designs.

Table 2: Performance of representative memory systems on `LongMemEval`. We report accuracy (%) on single-session user (SS-U), single-session assistant (SS-A), merged single-session recall (SSR), single-session preference (SS-P), knowledge update (KU), temporal reasoning (TR), multi-session reasoning (MS), and overall performance. SSR is the sample-weighted average of SS-U and SS-A. Overall is the sample-weighted average over the six original task categories. Best numerical results are in **bold** and runner-ups are underlined. The darker the cell, the better. Results from different reporting protocols are shown together for comparison and should not be treated as a single strict leaderboard.

Dataset	LongMemEval-S / LongMemEval							
Method	SS-U	SS-A	SSR	SS-P	KU	TR	MS	Overall
Other baselines: unified protocol in TiMem (GPT-4o-mini, LLM-judge accuracy)								
MemoryBank (▷)	50.0	9.8	32.1	0.0	21.8	0.0	0.0	11.5
A-MEM (▷)	82.9	87.5	84.9	39.3	72.8	36.1	40.3	55.4
Mem0 (▷)	94.3	51.8	75.4	50.0	78.7	49.2	66.2	65.0
MemoryOS (▷)	81.1	78.2	79.8	51.3	56.1	53.4	44.8	58.1
TiMem (▷)	95.7	82.1	89.7	63.3	86.2	68.4	**70.8**	76.9
Other baselines: MemOS evaluation suite (short-answer prompt)								
MIRIX (▷)	72.8	63.6	68.8	53.3	52.6	25.6	30.1	43.5
Mem0 (▷)	82.9	26.8	57.9	90.0	66.7	72.2	63.1	66.4
Supermemory[2]	85.7	58.9	73.8	90.0	55.1	44.4	52.6	58.4
MemU[3]	67.1	19.6	46.0	76.7	41.0	17.3	42.1	38.4
*Matched baseline–*SAGE *comparisons*								
MemOS (▷)	93.7	67.9	82.2	50.7	76.7	65.1	58.8	68.7
MemOS + SAGE	**97.1**	94.6	96.0	83.3	70.5	71.4	66.2	76.8
MemOS + SAGE (1 round)	96.5	**96.9**	**96.7**	85.2	73.4	71.9	68.2	78.2
memobase[1]	92.8	23.2	61.9	80.0	**89.7**	75.9	66.9	72.4
Memobase + SAGE	95.7	72.9	85.6	87.1	83.1	**79.1**	69.7	79.3
Memobase + SAGE (1 round)	95.5	74.8	86.3	**90.1**	86.0	78.6	68.9	**79.8**
Zep (▷)	92.9	75.0	84.9	53.3	74.4	54.1	47.4	63.2
Zep + SAGE	93.6	78.0	86.7	67.1	76.6	66.2	55.4	70.2
Zep + SAGE (1 round)	94.2	80.8	88.2	69.3	77.9	69.0	59.7	72.8

Datasets. We evaluate SAGE on eight benchmarks spanning three evaluation scenarios. The first category consists of general QA benchmarks and three multi-hop QA benchmarks, including NQ, PopQA, HotpotQA, 2WikiMultiHopQA, and MuSiQue, used to examine whether the system can recover bridge entities across documents and combine evidence and reasoning paths. The second category focuses on a practical e-commerce application scenario, using a Review-Based Question Answering Task: AmazonQA, to assess its value in real e-commerce applications with real noisy reviews. The third category comprises long-term agent memory datasets, including LongMemEval and HaluMem, used to test information extraction from long interaction histories, multi-session reasoning, temporal reasoning, knowledge updating, abstention, and operation-level hallucination. Table 19 summarizes the details of each dataset. Further details on baselines and metrics can be found in Appendix R.

5.1 END-TO-END EFFECTIVENESS

Multi-hop Question Answering Table 3 reports the main results on general QA benchmarks and three multi-hop QA benchmarks. Table 15 reports the results of retrieval performance on multi-hop QA benchmarks. It is worth mentioning that even when we directly test on NQ and PopQA using a model trained only on MuSiQue, HotpotQA, and 2WikiMultiHopQA, SAGE remains competitive on NQ and shows a clearer advantage on PopQA; detailed matched results are reported in Table 1. Full protocol is given in Appendix T.

Domain-specific Memory Table 16 reports the results on AmazonQA. SAGE consistently outperforms the neural baseline R-Net across all metrics, indicating strong cross-task generalization. After training on AmazonQA, Ours achieves substantial gains. Overall, training and interaction rounds steadily enhance performance, while the zero-shot results demonstrate promising transfer ability.

5.2 LONG-TERM AGENT MEMORY EVALUATION

[1] `memobase`: `https://github.com/memodb-io/memobase`.
[2] `Supermemory`: `https://github.com/supermemoryai/supermemory`.
[3] `MemU`: `https://github.com/NevaMind-AI/memU`.

Table 3: Results of multi-hop question answering (QA) performance. We report Exact Match (EM) and F1 score, both reported as percentages (%). Best results are in **bold** and runner-ups are underlined. The darker the cell, the better.

Dataset	HotpotQA		MuSiQue		2WikiMultiHopQA		Avg. Rank
Method	EM	F1	EM	F1	EM	F1	
BM25 (▷ *SIGIR'94*)	40.0	53.2	19.5	23.6	46.9	57.9	15.5
Contriever (▷ *TMLR'22*)	34.9	51.2	16.3	21.6	24.3	33.9	20.5
GTR (▷ *EMNLP'22*)	33.8	51.9	15.1	25.2	33.7	42.5	19.7
ColBERTv2 (▷ *NAACL'22*)	43.4	57.7	15.5	26.4	33.4	43.3	17.3
RAPTOR (▷ *ICLR'24*)	48.2	59.2	17.6	28.9	30.6	42.0	15.8
GraphRAG (▷ *arXiv'24*)	35.3	54.6	13.4	29.5	28.3	46.9	18.0
G-Retriever (▷ *NeurIPS'24*)	33.2	50.3	18.0	25.9	42.3	45.6	18.2
LightRAG (▷ *arXiv'24*)	36.8	48.3	18.1	27.5	45.1	49.5	16.7
HippoRAG (▷ *NeurIPS'24*)	41.8	55.0	19.2	29.8	46.6	59.5	13.8
HippoRAG 2 (▷ *ICML'25*)	57.3	69.6	33.9	40.5	**75.4**	78.2	4.7
SubgraphRAG (▷ *ICLR'25*)	52.6	65.8	33.0	39.2	71.0	76.2	7.5
PropRAG (▷ *EMNLP'25*)	57.2	70.1	**38.9**	41.3	73.0	79.9	3.8
GFM-RAG (▷ *NeurIPS'25*)	51.6	66.9	30.2	40.4	69.8	77.7	7.8
FLARE (▷ *EMNLP'23*)	48.7	60.6	16.2	28.4	46.7	65.4	12.7
Adaptive-RAG (▷ *NAACL'24*)	45.5	59.6	13.8	25.6	48.9	62.8	14.3
ColBERTv2 + IRCoT (▷ *ACL'23*)	45.5	58.4	19.1	30.5	35.4	45.1	14.4
HippoRAG + IRCoT (▷ *ACL'23*)	45.7	59.2	21.9	33.3	47.7	62.7	11.3
GFM-RAG + IRCoT (▷ *ACL'23*)	56.0	71.8	36.6	49.2	72.5	80.8	3.7
SAGE (ours)	51.3	72.7	28.0	47.3	64.5	75.3	7.5
ours +1 round	**60.4**	73.4	32.2	52.3	71.7	79.5	4.2
ours +2 round	56.1	**80.8**	34.2	**53.1**	74.0	80.4	**2.5**
ours + IRCoT	52.3	75.0	35.2	52.6	72.2	**82.9**	3.3

The LongMemEval results are shown in Table 2. The HaluMem results are shown in Table 17. A direct full-stack leaderboard is informative but does not cleanly isolate the contribution of a retrieval substrate: mature memory systems differ in memory extraction, consolidation and other engineering choices. We therefore perform a paired component-level intervention. For each host, we keep its non-retrieval memory pipeline and evaluation protocol fixed and replace its native memory-access/retrieval backend with SAGE. The +SAGE (1 ROUND) variant additionally performs one SAGE writer-reader evolution round under the same host protocol. Detailed information is documented in Appendix S.

Table 4: Retrieval efficiency comparison. We report retrieval time in seconds on HotpotQA, MuSiQue, and 2Wiki. For Time, lower is better. Best results are in **bold** and runner-ups are underlined. The darker the cell, the better.

Dataset	HotpotQA	MuSiQue	2Wiki
Method	Time↓	Time↓	Time↓
Single-step retrieval methods			
ColBERTv2	0.035	**0.030**	0.029
HippoRAG	0.255	0.251	0.158
LightRAG	0.861	1.109	0.911
GraphRAG (MS)	2.759	3.037	1.204
GFM-RAG	0.107	0.124	0.060
Iterative retrieval methods			
IRCoT + ColBERTv2	1.146	1.152	2.095
IRCoT + HippoRAG	3.162	3.104	3.441
SAGE	**0.032**	0.034	**0.019**

On LongMemEval, all three matched hosts improve in overall accuracy after the SAGE substitution and one evolution round. On HaluMem, the same intervention improves both upstream memory operations and downstream QA.

5.3 Ablations, Failure Modes, and Efficiency

As shown in Table 4, SAGE demonstrates a strong speed advantage. SAGE achieves the lowest retrieval-engine latency on HotpotQA and 2WikiMultiHopQA and is a close second to ColBERTv2 on MuSiQue, indicating strong potential for practical and large-scale deployment. To further analyze the interpretability of SAGE, we visualize the retrieved subgraph for a representative case, as shown in Figure 7. A detailed case study is provided in Appendix P.1. Detailed writer-side and reader-side ablations are provided in Appendix G and Appendix H. The coupling ablation in Table 26 separates the contribution of the main writer-reader components. The category-level KU/TR/MS statistics and concrete examples are reported in Appendix U. Table 4 reports *retrieval-engine* latency. A measured audit of training GPU-hours, LM call counts, query planning, NER, entity linking, retrieval, and answer-generation latency is provided in Appendix J.

6 Conclusion

We presented SAGE, a self-evolving agentic graph-memory engine that treats memory as a dynamic substrate for writing, reading, and continual improvement. Experiments show that SAGE improves evidence recovery, grounding, and retrieval efficiency, suggesting that self-evolving graph memory is a promising foundation for long-horizon language agents.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2024.

Sungguk Cha, DongWook Kim, Taeseung Hahn, Mintae Kim, Youngsub Han, and Byoung-Ki Jeon. Annotation-free reinforcement learning query rewriting via verifiable search reward. *arXiv preprint arXiv:2507.23242*, 2025.

Ding Chen, Simin Niu, Kehang Li, Peng Liu, Xiangping Zheng, Bo Tang, Xinchi Li, Feiyu Xiong, and Zhiyu Li. Halumem: Evaluating hallucinations in memory systems of agents. *arXiv preprint arXiv:2511.03506*, 2025a.

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. Dense x retrieval: What retrieval granularity should we use? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15159–15177, 2024.

Yiqun Chen, Lingyong Yan, Weiwei Sun, Xinyu Ma, Yi Zhang, Shuaiqiang Wang, Dawei Yin, Yiming Yang, and Jiaxin Mao. Improving retrieval-augmented generation through multi-agent reinforcement learning. *arXiv preprint arXiv:2501.15228*, 2025b.

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.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tian Ao, and Chao Huang. Lightrag: Simple and fast retrieval-augmented generation. *arXiv preprint arXiv:2410.05779*, 2(3), 2024.

Mansi Gupta, Nitish Kulkarni, Raghuveer Chanda, Anirudha Rayasam, and Zachary C Lipton. Amazonqa: A review-based question answering task. *arXiv preprint arXiv:1908.04364*, 2019.

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. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 21497–21515. PMLR, 2025.

Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh V Chawla, Thomas Laurent, Yann LeCun, Xavier Bresson, and Bryan Hooi. G-retriever: Retrieval-augmented generation for textual graph understanding and question answering. *Advances in Neural Information Processing Systems*, 37:132876–132907, 2024.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, 2020.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in llm agents via incremental multi-turn interactions. *arXiv preprint arXiv:2507.05257*, 2025.

Ziniu Hu, Yuxiao Dong, Kuansan Wang, Kai-Wei Chang, and Yizhou Sun. Gpt-gnn: Generative pre-training of graph neural networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1857–1867, 2020.

Zhengjun Huang, Zhoujin Tian, Qintian Guo, Fangyuan Zhang, Yingli Zhou, Di Jiang, Zeying Xie, and Xiaofang Zhou. Licomemory: Lightweight and cognitive agentic memory for efficient long-term reasoning. *arXiv preprint arXiv:2511.01448*, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*, 2021.

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 (Volume 1: Long Papers)*, pp. 7036–7050, 2024.

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. 25961–25970, 2025. doi: 10.18653/v1/2025.emnlp-main.1318.

Youmin Ko, Sungjong Seo, and Hyunjoon Kim. Cooperative retrieval-augmented generation for question answering: Mutual information exchange and ranking by contrasting layers. In *Advances in Neural Information Processing Systems*, 2025.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276.

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.

Kai Li, Xuanqing Yu, Ziyi Ni, Yi Zeng, Yao Xu, Zheqing Zhang, Xin Li, Jitao Sang, Xiaogang Duan, Xuelei Wang, et al. Timem: Temporal-hierarchical memory consolidation for long-horizon conversational agents. *arXiv preprint arXiv:2601.02845*, 2026a.

Mufei Li, Siqi Miao, and Pan Li. Simple is effective: The roles of graphs and large language models in knowledge-graph-based retrieval-augmented generation. *arXiv preprint arXiv:2410.20724*, 2024.

Yifei Li, Weidong Guo, Lingling Zhang, Rongman Xu, Muye Huang, Hui Liu, Lijiao Xu, Yu Xu, and Jun Liu. Locomo-plus: Beyond-factual cognitive memory evaluation framework for llm agents. *arXiv preprint arXiv:2602.10715*, 2026b.

Zhiyu Li, Shichao Song, Hanyu Wang, Simin Niu, Ding Chen, Jiawei Yang, Chenyang Xi, Huayi Lai, Jihao Zhao, Yezhaohui Wang, et al. Memos: An operating system for memory-augmented generation (mag) in large language models. *arXiv preprint arXiv:2505.22101*, 2025.

Jiawei Liu, Cheng Yang, Zhiyuan Lu, Junze Chen, Yibo Li, Mengmei Zhang, Ting Bai, Yuan Fang, Lichao Sun, Philip S Yu, et al. Graph foundation models: Concepts, opportunities and challenges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Linhao Luo, Zicheng Zhao, Gholamreza Haffari, Dinh Phung, Chen Gong, and Shirui Pan. Gfm-rag: graph foundation model for retrieval augmented generation. *arXiv preprint arXiv:2502.01113*, 2025.

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.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, 2023. doi: 10.18653/v1/2023.acl-long.546.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, et al. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9844–9855, 2022.

Charles Packer, Vivian Fang, Shishir G. Patil, Kevin Lin, Sarah Wooders, and Joseph E. Gonzalez. Memgpt: Towards llms as operating systems. *CoRR*, abs/2310.08560, 2023. doi: 10.48550/ARXIV.2310.08560. URL https://doi.org/10.48550/arXiv.2310.08560.

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.

Jiezhong Qiu, Qibin Chen, Yuxiao Dong, Jing Zhang, Hongxia Yang, Ming Ding, Kuansan Wang, and Jie Tang. Gcc: Graph contrastive coding for graph neural network pre-training. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1150–1160, 2020.

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.

Stephen E Robertson and Steve Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*, pp. 232–241. Springer, 1994.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. Colbertv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3715–3734, 2022.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 10014–10037, 2023.

Hong Ting Tsang, Jiaxin Bai, Haoyu Huang, Qiao Xiao, Tianshi Zheng, Baixuan Xu, Shujie Liu, and Yangqiu Song. Autograph-r1: End-to-end reinforcement learning for knowledge graph construction. *arXiv preprint arXiv:2510.15339*, 2025.

Jingjin Wang and Jiawei Han. PropRAG: Guiding retrieval with beam search over proposition paths. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6212–6227, 2025. doi: 10.18653/v1/2025.emnlp-main.317.

Yu Wang and Xi Chen. Mirix: Multi-agent memory system for llm-based agents. *arXiv preprint arXiv:2507.07957*, 2025.

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

Yaxiong Wu, Sheng Liang, Chen Zhang, Yichao Wang, Yongyue Zhang, Huifeng Guo, Ruiming Tang, and Yong Liu. From human memory to ai memory: A survey on memory mechanisms in the era of llms. *arXiv preprint arXiv:2504.15965*, 2025b.

Yaxiong Wu, Yongyue Zhang, Sheng Liang, and Yong Liu. Sgmem: Sentence graph memory for long-term conversational agents. *arXiv preprint arXiv:2509.21212*, 2025c.

Tianyang Xu, Haojie Zheng, Chengze Li, Haoxiang Chen, Yixin Liu, Ruoxi Chen, and Lichao Sun. Noderag: Structuring graph-based rag with heterogeneous nodes. *arXiv preprint arXiv:2504.11544*, 2025a.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110*, 2025b.

Chang Yang, Chuang Zhou, Yilin Xiao, Su Dong, Luyao Zhuang, Yujing Zhang, Zhu Wang, Zijin Hong, Zheng Yuan, Zhishang Xiang, et al. Graph-based agent memory: Taxonomy, techniques, and applications. *arXiv preprint arXiv:2602.05665*, 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.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *Advances in neural information processing systems*, 33: 5812–5823, 2020.

Xingtong Yu, Zechuan Gong, Chang Zhou, Yuan Fang, and Hui Zhang. Samgpt: Text-free graph foundation model for multi-domain pre-training and cross-domain adaptation. In *Proceedings of the ACM on Web Conference 2025*, pp. 1142–1153, 2025.

Juwei Yue, Chuanrui Hu, Jiawei Sheng, Zuyi Zhou, Wenyuan Zhang, Tingwen Liu, Li Guo, and Yafeng Deng. Hypermem: Hypergraph memory for long-term conversations. *arXiv preprint arXiv:2604.08256*, 2026.

Guibin Zhang, Muxin Fu, Guancheng Wan, Miao Yu, Kun Wang, and Shuicheng Yan. G-memory: Tracing hierarchical memory for multi-agent systems. *arXiv preprint arXiv:2506.07398*, 2025a.

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. *arXiv preprint arXiv:2510.10397*, 2025b.

Qinggang Zhang, Shengyuan Chen, Yuanchen Bei, Zheng Yuan, Huachi Zhou, Zijin Hong, Hao Chen, Yilin Xiao, Chuang Zhou, Junnan Dong, et al. A survey of graph retrieval-augmented generation for customized large language models. *arXiv preprint arXiv:2501.13958*, 2025c.

Yibo Zhao, Jiapeng Zhu, Ye Guo, Kangkang He, and Xiang Li. Eˆ 2graphrag: Streamlining graph-based rag for high efficiency and effectiveness. *arXiv preprint arXiv:2505.24226*, 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.

Algorithm 1: Writer-Reader Self-evolution Training for SAGE

Input: Training set $\mathcal{D}_{\text{train}}$, writer π_{θ_0}, GFM reader f_{ϕ_0}, self-evolution iterations T
Output: Trained writer π_{θ_T} and reader f_{ϕ_T}

for $t = 0, \ldots, T-1$ **do**
  // Writer update: fixed GFM reader as reward environment
  for *each sample* $x = (q, \mathcal{D}, \mathcal{D}^+, y) \in \mathcal{D}_{\text{train}}$ **do**
    Sample G graph construction trajectories $\{\tau_i\}_{i=1}^{G}$ from π_{θ_t};
    for $i = 1, \ldots, G$ **do**
      Obtain graph $\mathcal{G}_i$ and retrieve $P_k(q, \mathcal{G}_i)$ using f_{ϕ_t};
      Calculate return R_i;
    end
    Update writer π_{θ_t};
  end
  // Reader update: improved graphs as memory substrate
  Construct a set of graph memories $\{\mathcal{G}_x\}$ for the training corpus using $\pi_{\theta_{t+1}}$;
  Update GFM reader f_{ϕ_t} on $\{\mathcal{G}_x\}$;
end

A Additional Analysis of the Memory Writer

This appendix provides a detailed analysis of the memory writer experiments in the main text.

A.1 Reward Design and Writer Behavior

In Table 5, different RL rewards induce different writer behaviors. GFM-pretrained-only achieves Precision/Recall/Deducible of $0.838/0.818/0.510$, while GFM-finetuned achieves $0.824/0.813/0.512$, indicating that relying solely on supervised finetuning cannot stably improve the utility of graph memory for a frozen reader. This result is also consistent with our setup: the goal of the memory writer is not to reproduce a static graph format.

RL-Recall improves Precision and Recall to $0.889/0.835$, but Deducible drops to 0.502. This shows that rewarding only supporting context coverage encourages the writer to store more locally relevant evidence, but does not necessarily lead to a complete multi-hop reasoning chain. RL-F1 further raises Recall to 0.881, but Deducible is only 0.497, again indicating a gap between retrieval matching quality and answer deducibility: the reader hitting the supporting contexts does not guarantee that these contexts are organized in a way sufficient to support answer reasoning. In contrast, RL-Deduce achieves $0.861/0.892/0.517$, showing that using answer deducibility directly as feedback can encourage the writer to focus more on bridging entities, cross-document relations, and answer-relevant causal or attribute paths.

RL-Hybrid achieves Precision and Recall of $\mathbf{0.902}$ and $\mathbf{0.917}$, respectively, representing improvements of $+0.064$ and $+0.099$ over pretrained-only, while Deducible reaches 0.522. This indicates that hybrid rewards can mitigate the bias of a single reward: they both avoid the introduction of too much weakly relevant evidence caused by a pure recall reward and prevent a pure deducibility reward from overfavoring short paths or local answer clues. Hybrid + frozen answer API achieves the highest Deducible, at $\mathbf{0.526}$, but Precision and Recall drop to $0.832/0.874$. This suggests that stronger answer-side feedback can further improve reasoning usability, but it may also make the writer more conservative, writing only evidence directly related to the final answer and thereby sacrificing some supporting context coverage.

A.2 Cross-domain Transfer

Table 6 shows that the base writer trained on HotpotQA/MuSiQue has a certain degree of transferability to new domains, but training on the target domain remains very important. On GR-Bench, Base→GRBench achieves Precision/Recall/Deducible of $0.575/0.609/0.411$, while GR-Bench train→val improves to $0.794/0.833/0.596$. This improvement indicates that the writing strategy learned for multi-hop QA can transfer to structured product or domain graph memory tasks,

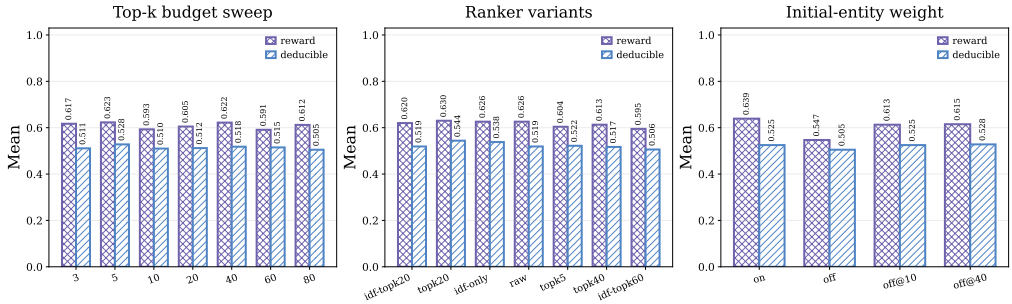


Figure 3: Freeze the sensitivity analysis on the reader side. The impact of the initial-entity weight on reward and Deducible is the most stable; the top-k and ranker variants exhibit a non-monotonic budget–noise trade-off.

but the entity types, attribute relations, and evidence granularity in the target domain still need to be re-adapted.

On HaluMem and LongMemEval, cross-domain differences are even more pronounced. The Base→HaluMem results are $0.230/0.448/0.299$, which improve to $0.312/0.708/0.438$ after training on the target domain; the Base→LongMemEval results are $0.232/0.376/0.475$, which improve to $0.377/0.439/0.531$ after training on the target domain. These results indicate that memory writing in agent memory tasks requires not only extracting explicit facts, but also maintaining user preferences, temporal order, state updates, and long-term consistency. In traditional multi-hop QA, supporting contexts often form a relatively static set of evidence centered around a single question, whereas information in long-term memory tasks changes over time and involves personalization, conflict updates, and context dependence. Therefore, although reader-aware RL feedback can provide transferable writing principles, interaction feedback from the target domain remains crucial for achieving stable performance.

A.3 Writing Protocol and Interaction Budget

Table 7 shows that the writing protocol significantly changes the trade-off among Precision, Recall, and Deducible. The results for Tight=True are $0.836/0.806/0.515$; the results for Tight=False are $0.845/0.851/0.506$. After relaxing the protocol, Recall improves noticeably, indicating that the writer can write more potentially relevant evidence; however, Deducible declines, suggesting that the additional evidence also contains more noise, redundant facts, or weakly related local information. Although this content may increase the coverage of supporting context, it can dilute the reasoning chain that truly supports the answer.

Iterative writing further highlights the role of the interaction budget. For Iterative, 12 turns, tight, Precision/Recall/Deducible are $0.852/0.829/0.516$; after increasing to 20 turns, Recall reaches the highest value of $\mathbf{0.881}$, indicating that multi-turn reader feedback helps the writer complete cross-document bridging paths. For Iterative, 24 turns, loose, Precision and Deducible reach $\mathbf{0.863}$ and $\mathbf{0.531}$, respectively, but Recall falls back to 0.826. This shows that more rounds of interaction are not simply "the longer, the better": the benefit comes from the writer revising the graph structure based on reader feedback, whereas when the protocol is too loose or the writing space becomes too large, the additional content may alter the reader's ranking, causing some gold supporting contexts to be pushed out of the top results.

A.4 Reader-side Sensitivity

Figure 3 analyzes the impact of the frozen reader setting on writer training results. First, the top-k budget sweep shows a non-monotonic trend: at $k = 5$, reward and Deducible are $0.623/0.528$, the better setting in this group; at $k = 40$, reward remains at 0.622, but Deducible drops to 0.518; at $k = 60$, reward further declines to 0.591. This indicates that expanding the retrieval budget does not necessarily lead to better reader feedback. Although a larger top-k improves potential coverage, it

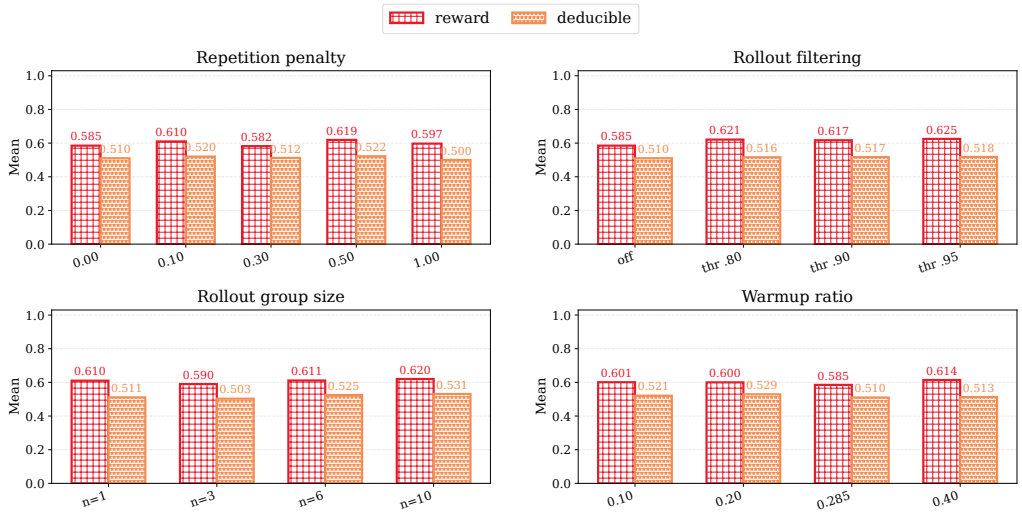


Figure 4: Training regularization and scaling analysis. A larger rollout group size and a moderate warmup ratio show a slight advantage in this batch of results, but the gains are smaller than the effects of reward design, the reader-side initial-entity weight, and the writing protocol.

also introduces more weakly related or redundant evidence, diluting the reasoning chain that truly supports the answer.

The ranker variants also reflect a similar coverage–noise trade-off. topk20 achieves the highest reward and Deducible, at $0.630/0.544$, respectively; idf-only and raw both have a reward of 0.626, but their Deducible scores are 0.538 and 0.519, respectively; idf-topk60 declines to $0.595/0.506$. This shows that the reader ranker cannot rely solely on entity overlap or on expanding the candidate set, but instead must strike a balance among entity matching, semantic relevance, and contextual compactness. For the writer, an overly weak ranker makes it difficult for effective graph structure to be read out, while an overly broad candidate space amplifies the negative impact of noisy writes.

Initial-entity weight is the most stable factor among the three groups of reader-side settings. When initial-entity weight is enabled, reward reaches 0.639 and Deducible is 0.525; when it is disabled, reward drops significantly to 0.547 and Deducible falls to 0.505. Even when the budget is increased after disabling it, the rewards for off@10 and off@40 recover only to 0.613 and 0.615. This indicates that the initial entity anchor is crucial in multi-hop graph retrieval: it helps the reader enter the correct local subgraph from the question entity and expand along the bridging relations written by the writer to the evidence supporting the answer. Without this anchor, simply increasing the retrieval budget cannot fully compensate for the deviation in graph traversal direction.

A.5 Training Stability and Regularization

Figure 4 shows the effects of training regularization and rollout settings on the writer. First, the repetition penalty affects both reward and Deducible, but the trend is not monotonic. Without any penalty, the result is $0.585/0.510$; with a penalty of 0.10, it improves to $0.610/0.520$; with a penalty of 0.50, it reaches the best result in this group at $0.619/0.522$; after further increasing it to 1.00, it drops to $0.597/0.500$. This indicates that a moderate penalty on repeated triples can suppress redundant edges and cyclic expressions, but an overly strong penalty may limit the writer's necessary restatement of key facts. Especially in multi-hop reasoning, the same bridging entity often needs to appear in multiple relational paths, so repetition is not always meaningless noise.

Rollout filtering brings consistent but limited gains. When filtering is disabled, reward/Deducible is $0.585/0.510$; after applying `thr_80`, `thr_90`, and `thr_95`, reward increases to 0.621, 0.617, and 0.625, respectively, while Deducible remains stable at 0.516–0.518. This suggests that filtering out low-quality rollouts can reduce the interference of noisy trajectories with policy updates, allowing the writer to learn effective writing strategies more stably. However, the differences between thresholds

are small, indicating that the main role of filtering is to remove obviously negative samples rather than determine the final performance ceiling.

Rollout group size and warmup ratio further affect training stability. As group size increases from $n = 1$ to $n = 10$, reward/Deducible rises from $0.610/0.511$ to $0.620/0.531$, indicating that a larger group size can provide more reliable relative preference estimates and help RL distinguish more accurately between effective and ineffective writing. The optimal warmup ratio appears at 0.20, where Deducible reaches 0.529; too little warmup may lead to unstable early policy updates, while too much warmup may delay the effect of the RL signal. Overall, these regularization and training scale settings can improve stability, but their gains are smaller than those from reward design, reader-side initial-entity weight, and the interaction protocol itself. This shows that the core improvement of the memory writer comes from reader-aware RL feedback: it forces the graph constructor to learn to preserve bridging entities, cross-document relations, and evidence chains that support answer derivation, while also reducing repetitive structures and irrelevant local facts.

B Signal-to-Noise Ratio and Retrieval Budget of Structurally Gated Propagation

This section analyzes the structural capability of the GFM memory reader in SAGE from the perspective of signal propagation and retrieval budget. Unlike graph-isomorphism expressivity analyses centered on k-WL, we focus on the following question: on noisy graph memories dynamically written by the memory writer, how do soft addressing, structurally gated propagation, context–schema dual-channel calibration, and entity-to-document projection jointly improve the ratio of query-relevant evidence signal to distractor noise, thereby reducing the top-k retrieval budget required to achieve a given level of evidence coverage?

B.1 Review of the SAGE-GFM Reader Formalization

Given a sample $x = (q, D, D^+, y)$, where q denotes the query, $D = \{d_i\}_{i=1}^N$ denotes the candidate memory fragments, and $D^+ \subseteq D$ denotes the gold evidence set supporting the answer y, the memory writer constructs a heterogeneous graph

$$G = W_\theta(q, D) = (V_E \cup V_D, E_{EE} \cup E_{ED}), \tag{5}$$

where V_E is the set of entity nodes, V_D is the set of memory-fragment nodes, E_{EE} denotes entity–entity relation edges, and E_{ED} denotes entity–text-fragment anchoring edges. The GFM memory reader outputs an entity distribution, a document distribution, and an optional activated subgraph:

$$f_\phi(q, G, D) = \big(p_\phi(e \mid q, G),\ p_\phi(d \mid q, G, D),\ G_q\big). \tag{6}$$

The reader first uses query planning and soft addressing to generate query-conditioned initial activation for entities. Let $s_e(q)$ denote the entry score of entity e, which integrates multiple cues such as explicit entities, aliases, pseudo-query similarity, answer type, hard constraints, and entity linking. The initial activation distribution is then given by

$$p_0(e \mid q) = \frac{\exp(s_e(q)/T_0)}{\sum_{v \in V_E} \exp(s_v(q)/T_0)}. \tag{7}$$

The initial entity representation is written as

$$h_e^{(0)} = \big(p_0(e \mid q)\big)^\eta W_q \,\mathrm{Emb}(q) + W_x x_e, \qquad 0 \le \eta \le 1. \tag{8}$$

The reader then constructs structural gates using node-level structural features, edge-pair structural features, and graph-level structural summaries. Specifically, let

$$\varphi(v) = \big(\log(1 + d_v),\ c_v,\ \kappa_v,\ \bar{d}_{\mathcal{N}(v)}\big), \tag{9}$$

$$\psi(u, v) = \big(|d_u - d_v|,\ |\mathcal{N}(u) \cap \mathcal{N}(v)|,\ \mathrm{Jaccard}(\mathcal{N}(u), \mathcal{N}(v))\big), \tag{10}$$

$$r_G = \big(\mathrm{mean}_{v \in V_E}\, \varphi(v);\ \mathrm{std}_{v \in V_E}\, \varphi(v);\ \mathrm{dens}(G)\big). \tag{11}$$

The edge structural context at layer l is

$$z_{uv}^{(l)} = \big(E_n^{(l)}(\varphi(u));\ E_n^{(l)}(\varphi(v));\ E_p^{(l)}(\psi(u, v));\ E_g^{(l)}(r_G)\big), \tag{12}$$

which generates the vector-valued gate

$$g_{uv}^{(l)} = 1 + \delta \tanh\left(\mathrm{MLP}_g^{(l)}(z_{uv}^{(l)})\right). \tag{13}$$

Let $\eta_{uv} \geq 0$ denote the normalized adjacency weight with self-loops. The message and node update are

$$m_{u\to v}^{(l)} = \eta_{uv} g_{uv}^{(l)} \odot W_m^{(l)} h_u^{(l-1)}, \tag{14}$$

$$h_v^{(l)} = \mathrm{LayerNorm}\left(h_v^{(l-1)} + \mathrm{PReLU}\left(b^{(l)} + \sum_{u\in\mathcal{N}(v)} m_{u\to v}^{(l)}\right)\right). \tag{15}$$

In addition, the reader combines a contextual calibration channel on the current graph with a cross-graph schema prior channel:

$$H(q, G) = H_{\mathrm{ctx}} + \beta_{\mathrm{sch}} H_{\mathrm{sch}}. \tag{16}$$

B.2 Recoverable Evidence Region and Effective Signal-to-Noise Ratio

Definition B.1 (Recoverable Evidence Region)**.** Fix a query q and the current memory graph G. Let $R_q \subseteq V_E$ denote the recoverable evidence region under query q, namely the set of entities jointly determined by the current graph structure, anchoring edges, and reader-reachable paths. This set contains nodes that support the answer, connect supporting documents, or serve as bridge entities. If the entity anchor set of document d is denoted by $A(d) \subseteq V_E$, then the anchor coverage of the current graph over the gold evidence is defined as

$$\rho_A = \frac{|\{d \in D^+ : A(d) \cap R_q \neq \varnothing\}|}{|D^+|}. \tag{17}$$

Definition B.2 (Query-Relevant Scalar Activation)**.** Let r_q be the direction induced by the query representation or the final scoring head. The nonnegative query-relevant activation of node v at layer l is defined as

$$a_v^{(l)} = \left[\langle r_q, h_v^{(l)}\rangle\right]_+, \tag{18}$$

where $[t]_+ = \max\{t, 0\}$. The evidence signal mass, noise mass, and effective signal-to-noise ratio at layer l are respectively defined as

$$S_l = \sum_{v\in R_q} a_v^{(l)}, \tag{19}$$

$$N_l = \sum_{v\in V_E\setminus R_q} a_v^{(l)}, \tag{20}$$

$$\mathrm{SNR}_l = \frac{S_l}{N_l}, \tag{21}$$

with the convention that $\mathrm{SNR}_l = +\infty$ when $N_l = 0$.

Remark B.3 (Scope of the Scalar-Channel Analysis). Equation (18) does not assume that the full vector update in Eq. (15) is a nonnegative linear recurrence in every coordinate. LayerNorm, PReLU, residual connections, and linear transformations may all change representation directions. We only analyze the query-relevant channel on which the final retrieval score depends, and absorb the additional effects caused by nonlinearities and directional shifts into a perturbation term. This avoids an overly strong coordinate-wise monotonicity assumption.

B.3 Aggregate Propagation Assumptions and Structural Gating Coefficients

Idealized analyses often assume that every evidence edge has a uniform lower gate bound g_+ and every noisy edge has a uniform upper gate bound g_-. However, in graph memories dynamically constructed by an LLM writer, edges may be missing, erroneous, or repeated; some evidence edges may be underestimated, while some distractor edges may receive high gates. We therefore adopt an aggregate propagation assumption.

Assumption B.4 (Query-Relevant Effective Propagation Operator)**.** Fix a query q, the current graph memory G, and the reader representation at layer l. For each layer $l \in \{1, \dots, L\}$, there exists a nonnegative matrix $T_l \in \mathbb{R}_{\geq 0}^{|V_E| \times |V_E|}$ and a nonnegative perturbation vector $\epsilon_l \in \mathbb{R}_{\geq 0}^{|V_E|}$ such that the query-relevant activation vector at layer l,

$$a^{(l)} = \left(a_v^{(l)}\right)_{v \in V_E},$$

is controlled by the previous-layer activation $a^{(l-1)}$ in the following coordinate-wise sense:

$$a^{(l)} \preceq T_l a^{(l-1)} + \epsilon_l. \tag{22}$$

Here, $\preceq$ denotes coordinate-wise inequality. The operator T_l denotes the effective propagation operator induced by the l-th layer on the query-relevant scalar channel. It absorbs the combined effects of normalized adjacency weights η_{uv}, structural gates $g_{uv}^{(l)}$, message projection $W_m^{(l)}$, context–schema representation composition, and final scoring-channel projection into a single nonnegative propagation kernel. In other words, $T_l(u, v)$ is the effective nonnegative contribution strength of the query-relevant activation of node v at the previous layer to node u at layer l.

The perturbation term ϵ_l absorbs residual effects that are difficult to exactly characterize by nonnegative linear propagation, including LayerNorm, PReLU, residual connections, vector-direction rotation, scoring-channel mismatch, and finite-parameter approximation error.

Furthermore, we only require the propagation process to preserve effective signal in the evidence region in an aggregate sense. We do not require the structural gate to perfectly distinguish every evidence edge from every noisy edge. The operator T_l may allow some evidence edges to be underestimated and some distractor edges to be overestimated; the aggregate propagation coefficients defined below only characterize the overall effect of these local errors on the evidence and noise regions.

Let

$$\bar{R}_q = V_E \setminus R_q.$$

Partition T_l according to the node sets R_q and $\bar{R}_q$:

$$T_l = \begin{pmatrix} T_{RR}^{(l)} & T_{R\bar{R}}^{(l)} \\ T_{\bar{R}R}^{(l)} & T_{\bar{R}\bar{R}}^{(l)} \end{pmatrix}. \tag{23}$$

Here, $T_{RR}^{(l)}$ denotes effective propagation within the evidence region, $T_{\bar{R}\bar{R}}^{(l)}$ denotes effective propagation within the noise region, $T_{\bar{R}R}^{(l)}$ denotes leakage propagation from the evidence region to the noise region, and $T_{R\bar{R}}^{(l)}$ denotes propagation from the noise region to the evidence region.

Definition B.5 (Aggregate Propagation Coefficients)**.** Given the effective propagation operator T_l at layer l and its block decomposition in Eq. (23), define the three aggregate propagation coefficients A_l, B_l, and C_l as follows.

A_l is the evidence-retention coefficient. It characterizes the minimum fraction of total mass that remains inside the evidence region R_q after any nonnegative evidence signal x propagates one layer within R_q. Formally, A_l is any nonnegative constant satisfying

$$A_l \leq \inf_{x \in \mathbb{R}_{\geq 0}^{|R_q|},\ \mathbf{1}^\top x > 0} \frac{\mathbf{1}^\top T_{RR}^{(l)} x}{\mathbf{1}^\top x}. \tag{24}$$

Equivalently, for any nonnegative evidence signal $x \in \mathbb{R}_{\geq 0}^{|R_q|}$ with $\mathbf{1}^\top x > 0$,

$$\mathbf{1}^\top T_{RR}^{(l)} x \geq A_l \, \mathbf{1}^\top x.$$

B_l is the noise self-propagation coefficient. It characterizes the maximum extent to which any nonnegative noise signal y can be retained or expanded after one layer of propagation inside the noise region $\bar{R}_q$. Formally, B_l is any nonnegative constant satisfying

$$B_l \geq \sup_{y \in \mathbb{R}_{\geq 0}^{|\bar{R}_q|},\ \mathbf{1}^\top y > 0} \frac{\mathbf{1}^\top T_{\bar{R}\bar{R}}^{(l)} y}{\mathbf{1}^\top y}. \tag{25}$$

Equivalently, for any nonnegative noise signal $y \in \mathbb{R}_{\geq 0}^{|\bar{R}_q|}$ with $\mathbf{1}^\top y > 0$,

$$\mathbf{1}^\top T_{\bar{R}\bar{R}}^{(l)} y \leq B_l \, \mathbf{1}^\top y.$$

C_l is the evidence-to-noise leakage coefficient. It characterizes the maximum fraction of an arbitrary nonnegative signal in the evidence region that can leak into the non-evidence region $\bar{R}_q$ after one layer of propagation. Formally, C_l is any nonnegative constant satisfying

$$C_l \geq \sup_{x \in \mathbb{R}_{\geq 0}^{|R_q|},\, \mathbf{1}^\top x > 0} \frac{\mathbf{1}^\top T_{\bar{R}R}^{(l)} x}{\mathbf{1}^\top x}. \tag{26}$$

Equivalently, for any nonnegative evidence signal $x \in \mathbb{R}_{\geq 0}^{|R_q|}$ with $\mathbf{1}^\top x > 0$,

$$\mathbf{1}^\top T_{\bar{R}R}^{(l)} x \leq C_l \, \mathbf{1}^\top x.$$

Finally, let

$$\xi_l = \mathbf{1}^\top \epsilon_{l,\bar{R}} \tag{27}$$

denote the total perturbation mass injected into the noise region $\bar{R}_q$ at layer l by nonlinearities, normalization, representation-direction shifts, and approximation errors. Here, $\epsilon_{l,\bar{R}}$ denotes the restriction of the perturbation vector ϵ_l to $\bar{R}_q$.

Lemma B.6 (Aggregate Propagation Recurrence)**.** *Under Assumption B.4 and Definition B.5, if $S_{l-1} > 0$ and $N_{l-1} \geq 0$, then layer l satisfies*

$$S_l \geq A_l S_{l-1}, \tag{28}$$

$$N_l \leq B_l N_{l-1} + C_l S_{l-1} + \xi_l. \tag{29}$$

Proof. Let $a_R^{(l-1)}$ and $a_{\bar{R}}^{(l-1)}$ be the restrictions of $a^{(l-1)}$ to R_q and $\bar{R}_q$, respectively. By the definition of the evidence-retention coefficient in Eq. (24), the total mass retained within the evidence region through $R_q \to R_q$ propagation is at least

$$\mathbf{1}^\top T_{RR}^{(l)} a_R^{(l-1)} \geq A_l \mathbf{1}^\top a_R^{(l-1)} = A_l S_{l-1}, \tag{30}$$

and thus $S_l \geq A_l S_{l-1}$.

On the other hand, the layer-l mass in the noise region can be upper-bounded by three terms: noise self-propagation, evidence leakage, and perturbation:

$$N_l \leq \mathbf{1}^\top T_{\bar{R}\bar{R}}^{(l)} a_{\bar{R}}^{(l-1)} + \mathbf{1}^\top T_{\bar{R}R}^{(l)} a_R^{(l-1)} + \mathbf{1}^\top \epsilon_{l,\bar{R}}. \tag{31}$$

Using Eqs. (25), (26), and (27), we obtain

$$N_l \leq B_l N_{l-1} + C_l S_{l-1} + \xi_l. \tag{32}$$

This proves the lemma. □

B.4 Realistic Aggregate Signal-to-Noise Ratio Bound

Theorem B.7 (Realistic Aggregate SNR Bound)**.** *Assume that for all $l = 1, \ldots, L$, there exist $A_l > 0$, $B_l \geq 0$, $C_l \geq 0$, and $\xi_l \geq 0$ such that the recurrences in Eqs.* (28)–(29) *hold. Let*

$$Q_l = \mathrm{SNR}_l^{-1} = \frac{N_l}{S_l}. \tag{33}$$

Then

$$Q_L \leq \left(\prod_{l=1}^{L} \frac{B_l}{A_l} \right) Q_0 + \sum_{i=1}^{L} \left(\frac{C_i}{A_i} + \frac{\xi_i}{A_i S_{i-1}} \right) \prod_{t=i+1}^{L} \frac{B_t}{A_t}. \tag{34}$$

Equivalently, if the right-hand side is finite, then

$$\mathrm{SNR}_L \geq \left[\left(\prod_{l=1}^{L} \frac{B_l}{A_l} \right) \mathrm{SNR}_0^{-1} + \sum_{i=1}^{L} \left(\frac{C_i}{A_i} + \frac{\xi_i}{A_i S_{i-1}} \right) \prod_{t=i+1}^{L} \frac{B_t}{A_t} \right]^{-1}. \tag{35}$$

The empty product is defined as 1.

Proof. By Lemma B.6, for any l,

$$S_l \geq A_l S_{l-1}, \qquad N_l \leq B_l N_{l-1} + C_l S_{l-1} + \xi_l. \tag{36}$$

Therefore,

$$Q_l = \frac{N_l}{S_l} \leq \frac{B_l N_{l-1} + C_l S_{l-1} + \xi_l}{A_l S_{l-1}} \tag{37}$$

$$= \frac{B_l}{A_l} Q_{l-1} + \frac{C_l}{A_l} + \frac{\xi_l}{A_l S_{l-1}}. \tag{38}$$

Let

$$r_l = \frac{B_l}{A_l}, \qquad d_l = \frac{C_l}{A_l} + \frac{\xi_l}{A_l S_{l-1}}. \tag{39}$$

Then

$$Q_l \leq r_l Q_{l-1} + d_l. \tag{40}$$

Expanding this first-order nonhomogeneous recurrence yields

$$Q_L \leq \left(\prod_{l=1}^{L} r_l \right) Q_0 + \sum_{i=1}^{L} d_i \prod_{t=i+1}^{L} r_t. \tag{41}$$

Substituting back r_l and d_l proves Eq. (34). Since $\mathrm{SNR}_L = Q_L^{-1}$, Eq. (35) follows. □

Corollary B.8 (Layer-Homogeneous Case)**.** *If $A_l = A > 0$, $B_l = B \geq 0$, $C_l = C \geq 0$, and $\xi_l = 0$, then*

$$\mathrm{SNR}_L \geq \left[\left(\frac{B}{A} \right)^L \mathrm{SNR}_0^{-1} + \frac{C}{A} \sum_{i=0}^{L-1} \left(\frac{B}{A} \right)^i \right]^{-1}. \tag{42}$$

If further $C = 0$, then

$$\mathrm{SNR}_L \geq \left(\frac{A}{B} \right)^L \mathrm{SNR}_0 \,. \tag{43}$$

Proof. Substituting $A_l = A$, $B_l = B$, $C_l = C$, and $\xi_l = 0$ into Theorem B.7 and simplifying the resulting geometric series gives the result. □

Corollary B.9 (Ideal Edge-Wise Gating as a Special Case)**.** *Suppose there exist constants $g_+ > g_- > 0$, $\alpha_+ > 0$, $\alpha_- > 0$, $g_0 \geq 0$, and $\lambda_{\mathrm{leak}} \geq 0$ such that the effective retention inside the evidence region is $A = g_+ \alpha_+$, the self-propagation inside the noise region is $B = g_- \alpha_-$, the evidence-to-noise leakage is $C = g_0 \lambda_{\mathrm{leak}}$, and $\xi_l = 0$. Then Theorem B.7 reduces to*

$$\mathrm{SNR}_L \geq \left[\left(\frac{g_- \alpha_-}{g_+ \alpha_+} \right)^L \mathrm{SNR}_0^{-1} + \frac{g_0 \lambda_{\mathrm{leak}}}{g_+ \alpha_+} \sum_{i=0}^{L-1} \left(\frac{g_- \alpha_-}{g_+ \alpha_+} \right)^i \right]^{-1}. \tag{44}$$

If $\lambda_{\mathrm{leak}} = 0$, then

$$\mathrm{SNR}_L \geq \left(\frac{g_+ \alpha_+}{g_- \alpha_-} \right)^L \mathrm{SNR}_0 \,. \tag{45}$$

B.5 Document Retrieval Budget

The final retrieval targets of the SAGE-GFM reader are memory fragments or documents. Therefore, we need to convert the entity-level SNR bound into a document-level top-k budget bound. Let the final document score be $S_D(d)$, and let the top-k retrieval result be

$$P_k(q, G) = \text{Top-}k_{d \in D} S_D(d). \tag{46}$$

Definition B.10 (ρ-Coverage Retrieval Budget)**.** Given $0 < \rho \leq \rho_A$, let

$$m_\rho = \lceil \rho |D^+| \rceil. \tag{47}$$

The minimum top-k budget required to achieve ρ-level gold evidence coverage is defined as

$$\mathcal{B}_\rho(q, G) = \min \left\{ k : \; |P_k(q, G) \cap D^+| \geq m_\rho \right\}. \tag{48}$$

Let τ_ρ^+ denote the m_ρ-th largest score among gold evidence documents, namely the gold score threshold required to achieve ρ-coverage.

Lemma B.11 (Quantile Retrieval Budget Bound)**.** *Let the total score mass of distractor documents be*

$$M_L^- = \sum_{d \in D \setminus D^+} S_D(d). \tag{49}$$

If $\tau_\rho^+ > 0$*, then*

$$\mathcal{B}_\rho(q, G) \le m_\rho + \frac{M_L^-}{\tau_\rho^+}. \tag{50}$$

Proof. Define the set of distractor documents whose scores are not lower than τ_ρ^+ as

$$\mathcal{N}_\rho = \{d \in D \setminus D^+ : S_D(d) \ge \tau_\rho^+\}. \tag{51}$$

For any $d \in \mathcal{N}_\rho$, we have $S_D(d) \ge \tau_\rho^+$, and hence

$$|\mathcal{N}_\rho| \tau_\rho^+ \le \sum_{d \in \mathcal{N}_\rho} S_D(d) \le M_L^-. \tag{52}$$

Thus $|\mathcal{N}_\rho| \le M_L^- / \tau_\rho^+$. To ensure that the top-k results contain at least m_ρ gold evidence documents, it suffices to include these m_ρ gold documents and all distractors whose scores are not lower than the threshold τ_ρ^+. Therefore,

$$\mathcal{B}_\rho(q, G) \le m_\rho + |\mathcal{N}_\rho| \le m_\rho + \frac{M_L^-}{\tau_\rho^+}. \tag{53}$$

This proves the lemma. □

To use entity-level SNR for document-level retrieval, we need to control the noise expansion introduced by entity-to-document projection.

Assumption B.12 (Projection Noise and Gold Score Concentration)**.** There exist constants $K_A \ge 0$, $\zeta_A \ge 0$, and $c_\rho \in (0, 1]$ such that the final document scores satisfy

$$M_L^- \le K_A N_L + \zeta_A, \tag{54}$$

$$\tau_\rho^+ \ge \frac{c_\rho}{m_\rho} S_L. \tag{55}$$

Here, K_A is the noise expansion factor of entity-to-document projection, ζ_A denotes the projection residual caused by incorrect anchors, missing anchors, or additional text-similarity terms, and c_ρ measures whether the evidence signal is effectively distributed over at least m_ρ gold documents.

Theorem B.13 (Realistic Signal–Noise–Budget Bound)**.** *Under the conditions of Theorem B.7, further assume that Assumption B.12 holds. Then*

$$\mathcal{B}_\rho(q, G) \le m_\rho + \frac{m_\rho K_A}{c_\rho} \mathrm{SNR}_L^{-1} + \frac{m_\rho \zeta_A}{c_\rho S_L}. \tag{56}$$

Substituting Theorem B.7 further yields the explicit upper bound

$$\begin{aligned} \mathcal{B}_\rho(q, G) \le m_\rho + \frac{m_\rho K_A}{c_\rho} \Bigg[\left(\prod_{l=1}^{L} \frac{B_l}{A_l} \right) \mathrm{SNR}_0^{-1} \\ + \sum_{i=1}^{L} \left(\frac{C_i}{A_i} + \frac{\xi_i}{A_i S_{i-1}} \right) \prod_{t=i+1}^{L} \frac{B_t}{A_t} \Bigg] + \frac{m_\rho \zeta_A}{c_\rho S_L}. \end{aligned} \tag{57}$$

Proof. By Lemma B.11,

$$\mathcal{B}_\rho(q, G) \le m_\rho + \frac{M_L^-}{\tau_\rho^+}. \tag{58}$$

By Assumption B.12,

$$\frac{M_L^-}{\tau_\rho^+} \le \frac{K_A N_L + \zeta_A}{(c_\rho / m_\rho) S_L} = \frac{m_\rho K_A}{c_\rho} \frac{N_L}{S_L} + \frac{m_\rho \zeta_A}{c_\rho S_L}. \tag{59}$$

Since $N_L / S_L = \mathrm{SNR}_L^{-1}$, Eq. (56) follows. Substituting the upper bound on SNR_L^{-1} from Theorem B.7 into Eq. (56) gives Eq. (57). □

Corollary B.14 (Full Evidence Recovery Budget). *If $\rho = 1$ and $\rho_A = 1$, then $m_\rho = |D^+|$, and $\mathcal{B}_\rho(q, G)$ reduces to the full evidence recovery budget. In this case, Theorem B.13 provides an upper bound on the top-k budget required to recover all gold evidence.*

B.6 Interpretation of the Theoretical Bound for the SAGE Design

Theorem B.7 and Theorem B.13 unify four reader design factors under the same retrieval-budget upper bound. To avoid relying only on intuitive discussion, we provide several direct monotonicity propositions.

Proposition 2 (Monotonicity of the Budget Bound). Fix m_ρ, K_A, c_ρ, ζ_A, S_L, and SNR_0, and define

$$\Gamma_L = \left(\prod_{l=1}^{L} \frac{B_l}{A_l}\right) \mathrm{SNR}_0^{-1} + \sum_{i=1}^{L} \left(\frac{C_i}{A_i} + \frac{\xi_i}{A_i S_{i-1}}\right) \prod_{t=i+1}^{L} \frac{B_t}{A_t}. \tag{60}$$

Then the budget upper bound

$$U_\rho = m_\rho + \frac{m_\rho K_A}{c_\rho} \Gamma_L + \frac{m_\rho \zeta_A}{c_\rho S_L} \tag{61}$$

is monotonically nondecreasing in Γ_L, K_A, and ζ_A, and monotonically nonincreasing in c_ρ and S_L. If all other terms in the products are fixed, decreasing any of B_l/A_l, C_l/A_l, or $\xi_l/(A_l S_{l-1})$ cannot increase U_ρ.

Proof. The partial derivatives of U_ρ with respect to Γ_L, K_A, and ζ_A are nonnegative, while the partial derivatives with respect to c_ρ and S_L are nonpositive. Moreover, Γ_L is a nonnegative linear or multiplicative combination of B_l/A_l, C_l/A_l, and $\xi_l/(A_l S_{l-1})$. When the other terms are fixed, decreasing any such nonnegative term cannot increase Γ_L. The proposition follows. □

Proposition 3 (Effect of Soft Addressing). If soft addressing increases the initial evidence signal S_0 and decreases the initial noise mass N_0, thereby increasing SNR_0, then, with all other coefficients fixed, the budget upper bound in Theorem B.13 does not increase. In particular, explicit entities, aliases, pseudo-queries, type constraints, hard constraints, and entity-linking signals in query planning improve the final budget bound whenever they increase S_0/N_0 in an aggregate sense.

Proposition 4 (Aggregate Advantage of Structural Gating). Compared with a reader without structural gating, suppose the structurally gated reader satisfies

$$\frac{B_l^{\mathrm{gate}}}{A_l^{\mathrm{gate}}} \le \frac{B_l^{\mathrm{plain}}}{A_l^{\mathrm{plain}}}, \qquad \frac{C_l^{\mathrm{gate}}}{A_l^{\mathrm{gate}}} \le \frac{C_l^{\mathrm{plain}}}{A_l^{\mathrm{plain}}}, \qquad \frac{\xi_l^{\mathrm{gate}}}{A_l^{\mathrm{gate}} S_{l-1}^{\mathrm{gate}}} \le \frac{\xi_l^{\mathrm{plain}}}{A_l^{\mathrm{plain}} S_{l-1}^{\mathrm{plain}}}. \tag{62}$$

Then the budget upper bound of the structurally gated reader is no larger than that of the ungated reader.

Proposition 5 (Stability Interpretation of the Context–Schema Dual Channel). Let δ_l denote the failure probability of the aggregate recurrences in Eqs. (28)–(29) at layer l. If the schema prior channel reduces the variance of cross-graph structural-role estimation and the context calibration channel reduces the current-graph adaptation error, so that δ_l decreases to δ_l' with $\delta_l' \le \delta_l$, then the probability lower bound under which Theorem B.7 and Theorem B.13 simultaneously hold improves from $1 - \sum_{l=1}^{L} \delta_l$ to $1 - \sum_{l=1}^{L} \delta_l'$.

The core quantities derived above are

$$\mathrm{SNR}_L^{-1} \le \left(\prod_{l=1}^{L} \frac{B_l}{A_l}\right) \mathrm{SNR}_0^{-1} + \sum_{i=1}^{L} \left(\frac{C_i}{A_i} + \frac{\xi_i}{A_i S_{i-1}}\right) \prod_{t=i+1}^{L} \frac{B_t}{A_t}, \tag{63}$$

and

$$\mathcal{B}_\rho(q, G) \le m_\rho + \frac{m_\rho K_A}{c_\rho} \mathrm{SNR}_L^{-1} + \frac{m_\rho \zeta_A}{c_\rho S_L}. \tag{64}$$

Equation (63) shows that soft addressing reduces the amount of noise that subsequent propagation must overcome by improving the initial SNR_0; structural gating improves aggregate evidence retention and noise suppression by increasing A_l and decreasing B_l and C_l; the context–schema dual

channel makes these aggregate inequalities more stable on dynamic graph memories by reducing cross-graph structural-role estimation error; and entity-to-document projection converts entity-level SNR into document-level budget efficiency by decreasing K_A and ζ_A and increasing c_ρ. Equation (64) further shows that the advantage of SAGE-GFM does not rely on perfect edge-wise classification or zero-leakage assumptions. As long as evidence-retention dominance is achieved in an aggregate or high-probability sense, i.e., B_l/A_l and C_l/A_l are sufficiently small, the reader can improve query-relevant SNR and reduce the top-k retrieval budget required to achieve a given level of evidence coverage.

C Target Graph Calibration and Cross-graph Structural Priors

C.1 Structural Role Decomposition Assumption

Definition C.1 (Structural role mapping)**.** Given a graph G, a mapping

$$\rho_G : V(G) \to \mathcal{R} \tag{65}$$

is called a structural role mapping, where $\mathcal{R}$ is the structural role space. $\rho_G(v)$ can be jointly determined by $\varphi_G(v)$, the structural statistics of edges incident to v, local community-boundary statistics, and other graph-structure summaries. Typical structural roles include hub, bridge, community core, boundary node, and noisy shortcut.

Definition C.2 (Target graph reading risk)**.** Fix a target graph G. Let $\mathcal{D}_G$ be the query–node sampling distribution on the target graph, and let $f^\star_G(q, v)$ be the ideal evidence relevance function. For any measurable function f, define the squared risk as

$$\mathcal{R}_G(f) = \mathbb{E}_{(q,v)\sim\mathcal{D}_G}\left[\left(f(q, v, G) - f^\star_G(q, v)\right)^2\right]. \tag{66}$$

Assumption C.3 (Context–schema decomposability)**.** For every target graph G, the ideal reading function can be decomposed as

$$f^\star_G(q, v) = f^\star_{\mathrm{sch}}(q, \rho_G(v)) + f^\star_{\mathrm{ctx},G}(q, v), \tag{67}$$

where $f^\star_{\mathrm{sch}}$ denotes the cross-graph shared structural reading rule, and $f^\star_{\mathrm{ctx},G}$ denotes the target-graph residual induced by the current writer, current domain, current entity naming, local noise, and writing style.

Equation (67) corresponds exactly to the structural design of SAGE: H_{sch} is used to approximate $f^\star_{\mathrm{sch}}$, and H_{ctx} is used to approximate $f^\star_{\mathrm{ctx},G}$. The next subsection gives the risk meaning of this decomposition.

C.2 Approximation Risk of Context–schema Decomposition

Theorem C.4 (Context–schema decomposition reduces target-graph approximation risk)**.** *Suppose Assumption 2.3 holds. Let $\mathcal{H}_{\mathrm{sch}}$ be the schema function class, let $\mathcal{H}_{\mathrm{ctx},G}$ be the target-graph context function class, and define the sum class*

$$\mathcal{H}_{\mathrm{sch}} + \mathcal{H}_{\mathrm{ctx},G} = \{f_s + f_c : f_s \in \mathcal{H}_{\mathrm{sch}}, f_c \in \mathcal{H}_{\mathrm{ctx},G}\}. \tag{68}$$

If there exist $\epsilon_{\mathrm{sch}}, \epsilon_{\mathrm{ctx}} \geq 0$ such that

$$\inf_{f_s\in\mathcal{H}_{\mathrm{sch}}} \mathbb{E}\left[\left(f_s(q, \rho_G(v)) - f^\star_{\mathrm{sch}}(q, \rho_G(v))\right)^2\right] \leq \epsilon_{\mathrm{sch}}, \tag{69}$$

$$\inf_{f_c\in\mathcal{H}_{\mathrm{ctx},G}} \mathbb{E}\left[\left(f_c(q, v, G) - f^\star_{\mathrm{ctx},G}(q, v)\right)^2\right] \leq \epsilon_{\mathrm{ctx}}, \tag{70}$$

where both expectations are over $(q, v) \sim \mathcal{D}_G$, then

$$\inf_{f\in\mathcal{H}_{\mathrm{sch}}+\mathcal{H}_{\mathrm{ctx},G}} \mathcal{R}_G(f) \leq 2\epsilon_{\mathrm{sch}} + 2\epsilon_{\mathrm{ctx}}. \tag{71}$$

Proof. Take any $\alpha > 0$. By the two approximation error conditions, there exist $\hat{f}_s \in \mathcal{H}_{\mathrm{sch}}$ and $\hat{f}_c \in \mathcal{H}_{\mathrm{ctx},G}$ such that

$$\mathbb{E}[(\hat{f}_s - f^\star_{\mathrm{sch}})^2] \leq \epsilon_{\mathrm{sch}} + \alpha, \qquad \mathbb{E}[(\hat{f}_c - f^\star_{\mathrm{ctx},G})^2] \leq \epsilon_{\mathrm{ctx}} + \alpha. \tag{72}$$

Let $\hat{f} = \hat{f}_s + \hat{f}_c$. By the decomposition in Eq. (67), we have

$$\hat{f} - f_G^\star = (\hat{f}_s - f_{\mathrm{sch}}^\star) + (\hat{f}_c - f_{\mathrm{ctx},G}^\star). \tag{73}$$

Using $(a+b)^2 \le 2a^2 + 2b^2$, we obtain

$$\begin{aligned}\mathcal{R}_G(\hat{f}) &= \mathbb{E}[(\hat{f} - f_G^\star)^2] \\ &\le 2\mathbb{E}[(\hat{f}_s - f_{\mathrm{sch}}^\star)^2] + 2\mathbb{E}[(\hat{f}_c - f_{\mathrm{ctx},G}^\star)^2] \\ &\le 2\epsilon_{\mathrm{sch}} + 2\epsilon_{\mathrm{ctx}} + 4\alpha. \end{aligned} \tag{74}$$

Since $\alpha > 0$ is arbitrary, taking the infimum yields the conclusion. □

Proposition 6 (Residual bias of schema-only models). Further assume that $L_2(\mathcal{D}_G)$ is a Hilbert space, $\mathcal{H}_{\mathrm{sch}}$ is a closed linear subspace of it, and $f_{\mathrm{sch}}^\star \in \mathcal{H}_{\mathrm{sch}}$. If only a schema-only model $f_s \in \mathcal{H}_{\mathrm{sch}}$ is used, then

$$\inf_{f_s \in \mathcal{H}_{\mathrm{sch}}} \mathcal{R}_G(f_s) = \mathrm{dist}^2_{L_2(\mathcal{D}_G)}(f_{\mathrm{ctx},G}^\star, \mathcal{H}_{\mathrm{sch}}). \tag{75}$$

Therefore, as long as the target-graph residual $f_{\mathrm{ctx},G}^\star$ does not belong to $\mathcal{H}_{\mathrm{sch}}$, a schema-only reader has an irreducible target-graph bias.

Proof. By $f_{\mathrm{sch}}^\star \in \mathcal{H}_{\mathrm{sch}}$ and the linearity of $\mathcal{H}_{\mathrm{sch}}$, any $f_s \in \mathcal{H}_{\mathrm{sch}}$ can be written as $f_s = f_{\mathrm{sch}}^\star + g$, where $g \in \mathcal{H}_{\mathrm{sch}}$. Thus,

$$\begin{aligned}\mathcal{R}_G(f_s) &= \left\| f_s - f_{\mathrm{sch}}^\star - f_{\mathrm{ctx},G}^\star \right\|^2_{L_2(\mathcal{D}_G)} \\ &= \left\| g - f_{\mathrm{ctx},G}^\star \right\|^2_{L_2(\mathcal{D}_G)}. \end{aligned} \tag{76}$$

Taking the infimum over $f_s \in \mathcal{H}_{\mathrm{sch}}$ is equivalent to taking the infimum over $g \in \mathcal{H}_{\mathrm{sch}}$, and Eq. (75) follows from the definition of distance. □

Remark C.5. Proposition 2.5 shows that the cross-graph schema prior can only characterize cross-graph shared structural rules and cannot replace target graph calibration. In contrast, the role of the target graph calibration channel H_{ctx} is to absorb $f_{\mathrm{ctx},G}^\star$, namely the local noise, entity granularity, relation style, and domain residual of the graph generated by the current writer.

C.3 Sample Complexity Advantage of Schema Prior

Lemma C.6 (Uniform convergence for bounded loss classes)**.** *Let $\mathcal{F}$ be a function class, and let the loss $\ell_f(z) \in [0,1]$. Given n independent samples $S = \{z_i\}_{i=1}^n$, define the true risk $R(f) = \mathbb{E}[\ell_f(z)]$ and the empirical risk $\widehat{R}_S(f) = n^{-1}\sum_{i=1}^n \ell_f(z_i)$. Then, with probability at least $1-\delta$, for all $f \in \mathcal{F}$ simultaneously,*

$$R(f) \le \widehat{R}_S(f) + 2\,\mathrm{Rad}_n(\ell \circ \mathcal{F}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}, \tag{77}$$

where $\mathrm{Rad}_n(\ell \circ \mathcal{F})$ is an upper bound on the empirical Rademacher complexity of the loss-composed class.

Proof. This is a direct result of the standard symmetrization and McDiarmid/Hoeffding concentration inequalities for bounded loss function classes. Specifically, let

$$Z(S) = \sup_{f \in \mathcal{F}} \left| R(f) - \widehat{R}_S(f) \right|. \tag{78}$$

By symmetrization, $\mathbb{E}Z(S) \le 2\,\mathrm{Rad}_n(\ell \circ \mathcal{F})$. Moreover, since changing one sample can change $Z(S)$ by at most $1/n$, McDiarmid's inequality gives

$$Z(S) \le \mathbb{E}Z(S) + 3\sqrt{\frac{\log(2/\delta)}{2n}} \tag{79}$$

with probability at least $1-\delta$. Combining the two inequalities gives the conclusion. □

Theorem C.7 (Schema prior reduces the sample complexity of target-graph adaptation)**.** *Fix a target graph G, and suppose the number of supervised samples available for reader calibration on the target graph is n_G. Let $\mathcal{H}_{\text{full}}$ be the full reader class that needs to be learned on the target graph when no schema prior is used; let $\mathcal{H}_{\text{res}}$ be the residual class that only needs to be learned given a schema prior f_s, with the combined model $f = f_s + f_r$, $f_r \in \mathcal{H}_{\text{res}}$. Let $\hat{f}_{\text{full}}$ and $\hat{f}_{\text{res}}$ be the empirical risk minimizers over the two classes, respectively. Then, with probability at least $1-\delta$,*

$$\mathcal{R}_G(\hat{f}_{\text{full}}) \le \inf_{f \in \mathcal{H}_{\text{full}}} \mathcal{R}_G(f) + 4\,\mathrm{Rad}_{n_G}(\ell \circ \mathcal{H}_{\text{full}}) + 6\sqrt{\frac{\log(4/\delta)}{2n_G}}, \tag{80}$$

$$\mathcal{R}_G(f_s + \hat{f}_{\text{res}}) \le \inf_{f_r \in \mathcal{H}_{\text{res}}} \mathcal{R}_G(f_s + f_r) + 4\,\mathrm{Rad}_{n_G}(\ell \circ (f_s + \mathcal{H}_{\text{res}})) + 6\sqrt{\frac{\log(4/\delta)}{2n_G}}. \tag{81}$$

If the complexities satisfy

$$\mathrm{Rad}_{n_G}(\ell \circ \mathcal{H}_{\text{full}}) = \widetilde{O}\left(\sqrt{\frac{d_{\text{full}}}{n_G}}\right), \qquad \mathrm{Rad}_{n_G}(\ell \circ (f_s + \mathcal{H}_{\text{res}})) = \widetilde{O}\left(\sqrt{\frac{d_{\text{res}}}{n_G}}\right), \tag{82}$$

and $d_{\text{res}} \ll d_{\text{full}}$, then the schema prior reduces target-graph learning from estimating the full reading function to estimating the residual, and lowers the estimation error term for target-graph adaptation.

Proof. For any function class $\mathcal{F}$, let $\hat{f}$ be the empirical risk minimizer and let $f^\circ \in \arg\inf_{f\in\mathcal{F}} R(f)$ be the true risk minimizer. By empirical optimality, $\widehat{R}(\hat{f}) \le \widehat{R}(f^\circ)$, so

$$\begin{aligned} R(\hat{f}) - R(f^\circ) &= R(\hat{f}) - \widehat{R}(\hat{f}) + \widehat{R}(\hat{f}) - \widehat{R}(f^\circ) + \widehat{R}(f^\circ) - R(f^\circ) \\ &\le 2 \sup_{f\in\mathcal{F}} \left| R(f) - \widehat{R}(f) \right|. \end{aligned} \tag{83}$$

Applying Lemma 2.7 to $\mathcal{F} = \mathcal{H}_{\text{full}}$ and $\mathcal{F} = f_s + \mathcal{H}_{\text{res}}$, respectively, and combining the two probability events by a union bound, gives Eq. (80) and Eq. (81). The complexity-order conclusion follows by substituting the corresponding Rademacher upper bounds. □

Remark C.8. Theorem 2.8 gives the key theoretical motivation for the schema prior: in self-evolving memory, each new graph generated by the writer usually provides only limited supervised signal. If the reader relearns the full reading function from the target graph in every round, high-variance estimation arises; the schema prior fixes or strongly regularizes the cross-graph shared structural rules, so that target-graph calibration only needs to learn the residual, thereby reducing sample complexity.

D Writer-induced Graph Distribution Shift and Target Graph Calibration

D.1 Writer-induced Dynamic Graph Distribution

The writer parameter θ induces a graph distribution $P_\theta(G \mid q, D)$ on the sample (q, D). For notational simplicity, denote the joint distribution of (q, D, D^+, y, G) by Π_θ. Given the reader parameter ϕ, define the reader risk as

$$\mathcal{L}_R(\phi;\theta) = \mathbb{E}_{(q,D,D^+,y,G)\sim\Pi_\theta}\left[\ell_R(R_\phi(q,G,D), D^+, y)\right]. \tag{84}$$

Here, ℓ_R can be supporting-entity BCE, multi-positive ranking loss, document recall loss, or a combination thereof.

Proposition 7 (Writer updates cause reader distribution shift). Assume $0 \le \ell_R \le 1$. For any fixed ϕ and any writer parameters θ, θ', we have

$$|\mathcal{L}_R(\phi;\theta') - \mathcal{L}_R(\phi;\theta)| \le \mathrm{TV}(\Pi_{\theta'}, \Pi_\theta), \tag{85}$$

where TV is the total variation distance. If, further, $\ell_R(R_\phi(q,G,D), D^+, y)$ is L_ℓ-Lipschitz with respect to the graph variable under some graph metric $d_{\mathcal{G}}$, then

$$|\mathcal{L}_R(\phi;\theta') - \mathcal{L}_R(\phi;\theta)| \le L_\ell W_1(\Pi_{\theta'}, \Pi_\theta), \tag{86}$$

where W_1 is the first-order Wasserstein distance induced by $d_{\mathcal{G}}$.

Proof. For the bounded loss case, let $h_\phi(q, D, D^+, y, G) = \ell_R(R_\phi(q, G, D), D^+, y) \in [0, 1]$. Then

$$|\mathcal{L}_R(\phi; \theta') - \mathcal{L}_R(\phi; \theta)| = \left| \int h_\phi \, d\Pi_{\theta'} - \int h_\phi \, d\Pi_\theta \right| \leq \mathrm{TV}(\Pi_{\theta'}, \Pi_\theta), \tag{87}$$

where the last step follows from the dual definition of total variation. If h_ϕ is L_ℓ-Lipschitz, the Wasserstein upper bound follows from the Kantorovich–Rubinstein duality. □

Corollary D.1 (Necessity of target graph calibration)**.** *If updating the writer from θ to θ' causes $\mathrm{TV}(\Pi_{\theta'}, \Pi_\theta)$ to be non-negligible, then the risk of a fixed reader ϕ on the new graph distribution may increase. Therefore, target graph calibration of the reader, namely $\phi \mapsto \phi'$ to reduce $\mathcal{L}_R(\phi'; \theta')$, is a necessary mechanism for handling writer-induced graph distribution shift.*

Proof. By Proposition 3.1, writer distribution shift can directly change the new-distribution risk of the fixed reader. If the reader is not updated, there is no optimization mechanism to offset this drift term. Target graph calibration is precisely re-optimization of $\mathcal{L}_R(\cdot; \theta')$, and is therefore a natural step for reducing the risk on the new graph. □

E Reader Stability under Dynamic Graph Evolution

E.1 Realistic Graph Evolution Distance

Real graphs often contain hubs, node additions and deletions, alias merges, anchor rewrites, and structural statistics that are not globally Lipschitz. Therefore, we use the augmented graph drift actually perceived by the reader to measure graph evolution.

Definition E.1 (Padding alignment and presence bit)**.** Given two consecutive-round graphs G and G', align them through persistent memory ids to a common node universe $\bar{V} = V(G) \cup V(G')$. If a node exists only in one graph, it is treated as an isolated padding node in the other graph, and a presence bit is added to its features. The aligned node feature matrices are still denoted by X, X'.

Definition E.2 (Augmented graph drift)**.** Let A, A' be self-looped row-normalized adjacency matrices, let S_q, S'_q be the entry score vectors before soft addressing, and let B, B' be row-normalized entity-to-document anchoring matrices. Define

$$\Delta_X = \|X - X'\|_{2,\infty}, \qquad \Delta_A = \|A - A'\|_\infty, \qquad \Delta_{\mathrm{seed}} = \left\|S_q - S'_q\right\|_\infty, \qquad \Delta_B = \|B - B'\|_\infty, \tag{88}$$

where

$$\|H\|_{2,\infty} = \max_{v \in \bar{V}} \|h_v\|_2, \qquad \|A\|_\infty = \max_v \sum_u |A_{vu}|. \tag{89}$$

For the gate input $z_{uv}^{(l)}$ at layer l, define the weighted structural drift as

$$\Delta_Z^{(l)} = \max_v \sum_u A'_{vu} \left\| z_{uv}^{(l)} - z_{uv}^{\prime(l)} \right\|_2, \qquad \Delta_Z = \max_{1 \leq l \leq L} \Delta_Z^{(l)}. \tag{90}$$

The total augmented graph drift is defined as

$$\Delta_{\mathrm{aug}}(G, G'; q) = \Delta_X + \Delta_A + \Delta_{\mathrm{seed}} + \Delta_Z + \Delta_B. \tag{91}$$

E.2 Stability Assumptions

Assumption E.3 (Normalized adjacency)**.** For all considered graphs, $A_{vu} \geq 0$ and

$$\sum_u A_{vu} = 1, \qquad \forall v \in \bar{V}. \tag{92}$$

Therefore, $\|A\|_\infty = 1$. This assumption allows high-degree hubs to exist, but prevents single-layer propagation from being unboundedly amplified by node degree.

Assumption E.4 (Trajectory-local boundedness)**.** For graph pairs G, G' on the training and inference trajectories, there exist constants B_l such that

$$\left\| H^{(l)}(q, G) \right\|_{2,\infty} \leq B_l, \qquad \left\| H^{(l)}(q, G') \right\|_{2,\infty} \leq B_l, \qquad l = 0, \ldots, L. \tag{93}$$

Assumption E.5 (Locally Lipschitz modules)**.** In a neighborhood of the training trajectory, the l-th layer satisfies

$$\left\|W_m^{(l)}\right\|_2 \le M_l, \tag{94}$$

$$\left\|\mathrm{MLP}_g^{(l)}(z) - \mathrm{MLP}_g^{(l)}(z')\right\|_\infty \le L_{g,l} \left\|z - z'\right\|_2 . \tag{95}$$

The Lipschitz constant of PReLU is L_σ, and the Lipschitz constant of LayerNorm with a numerical stabilizer in this trajectory neighborhood is $L_{\mathrm{LN},l}$.

Assumption E.6 (Local Lipschitzness of score head and projection)**.** The entity score head satisfies

$$\left\|s_E(q,G) - s_E(q,G')\right\|_\infty \le L_E \left\|H(q,G) - H(q,G')\right\|_{2,\infty} . \tag{96}$$

Meanwhile, $\left\|s_E(q,G)\right\|_\infty \le S_E$, and B is row-normalized, so $\left\|B\right\|_\infty \le 1$.

E.3 Stability of Soft Addressing and Initial Representation

Lemma E.7 (Softmax and pre-activation stability)**.** *Let*

$$p = \mathrm{softmax}(S/T_0), \qquad p' = \mathrm{softmax}(S'/T_0). \tag{97}$$

Then

$$\left\|p - p'\right\|_\infty \le \frac{1}{T_0} \left\|S - S'\right\|_\infty . \tag{98}$$

Furthermore, let $a_v = (p_v + \epsilon_p)^\eta$ *and* $a'_v = (p'_v + \epsilon_p)^\eta$*, where* $0 < \eta \le 1$*. Then*

$$\left\|a - a'\right\|_\infty \le \frac{\eta \epsilon_p^{\eta-1}}{T_0} \left\|S - S'\right\|_\infty . \tag{99}$$

Proof. The Jacobian of softmax is $J(z) = \mathrm{diag}(p) - pp^\top$. For any row i,

$$\sum_j |J_{ij}(z)| = 2p_i(1 - p_i) \le 1. \tag{100}$$

Thus, $\left\|J(z)\right\|_{\infty\to\infty} \le 1$. By the mean value theorem and setting $z = S/T_0$, we obtain

$$\left\|p - p'\right\|_\infty \le \frac{1}{T_0} \left\|S - S'\right\|_\infty . \tag{101}$$

The function $t \mapsto (t + \epsilon_p)^\eta$ is Lipschitz on $[0,1]$, with constant at most $\eta\epsilon_p^{\eta-1}$. Composing the two inequalities gives the conclusion. □

Lemma E.8 (Initial node representation stability)**.** *Let* $u_q = W_q \,\mathrm{Emb}(q)$*, and define*

$$h_v^{(0)} = a_v(q) u_q + W_x x_v. \tag{102}$$

Then

$$\left\|H^{(0)}(q,G) - H^{(0)}(q,G')\right\|_{2,\infty} \le C_{\mathrm{init}}(\Delta_{\mathrm{seed}} + \Delta_X), \tag{103}$$

where

$$C_{\mathrm{init}} = \frac{\eta \epsilon_p^{\eta-1}}{T_0} \left\|u_q\right\|_2 + \left\|W_x\right\|_2 . \tag{104}$$

Proof. For any node v,

$$h_v^{(0)} - h_v'^{(0)} = (a_v - a'_v) u_q + W_x(x_v - x'_v). \tag{105}$$

Taking the norm and applying Lemma 4.8 gives

$$\left\|h_v^{(0)} - h_v'^{(0)}\right\|_2 \le \frac{\eta \epsilon_p^{\eta-1}}{T_0} \left\|u_q\right\|_2 \Delta_{\mathrm{seed}} + \left\|W_x\right\|_2 \Delta_X. \tag{106}$$

Taking the maximum over v gives the conclusion. □

E.4 Single-layer Stability of Structurally Gated Propagation

Lemma E.9 (Boundedness and stability of structural gate)**.** *The structural gate of layer l,*

$$g_{uv}^{(l)} = 1 + \delta \tanh(\mathrm{MLP}_g^{(l)}(z_{uv}^{(l)})), \tag{107}$$

satisfies

$$\left\| g_{uv}^{(l)} \right\|_\infty \le 1 + \delta, \tag{108}$$

and

$$\left\| g_{uv}^{(l)} - g_{uv}^{\prime(l)} \right\|_\infty \le \delta L_{g,l} \left\| z_{uv}^{(l)} - z_{uv}^{\prime(l)} \right\|_2. \tag{109}$$

Proof. Since the range of $\tanh$ is contained in $[-1, 1]$, the first statement follows immediately. Moreover, because $\tanh$ is 1-Lipschitz and $\mathrm{MLP}_g^{(l)}$ is $L_{g,l}$-Lipschitz in the trajectory neighborhood,

$$\begin{aligned} \left\| g_{uv}^{(l)} - g_{uv}^{\prime(l)} \right\|_\infty &\le \delta \left\| \mathrm{MLP}_g^{(l)}(z_{uv}^{(l)}) - \mathrm{MLP}_g^{(l)}(z_{uv}^{\prime(l)}) \right\|_\infty \\ &\le \delta L_{g,l} \left\| z_{uv}^{(l)} - z_{uv}^{\prime(l)} \right\|_2. \end{aligned} \tag{110}$$

□

Lemma E.10 (Single-layer stability of structurally gated propagation)**.** *Define*

$$D_l = \left\| H^{(l)}(q, G) - H^{(l)}(q, G') \right\|_{2,\infty}. \tag{111}$$

Under Assumptions 4.5–4.7, the l-th propagation layer satisfies

$$D_l \le \alpha_l D_{l-1} + \beta_l^A \Delta_A + \beta_l^Z \Delta_Z^{(l)}, \tag{112}$$

where one can take

$$\alpha_l = L_{\mathrm{LN},l}\big(1 + L_\sigma(1+\delta)M_l\big), \tag{113}$$

$$\beta_l^A = L_{\mathrm{LN},l} L_\sigma (1+\delta) M_l B_{l-1}, \qquad \beta_l^Z = L_{\mathrm{LN},l} L_\sigma \delta L_{g,l} M_l B_{l-1}. \tag{114}$$

Proof. Write

$$M_v = \sum_u A_{vu} g_{uv} \odot W h_u, \qquad M_v' = \sum_u A'_{vu} g'_{uv} \odot W h'_u, \tag{115}$$

where the layer index l is omitted. Adding and subtracting intermediate terms gives

$$\begin{aligned} M_v - M_v' = &\sum_u A_{vu} g_{uv} \odot W(h_u - h'_u) \\ &+ \sum_u (A_{vu} - A'_{vu}) g_{uv} \odot W h'_u \\ &+ \sum_u A'_{vu} (g_{uv} - g'_{uv}) \odot W h'_u. \end{aligned} \tag{116}$$

The first term is controlled by row-normalization, $\|g_{uv}\|_\infty \le 1 + \delta$, and $\|W\|_2 \le M_l$:

$$\left\| \sum_u A_{vu} g_{uv} \odot W(h_u - h'_u) \right\|_2 \le (1+\delta) M_l D_{l-1}. \tag{117}$$

The second term satisfies

$$\left\| \sum_u (A_{vu} - A'_{vu}) g_{uv} \odot W h'_u \right\|_2 \le (1+\delta) M_l B_{l-1} \sum_u |A_{vu} - A'_{vu}| \le (1+\delta) M_l B_{l-1} \Delta_A. \tag{118}$$

For the third term, by Lemma 4.10,

$$\begin{aligned} \left\| \sum_u A'_{vu} (g_{uv} - g'_{uv}) \odot W h'_u \right\|_2 &\le \delta L_{g,l} M_l B_{l-1} \sum_u A'_{vu} \left\| z_{uv}^{(l)} - z_{uv}^{\prime(l)} \right\|_2 \\ &\le \delta L_{g,l} M_l B_{l-1} \Delta_Z^{(l)}. \end{aligned} \tag{119}$$

Therefore,

$$\|M_v - M'_v\|_2 \le (1+\delta)M_l D_{l-1} + (1+\delta)M_l B_{l-1}\Delta_A + \delta L_{g,l} M_l B_{l-1}\Delta_Z^{(l)}. \tag{120}$$

By the L_σ-Lipschitz property of PReLU, the residual structure, and the $L_{\mathrm{LN},l}$-Lipschitz property of LayerNorm,

$$\left\|h_v^{(l)} - h_v'^{(l)}\right\|_2 \le L_{\mathrm{LN},l}\left(\left\|h_v^{(l-1)} - h_v'^{(l-1)}\right\|_2 + L_\sigma \|M_v - M'_v\|_2\right). \tag{121}$$

Taking the maximum over v gives Eq. (112). □

E.5 Stability of Representations, Scores, and Retrieval Sets

Theorem E.11 (Local stability of structurally gated representations to augmented graph drift)**.** *Under Assumptions 4.5–4.7, L-layer structurally gated propagation satisfies*

$$D_L \le \left(\prod_{l=1}^{L}\alpha_l\right) D_0 + \sum_{t=1}^{L}\left(\prod_{l=t+1}^{L}\alpha_l\right)\left(\beta_t^A\Delta_A + \beta_t^Z\Delta_Z^{(t)}\right). \tag{122}$$

Therefore, there exists a constant $C_H > 0$ such that

$$\left\|H^{(L)}(q,G) - H^{(L)}(q,G')\right\|_{2,\infty} \le C_H(\Delta_X + \Delta_{\mathrm{seed}} + \Delta_A + \Delta_Z). \tag{123}$$

Proof. By Lemma 4.11, the recursion in Eq. (112) holds. Unrolling the recursion layer by layer gives Eq. (122). By Lemma 4.9,

$$D_0 \le C_{\mathrm{init}}(\Delta_X + \Delta_{\mathrm{seed}}). \tag{124}$$

Substituting $\Delta_Z^{(t)} \le \Delta_Z$ and merging all layer-related constants into C_H gives the conclusion. □

Theorem E.12 (Stability of context/schema dual channels)**.** *If the two channels respectively satisfy*

$$\|H_{\mathrm{ctx}}(q,G) - H_{\mathrm{ctx}}(q,G')\|_{2,\infty} \le C_{\mathrm{ctx}}\Delta_{\mathrm{aug}}, \tag{125}$$

$$\|H_{\mathrm{sch}}(q,G) - H_{\mathrm{sch}}(q,G')\|_{2,\infty} \le C_{\mathrm{sch}}\Delta_{\mathrm{aug}}, \tag{126}$$

then the additive fusion satisfies

$$\|H(q,G) - H(q,G')\|_{2,\infty} \le (C_{\mathrm{ctx}} + |\beta_{\mathrm{sch}}|\, C_{\mathrm{sch}})\Delta_{\mathrm{aug}}. \tag{127}$$

If a normalized or gated convex fusion theoretical form is adopted,

$$H_\lambda(q,G) = (1-\lambda)H_{\mathrm{ctx}}(q,G) + \lambda H_{\mathrm{sch}}(q,G), \qquad 0 \le \lambda \le 1, \tag{128}$$

then

$$\|H_\lambda(q,G) - H_\lambda(q,G')\|_{2,\infty} \le \big((1-\lambda)C_{\mathrm{ctx}} + \lambda C_{\mathrm{sch}}\big)\Delta_{\mathrm{aug}}. \tag{129}$$

In particular, if $C_{\mathrm{sch}} < C_{\mathrm{ctx}}$, increasing λ decreases this worst-case stability upper bound.

Proof. The additive fusion case follows directly from the triangle inequality:

$$\begin{aligned}\|H(G) - H(G')\|_{2,\infty} &\le \|H_{\mathrm{ctx}}(G) - H_{\mathrm{ctx}}(G')\|_{2,\infty} + |\beta_{\mathrm{sch}}|\,\|H_{\mathrm{sch}}(G) - H_{\mathrm{sch}}(G')\|_{2,\infty} \\ &\le (C_{\mathrm{ctx}} + |\beta_{\mathrm{sch}}|\, C_{\mathrm{sch}})\Delta_{\mathrm{aug}}.\end{aligned} \tag{130}$$

The convex fusion case is analogous:

$$\|H_\lambda(G) - H_\lambda(G')\|_{2,\infty} \le (1-\lambda)C_{\mathrm{ctx}}\Delta_{\mathrm{aug}} + \lambda C_{\mathrm{sch}}\Delta_{\mathrm{aug}}. \tag{131}$$

If $C_{\mathrm{sch}} < C_{\mathrm{ctx}}$, the right-hand side is monotonically decreasing in λ. □

Remark E.13. Equation (127) does not claim that the schema prior necessarily reduces the worst-case Lipschitz constant of additive fusion; its main theoretical role also lies in risk decomposition and sample complexity reduction. If additional gating, normalization, or regularization constraints are adopted in the implementation, then Eq. (129) shows that the schema channel can also serve as a low-sensitivity channel to reduce graph evolution drift.

Theorem E.14 (Stability of entity scores and document scores)**.** *Under Assumption 4.7, there exist constants $C_E, C_D > 0$ such that*

$$\|s_E(q,G) - s_E(q,G')\|_\infty \le C_E \Delta_{\mathrm{aug}}, \tag{132}$$

$$\|s_D(q,G) - s_D(q,G')\|_\infty \le C_D \Delta_{\mathrm{aug}}. \tag{133}$$

For additive fusion, one can take

$$C_E = L_E(C_{\mathrm{ctx}} + |\beta_{\mathrm{sch}}|\, C_{\mathrm{sch}}), \qquad C_D = C_E + S_E. \tag{134}$$

Proof. The entity score upper bound follows from the Lipschitz property of the score head and Theorem 4.13. For document projection, $s_D = B s_E$, and therefore

$$s_D(G) - s_D(G') = B_G(s_E(G) - s_E(G')) + (B_G - B_{G'}) s_E(G'). \tag{135}$$

Taking the ℓ_∞ norm and using $\|B_G\|_\infty \le 1$ gives

$$\|s_D(G) - s_D(G')\|_\infty \le \|s_E(G) - s_E(G')\|_\infty + \|B_G - B_{G'}\|_\infty \|s_E(G')\|_\infty . \tag{136}$$

Using $\|s_E(G')\|_\infty \le S_E$ and $\Delta_B \le \Delta_{\mathrm{aug}}$ gives the document score stability. □

Theorem E.15 (Boundary stability of hard top-k)**.** *Let $s = s_D(q,G)$ and $s' = s_D(q,G')$, and suppose*

$$\|s - s'\|_\infty \le \epsilon_s. \tag{137}$$

Let $t_k = s_{(k)}$ be the k-th largest score in s, and define the boundary set

$$\mathcal{B}_{k,2\epsilon_s}(s) = \{d : |s_d - t_k| \le 2\epsilon_s\}. \tag{138}$$

Then

$$\text{Top-}k(s) \triangle \text{Top-}k(s') \subseteq \mathcal{B}_{k,2\epsilon_s}(s), \tag{139}$$

and hence

$$|\text{Top-}k(s) \triangle \text{Top-}k(s')| \le |\mathcal{B}_{k,2\epsilon_s}(s)|\,. \tag{140}$$

In particular, if $s_{(k)} - s_{(k+1)} > 2\epsilon_s$, then Top-$k(s) =$ Top-$k(s')$.

Proof. Take any $i \in \text{Top-}k(s) \setminus \text{Top-}k(s')$. Since i drops out of the top-k, there exists $j \notin \text{Top-}k(s)$ such that $j \in \text{Top-}k(s')$ and $s'_j \ge s'_i$. By the perturbation bound,

$$s_j + \epsilon_s \ge s'_j \ge s'_i \ge s_i - \epsilon_s, \tag{141}$$

so $s_i \le s_j + 2\epsilon_s \le t_k + 2\epsilon_s$. Since $i \in \text{Top-}k(s)$, we have $s_i \ge t_k$, and hence $i \in \mathcal{B}_{k,2\epsilon_s}(s)$. A symmetric argument for $j \in \text{Top-}k(s') \setminus \text{Top-}k(s)$ gives $j \in \mathcal{B}_{k,2\epsilon_s}(s)$. Therefore, the symmetric difference is contained in the boundary set. If $s_{(k)} - s_{(k+1)} > 2\epsilon_s$, boundary exchange cannot occur, and the top-k set remains unchanged. □

Corollary E.16 (Top-k boundary stability under graph evolution)**.** *By Theorem 4.15, taking $\epsilon_s = C_D \Delta_{\mathrm{aug}}(G, G'; q)$ yields*

$$P_k(q,G) \triangle P_k(q,G') \subseteq \mathcal{B}_{k,2C_D\Delta_{\mathrm{aug}}}(s_D(q,G)). \tag{142}$$

Therefore, the instability of hard top-k is restricted to candidates near the original score boundary.

Theorem E.17 (Stability of soft retrieval distribution)**.** *Let*

$$\pi_D(q,G) = \mathrm{softmax}(s_D(q,G)/\tau). \tag{143}$$

If $\|s_D(q,G) - s_D(q,G')\|_\infty \le \epsilon_s$, then

$$\|\pi_D(q,G) - \pi_D(q,G')\|_1 \le \frac{2}{\tau}\epsilon_s. \tag{144}$$

Therefore,

$$\|\pi_D(q,G) - \pi_D(q,G')\|_1 \le \frac{2C_D}{\tau}\Delta_{\mathrm{aug}}(G, G'; q). \tag{145}$$

Proof. The Jacobian of softmax is $J(z) = \mathrm{diag}(\pi) - \pi\pi^\top$. For any perturbation r,

$$J(z)r = \pi \odot (r - \mathbb{E}_\pi r). \tag{146}$$

If $\|r\|_\infty \le 1$, then $|r_i - \mathbb{E}_\pi r| \le 2$, so

$$\|J(z)r\|_1 \le \sum_i \pi_i \, |r_i - \mathbb{E}_\pi r| \le 2. \tag{147}$$

Thus, the $\ell_\infty \to \ell_1$ Lipschitz constant of softmax is at most 2. Since the input is s_D/τ, we obtain

$$\|\pi_D(s) - \pi_D(s')\|_1 \le \frac{2}{\tau} \|s - s'\|_\infty . \tag{148}$$

Substituting Theorem 4.15 gives the conclusion. □

Theorem E.18 (High-probability graph evolution stability)**.** *If the writer's single-round graph update satisfies*

$$\mathbb{P}[\Delta_{\mathrm{aug}}(G, G'; q) > \epsilon] \le \delta, \tag{149}$$

then

$$\mathbb{P}\left[\|s_D(q, G) - s_D(q, G')\|_\infty > C_D\epsilon\right] \le \delta. \tag{150}$$

If $\mathbb{E}[\Delta_{\mathrm{aug}}(G, G'; q)] \le \bar{\epsilon}$*, then*

$$\mathbb{E}\left[\|s_D(q, G) - s_D(q, G')\|_\infty\right] \le C_D\bar{\epsilon}. \tag{151}$$

Proof. By Theorem 4.15, for any graph pair, we have

$$\|s_D(q, G) - s_D(q, G')\|_\infty \le C_D \Delta_{\mathrm{aug}}(G, G'; q). \tag{152}$$

Therefore, the event $\{\|s_D(q, G) - s_D(q, G')\|_\infty > C_D\epsilon\}$ implies the event $\{\Delta_{\mathrm{aug}}(G, G'; q) > \epsilon\}$, so the probability upper bound follows immediately. The expectation conclusion follows by taking expectations on both sides of the deterministic inequality. □

E.6 Local Influence Cone

Proposition 8 (Influence cone of local graph updates). Suppose the writer only changes nodes, edges, anchors, or attributes on a primitive set $\mathcal{U}$. Suppose that the structural gate input $z_{uv}^{(l)}$, except for the graph-level summary, only depends on a local neighborhood of radius r_z, and that G and G' are exactly the same outside $\mathcal{N}_{L+r_z}(\mathcal{U})$. If graph-level summary drift is ignored, then for any

$$v \notin \mathcal{N}_{L+r_z}(\mathcal{U}), \tag{153}$$

we have

$$h_v^{(L)}(q, G) = h_v^{(L)}(q, G'). \tag{154}$$

If the graph-level summary drift is $\rho_g = \|r_G - r_{G'}\|_2$, then there exists a constant C_g such that

$$\left\| h_v^{(L)}(q, G) - h_v^{(L)}(q, G') \right\|_2 \le C_g \rho_g. \tag{155}$$

Proof. First consider the case without graph-level summary drift. We induct on the layer index l. For $l = 0$, if $v \notin \mathcal{N}_{L+r_z}(\mathcal{U})$, then its node features, presence bit, and seed score are all identical, so $h_v^{(0)}(G) = h_v^{(0)}(G')$. Suppose that at layer $l-1$, all nodes whose distance from $\mathcal{U}$ exceeds $L + r_z - (l-1)$ have identical representations. If $v \notin \mathcal{N}_{L+r_z-l}(\mathcal{U})$, then all its one-hop neighbors u do not belong to $\mathcal{N}_{L+r_z-(l-1)}(\mathcal{U})$; by the induction hypothesis, $h_u^{(l-1)}(G) = h_u^{(l-1)}(G')$. Meanwhile, the radius-$r_z$ local structural contexts of all relevant edges are also identical, so the gate, message multiset, and aggregation result are identical, and hence $h_v^{(l)}(G) = h_v^{(l)}(G')$. Taking $l = L$ gives the first conclusion. If graph-level summary drift exists, then the gate input has an additional uniform perturbation term ρ_g, and a $C_g\rho_g$-type upper bound follows from Lemma 4.10 and the recursion in Theorem 4.12. □

F Theoretical Motivation of the Self-evolving Writer-Reader Loop

F.1 Joint Memory Utility

The reader-aware writer reward can consist of evidence coverage, precision, deducibility, and answer utility. Abstractly, define the joint memory utility as

$$\mathcal{J}(\theta,\phi)=\mathbb{E}_{(q,D,D^+,y)}\left[U\left(R_\phi(q,W_\theta(q,D),D),D^+,y\right)\right], \tag{156}$$

where U can be taken as

$$U=\frac{\alpha r_{\mathrm{rec}}+\beta r_{\mathrm{pre}}+\gamma r_{\mathrm{ded}}}{\alpha+\beta+\gamma}-\lambda_{\mathrm{rep}}\rho_{\mathrm{rep}}+\lambda_{\mathrm{fmt}}r_{\mathrm{fmt}}, \tag{157}$$

or an extended form including answer-level reward. This definition places the writer's graph construction quality and the reader's graph reading ability under the same objective.

F.2 Approximate Coordinate Improvement

Theorem F.1 (The self-evolution process is approximate coordinate improvement on joint utility)**.** *Suppose that the writer update at round r satisfies*

$$\mathcal{J}(\theta^{(r+1)},\phi^{(r)})\geq\mathcal{J}(\theta^{(r)},\phi^{(r)})+\Delta_W^{(r)}-\epsilon_W^{(r)}, \tag{158}$$

and the reader update satisfies

$$\mathcal{J}(\theta^{(r+1)},\phi^{(r+1)})\geq\mathcal{J}(\theta^{(r+1)},\phi^{(r)})+\Delta_R^{(r)}-\epsilon_R^{(r)}. \tag{159}$$

Then one full round of writer-reader self-evolution satisfies

$$\mathcal{J}(\theta^{(r+1)},\phi^{(r+1)})-\mathcal{J}(\theta^{(r)},\phi^{(r)})\geq\Delta_W^{(r)}+\Delta_R^{(r)}-\epsilon_W^{(r)}-\epsilon_R^{(r)}. \tag{160}$$

Therefore, as long as $\Delta_W^{(r)}+\Delta_R^{(r)}>\epsilon_W^{(r)}+\epsilon_R^{(r)}$, the joint memory utility improves in that round.

Proof. By telescoping decomposition,

$$\begin{aligned}&\mathcal{J}(\theta^{(r+1)},\phi^{(r+1)})-\mathcal{J}(\theta^{(r)},\phi^{(r)})\\&=\left[\mathcal{J}(\theta^{(r+1)},\phi^{(r+1)})-\mathcal{J}(\theta^{(r+1)},\phi^{(r)})\right]+\left[\mathcal{J}(\theta^{(r+1)},\phi^{(r)})-\mathcal{J}(\theta^{(r)},\phi^{(r)})\right].\end{aligned} \tag{161}$$

Substituting Eq. (158) and Eq. (159), respectively, gives the conclusion. □

F.3 Reader Reward Bias and Calibration Benefit

Definition F.2 (True utility and reader surrogate reward)**.** Let $U^\star(G)$ denote the true utility of graph memory G with respect to the downstream task, and let $\widehat{U}_\phi(G)$ denote the surrogate reward constructed from the readout result of reader R_ϕ. We say that the reader reward bias is at most ϵ_ϕ if, for all considered graphs G,

$$\left|\widehat{U}_\phi(G)-U^\star(G)\right|\leq\epsilon_\phi. \tag{162}$$

Theorem F.3 (Surrogate reward improvement to true utility improvement)**.** *If the reader reward bias is at most ϵ_ϕ, and the writer update improves the surrogate reward by*

$$\widehat{U}_\phi(G_{\theta'})-\widehat{U}_\phi(G_\theta)\geq\Delta, \tag{163}$$

then the true utility satisfies

$$U^\star(G_{\theta'})-U^\star(G_\theta)\geq\Delta-2\epsilon_\phi. \tag{164}$$

Proof. By the bias assumption,

$$U^\star(G_{\theta'})\geq\widehat{U}_\phi(G_{\theta'})-\epsilon_\phi,\qquad U^\star(G_\theta)\leq\widehat{U}_\phi(G_\theta)+\epsilon_\phi. \tag{165}$$

Subtracting the two inequalities gives

$$U^\star(G_{\theta'})-U^\star(G_\theta)\geq\widehat{U}_\phi(G_{\theta'})-\widehat{U}_\phi(G_\theta)-2\epsilon_\phi\geq\Delta-2\epsilon_\phi. \tag{166}$$

□

Corollary F.4 (Reader calibration reduces writer optimization bias)**.** *If the reader is calibrated from ϕ to ϕ' and reduces the reward bias from ϵ_ϕ to $\epsilon_{\phi'}$, where $\epsilon_{\phi'} < \epsilon_\phi$, then for the same surrogate reward improvement Δ, the lower bound on true utility improvement increases by*

$$2(\epsilon_\phi - \epsilon_{\phi'}). \tag{167}$$

Proof. By Theorem 5.3, the true utility improvement lower bound before calibration is $\Delta - 2\epsilon_\phi$, and after calibration it is $\Delta - 2\epsilon_{\phi'}$. Subtracting the two gives the result. □

F.4 Irreducible Bottlenecks of Single-sided Updates

Proposition 9 (Lower-bound bottlenecks of single-sided updates)*.* Assume that the overall error can be decomposed as

$$\mathcal{E}(\theta, \phi) = \mathcal{E}_{\mathrm{write}}(\theta) + \mathcal{E}_{\mathrm{read}}(\phi; \theta) + \epsilon_{\mathrm{int}}(\theta, \phi), \tag{168}$$

where all terms are nonnegative. If only the reader is updated, i.e., $\phi \mapsto \phi'$ while θ is fixed, then

$$\mathcal{E}(\theta, \phi') \geq \mathcal{E}_{\mathrm{write}}(\theta). \tag{169}$$

If only the writer is updated, i.e., $\theta \mapsto \theta'$ while ϕ is fixed, then

$$\mathcal{E}(\theta', \phi) \geq \mathcal{E}_{\mathrm{read}}(\phi; \theta'). \tag{170}$$

Therefore, reader-only updates cannot compensate for evidence chains that the writer has not written; writer-only updates cannot guarantee that a fixed reader can read out the evidence structures in the new graph distribution.

Proof. By the decomposition in Eq. (168) and the nonnegativity of all terms,

$$\mathcal{E}(\theta, \phi') = \mathcal{E}_{\mathrm{write}}(\theta) + \mathcal{E}_{\mathrm{read}}(\phi'; \theta) + \epsilon_{\mathrm{int}}(\theta, \phi') \geq \mathcal{E}_{\mathrm{write}}(\theta). \tag{171}$$

The second inequality is analogous. □

F.5 Stability of Closed-loop Graph Evolution and Parameter Updates

Theorem F.5 (Score drift control under multi-round self-evolution)**.** *Let the graph at round r be $G^{(r)}$, and the reader parameter be $\phi^{(r)}$. If single-step graph stability satisfies*

$$\left\| s_D(q, G^{(r+1)}; \phi^{(r)}) - s_D(q, G^{(r)}; \phi^{(r)}) \right\|_\infty \leq C_D \Delta_r, \tag{172}$$

where $\Delta_r = \Delta_{\mathrm{aug}}(G^{(r)}, G^{(r+1)}; q)$; and if the score is locally Lipschitz with respect to the parameter:

$$\left\| s_D(q, G; \phi) - s_D(q, G; \phi') \right\|_\infty \leq C_\phi \left\| \phi - \phi' \right\|_2, \tag{173}$$

then

$$\begin{aligned} &\left\| s_D(q, G^{(T)}; \phi^{(T)}) - s_D(q, G^{(0)}; \phi^{(0)}) \right\|_\infty \\ &\quad \leq \sum_{r=0}^{T-1} \left(C_D \Delta_r + C_\phi \left\| \phi^{(r+1)} - \phi^{(r)} \right\|_2 \right). \end{aligned} \tag{174}$$

Proof. For each r, adding and subtracting the intermediate term $s_D(q, G^{(r+1)}; \phi^{(r)})$ gives

$$\begin{aligned} &\left\| s_D(q, G^{(r+1)}; \phi^{(r+1)}) - s_D(q, G^{(r)}; \phi^{(r)}) \right\|_\infty \\ &\quad \leq \left\| s_D(q, G^{(r+1)}; \phi^{(r+1)}) - s_D(q, G^{(r+1)}; \phi^{(r)}) \right\|_\infty + \left\| s_D(q, G^{(r+1)}; \phi^{(r)}) - s_D(q, G^{(r)}; \phi^{(r)}) \right\|_\infty \\ &\quad \leq C_\phi \left\| \phi^{(r+1)} - \phi^{(r)} \right\|_2 + C_D \Delta_r. \end{aligned} \tag{175}$$

Summing over $r = 0, \ldots, T-1$ and using the triangle inequality gives the conclusion. □

Table 5: Training results for the memory writer. *GFM-pretrained-only* refers to using rewards fed back only by the pretrained memory reader, while *GFM-finetuned* further refers to using the fine-tuned memory reader.

Methods	Prec.↑	Recall↑	Deducible↑
GFM-pretrained-only	0.838	0.818	0.510
GFM-finetuned	0.824	0.813	0.512
RL-Recall	0.889	0.835	0.502
RL-F1	0.839	0.881	0.497
RL-Deduce	0.861	0.892	0.517
RL-Hybrid	**0.902**	**0.917**	0.522
Hybrid + frozen answer API	0.832	0.874	**0.526**

Corollary F.6 (High-probability multi-round stability)**.** *If* $\mathbb{P}[\Delta_r > \epsilon_r] \le \delta_r$*, then with probability at least* $1 - \sum_{r=0}^{T-1} \delta_r$*,*

$$\begin{aligned} &\left\| s_D(q, G^{(T)}; \phi^{(T)}) - s_D(q, G^{(0)}; \phi^{(0)}) \right\|_\infty \\ &\le \sum_{r=0}^{T-1} \left(C_D \epsilon_r + C_\phi \left\| \phi^{(r+1)} - \phi^{(r)} \right\|_2 \right). \end{aligned} \tag{176}$$

Proof. By a union bound, the event $\{\forall r, \Delta_r \le \epsilon_r\}$ holds with probability at least $1 - \sum_r \delta_r$. Applying Theorem 5.7 on this event gives the conclusion. □

G Analysis of the Memory Writer

This section further analyzes the memory writer while keeping the memory reader fixed. The experiments mainly use HotpotQA and MuSiQue. To further examine domain transfer capability, we also evaluate the trained writing policy on GRBench-Amazon, HaluMem-Medium, and LongMemEval-Oracle (Wu et al., 2025a; Chen et al., 2025a). We primarily report Precision, Recall, and Deducible: the first two measure whether the text retrieved by the reader covers the gold supporting contexts, while Deducible is determined by a judge as to whether the standard answer can be inferred from the retrieved context, thus more directly reflecting whether the graph memory is usable for reasoning.

Table 5 shows that different rewards have different preferences. Overall, RL-Hybrid achieves the best overall results, indicating that hybrid rewards can simultaneously constrain the selectivity and coverage of graph writing. *Hybrid + frozen answer API* achieves the highest Deducible but slightly lower retrieval Precision/Recall, suggesting that answer-side feedback helps improve reasoning usability, but may also make the writer more conservatively inclined to write evidence that directly supports the answer. Table 6 shows that a writer learned on HotpotQA/MuSiQue can transfer to GRBench, HaluMem, and LongMemEval, but continued training on the target domain still brings significant improvements. This indicates that the memory structure in agent memory scenarios is not entirely similar to that in traditional multi-hop QA, so target-domain feedback remains crucial. Table 7 further shows that the writing protocol and interaction budget affect the trade-off between coverage and noise. Relaxing the tight prompt can improve Recall, but reduces Deducible. Increasing the number of interaction steps can help complete cross-document bridging paths, whereas an overly large interaction budget or a loose writing protocol introduces redundant or noisy memory entries and can reduce downstream retrieval quality. More detailed reader-side sensitivity, training stability, and regularization analyses are provided in Appendix A.

H Ablation Study of the Memory Reader

We conduct ablation studies to isolate the contribution of each major component in the memory reader. All variants use the same writer-produced graph memory and the same retrieval budget unless the ablated component directly changes the retrieval mechanism. The ablations are organized around four questions: (1) whether structured query planning and global soft addressing are necessary for

Table 6: Cross-dataset memory writing results. "Base→Target" indicates direct evaluation on the target domain after training on the HotpotQA/MuSiQue base; "Target train→val" indicates training and validation on the target domain.

Settings	Prec.↑	Recall↑	Deducible↑
Base→GRBench	0.575	0.609	0.411
GRBench train→val	0.794	0.833	0.596
Base→HaluMem	0.230	0.448	0.299
HaluMem train→val	0.312	0.708	0.438
Base→LongMemEval	0.232	0.376	0.475
LongMem train→LongMemEval	0.377	0.439	0.531

Table 7: Ablation of writing protocols and interaction budgets. Tight=True indicates that the writer performs a single-round graph write under a stricter evidence budget, meaning the reader exposes only fewer, higher-confidence candidate pieces of evidence to the writer; Tight=False indicates that this evidence budget is relaxed, allowing the writer to access a broader candidate context. Iterative indicates that multi-round interactive writer-reader writing is enabled: the writer first writes the initial graph memory, the reader then returns retrieval feedback based on the current graph, and the writer continues to supplement or revise the graph structure. Here, 12/20/24 turns indicates the maximum number of interaction rounds allowed, and tight/loose indicates that strict or relaxed evidence budget constraints are still used during this multi-round interaction process.

Settings	Prec.↑	Recall↑	Deducible↑
Tight=True	0.836	0.806	0.515
Tight=False	0.845	0.851	0.506
Iterative, 12 turns, tight	0.852	0.829	0.516
Iterative, 20 turns, tight	0.835	**0.881**	0.522
Iterative, 24 turns, loose	**0.863**	0.826	**0.531**

recovering evidence from fragmented cues; (2) whether structurally gated propagation improves over uniform graph propagation; (3) whether cross-graph structural priors and target-graph calibration are both needed for evolving graph memory; and (4) whether reader training and entity-to-document projection are important for converting entity-level activation into document-level retrieval.

Table 8 summarizes the results. The first group evaluates how the reader handles fragmented cues. Removing structured query planning, alias/constraint cues, or global soft addressing forces the reader to rely more heavily on surface-level query matches or a small number of anchor entities, directly testing whether the complete evidence chain can still be recovered when distant bridge nodes are not initially activated. The second group studies whether the reader uses graph structure in a learned and selective way. Removing structural gates or replacing them with uniform message passing tests whether treating hub edges, bridge edges, redundant edges, and noisy shortcuts similarly harms retrieval. The third group examines the context–schema decomposition: the schema channel captures transferable structural reading patterns, while the context channel adapts the reader to the current writer-produced graph. The last group evaluates the selector, entity-to-document projection, GFM pre-training, and supervised retrieval fine-tuning.

The ablation results show the relative contribution of each memory-reader component. First, structured query planning and global soft addressing are important for fragmented-cue retrieval: removing them noticeably weakens performance, especially on MuSiQue and 2WikiMultiHopQA, where evidence chains are more likely to depend on implicit bridge entities. Second, structurally gated propagation consistently improves over uniform message passing, indicating that graph structure should be used selectively rather than as a fixed expansion rule. Among the structural inputs, graph-level summaries have a relatively smaller effect, while node-level and edge-pair features are more important for recognizing hubs, bridges, and cross-community evidence paths. Third, both the schema prior and context calibration channels contribute to performance, suggesting that the reader benefits from preserving transferable structural priors while adapting to the current writer-produced graph. Finally, supervised retrieval fine-tuning is essential for aligning the GFM reader with document-

Table 8: Ablation study of the memory reader on multi-hop QA retrieval. We report document-level Recall (%) at top-2 and top-5. All variants use the same writer-produced graph memory unless otherwise specified.

Reader Variant	HotpotQA		MuSiQue		2WikiMultiHopQA	
	R@2	R@5	R@2	R@5	R@2	R@5
SAGE	**65.1**	**77.6**	**43.2**	**53.1**	**83.6**	**88.6**
Query planning and global addressing						
SAGE w/o Structured Query Planning	62.7	75.1	40.4	50.1	80.7	86.6
SAGE w/o Global Soft Addressing	59.3	72.5	37.6	47.4	75.9	83.1
SAGE w/o Alias and Constraint Cues	63.0	75.8	41.0	50.8	80.8	86.9
SAGE w/ Anchor-only Initialization	58.6	71.4	36.8	46.5	74.2	82.4
Structurally conditioned propagation						
SAGE w/o Structural Gate	60.4	73.2	39.2	48.7	78.1	84.9
SAGE w/o Node Structural Features	62.1	74.8	40.5	50.0	80.0	86.0
SAGE w/o Edge-pair Structural Features	61.5	74.0	40.1	49.4	79.1	85.6
SAGE w/o Graph-level Summary	63.2	75.9	41.6	51.0	81.3	87.0
SAGE w/ Uniform Message Passing	58.9	71.8	37.5	46.9	75.3	82.9
Cross-graph priors and target-graph calibration						
SAGE w/o Schema Prior Channel	62.4	75.0	40.9	50.6	80.4	86.4
SAGE w/o Context Calibration Channel	61.8	74.3	40.0	49.8	79.5	85.9
SAGE w/o Context–Schema Fusion	60.7	73.5	39.1	48.6	77.8	84.7
Selector, projection, and reader training						
SAGE w/o Controlled Entity-to-Document Projection	60.9	73.9	38.7	48.2	77.2	84.4
SAGE w/o Query-conditioned Selector	63.9	76.5	41.9	52.0	82.2	87.6
SAGE w/ Vanilla GNN Reader	57.2	70.6	36.3	45.2	72.8	80.7

level evidence retrieval, whereas GFM pre-training provides transferable structural initialization that improves stability across datasets.

I Implementation Details of the Query-conditioned Subgraph Selection Regularizer

We provide the implementation details of the query-conditioned subgraph selector. In addition to the base entity scoring, it further learns a soft gating probability $\pi_e(q)$, which characterizes whether entity e should enter the reading subgraph of the current query q. This module performs lightweight reweighting of the final entity score, and constrains the reading subgraph through several structural regularizers during training.

Query-conditioned Selection Probability. Given a query q and a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, let $\mathbf{h}_e \in \mathbb{R}^d$ denote the representation of entity node $e \in \mathcal{V}$ after propagation by the GFM backbone, and let $\mathbf{z}_q \in \mathbb{R}^d$ denote the query representation output by the query encoder and then linearly projected. Here, d is the hidden dimension. The selector first projects the node representation and the query representation into the same selector space:

$$\mathbf{u}_e = W_n \mathbf{h}_e, \qquad \mathbf{v}_q = W_s \mathbf{z}_q, \tag{177}$$

where $W_n \in \mathbb{R}^{d_s \times d}$ is the node-side projection matrix, $W_s \in \mathbb{R}^{d_s \times d}$ is the query-side projection matrix, and d_s is the hidden dimension of the selector space. In our implementation, we set $d_s = d$, but the two do not have to be equal. Then, the selector obtains the selection logit of entity e with respect to query q through a scaled inner product:

$$\zeta_e(q) = \frac{\mathbf{u}_e^\top \mathbf{v}_q}{T_s}, \tag{178}$$

where $T_s > 0$ is the selector temperature coefficient. The final soft selection probability is defined as

$$\pi_e(q) = \mathrm{sigmoid}(\zeta_e(q)) = \frac{1}{1 + \exp(-\zeta_e(q))}. \tag{179}$$

Here, $\pi_e(q) \in (0, 1)$ can be understood as the soft probability that entity e is included in the reading subgraph of the current query. During training, we directly use $\pi_e(q)$ for differentiable optimization;

during inference, we can either continue to use the soft probability for reweighting, or obtain a discrete subgraph according to a threshold τ_π:

$$\mathcal{V}_q = \{e \in \mathcal{V} \mid \pi_e(q) > \tau_\pi\}, \qquad \mathcal{E}_q = \{(u,v) \in \mathcal{E} \mid u, v \in \mathcal{V}_q\}. \tag{180}$$

Let $a_e(q)$ denote the base entity score given by the GFM backbone reader. This score is usually obtained from the similarity between the node representation and the query representation, for example

$$a_e(q) = \mathbf{h}_e^\top \mathbf{z}_q. \tag{181}$$

Finally, we have:

$$a_e^{\text{final}}(q) = a_e(q) + \lambda_s \zeta_e(q), \tag{182}$$

where $\lambda_s \geq 0$ controls the influence of the selector logit on the final entity ranking.

Query–Subgraph Contrastive Regularizer. Using only Eq. (182) to fuse the selector score can easily lead to two types of degeneration: first, the selector may assign high probabilities to most nodes, thereby degenerating into full-graph activation; second, the selector may only learn local high-frequency entities, without forming a subgraph representation that is consistent with the overall semantics of the query. To this end, we first construct a query-conditioned subgraph representation weighted by the selection probabilities:

$$\bar{\mathbf{h}}_\pi(q) = \frac{\sum_{e \in \mathcal{V}} \pi_e(q) \mathbf{h}_e}{\sum_{e \in \mathcal{V}} \pi_e(q) + \epsilon}, \tag{183}$$

where $\epsilon > 0$ is a numerical stability term, which avoids an excessively small denominator when all $\pi_e(q)$ are close to 0. $\bar{\mathbf{h}}_\pi(q)$ can be understood as the semantic center of the soft subgraph activated by the current selector.

For a mini-batch $\mathcal{B} = \{q_i\}_{i=1}^B$, we treat $(\bar{\mathbf{h}}_\pi(q_i), \mathbf{z}_{q_i})$ within the same sample as a positive pair, and treat $(\bar{\mathbf{h}}_\pi(q_i), \mathbf{z}_{q_j})$, $j \neq i$, as in-batch negative pairs. The query–subgraph contrastive loss is defined as

$$\Omega_{\text{nce}} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp\left(\text{sim}(\bar{\mathbf{h}}_\pi(q_i), \mathbf{z}_{q_i})/T_n\right)}{\sum_{j=1}^{B} \exp\left(\text{sim}(\bar{\mathbf{h}}_\pi(q_i), \mathbf{z}_{q_j})/T_n\right)}, \tag{184}$$

where $T_n > 0$ is the contrastive learning temperature coefficient, and $\text{sim}(\cdot,\cdot)$ is the similarity function. In implementation, we usually apply ℓ_2 normalization to $\bar{\mathbf{h}}_\pi(q)$ and $\mathbf{z}_q$, and use inner-product similarity, so that $\text{sim}(\bar{\mathbf{h}}_\pi(q), \mathbf{z}_q) = \frac{\bar{\mathbf{h}}_\pi(q)^\top \mathbf{z}_q}{\|\bar{\mathbf{h}}_\pi(q)\|_2 \|\mathbf{z}_q\|_2}$. This term encourages the soft subgraph activated by the selector to semantically represent the current query, rather than only selecting nodes with high frequency or high centrality in the graph.

Size Regularizer. To prevent the selector from improving recall by activating a large number of nodes, we use the average selection probability as a size penalty:

$$\Omega_{\text{size}} = \frac{1}{|\mathcal{V}|} \sum_{e \in \mathcal{V}} \pi_e(q). \tag{185}$$

This term approximately represents the expected proportion of activated nodes. Minimizing Ω_{size} pushes the model to select a smaller reading subgraph. However, this term cannot be used alone; otherwise, the selector may degenerate into selecting too few nodes or even no nodes. Therefore, it needs to be jointly optimized with Ω_{nce} and the main retrieval loss: the former ensures query relevance, while the latter ensures that the selected structure can still support correct entity and document recall.

Connectivity Smoothing Regularizer. In addition to controllable size, an effective reading subgraph should also have local structural coherence. If the selection probabilities of adjacent nodes differ too much, the model may form several isolated activated points, making multi-hop paths difficult to explicitly utilize. To this end, we use a smoothing penalty on edges:

$$\Omega_{\text{con}} = \frac{1}{|\mathcal{E}|} \sum_{(u,v) \in \mathcal{E}} \left(\pi_u(q) - \pi_v(q)\right)^2. \tag{186}$$

Here, (u, v) is a directed or undirected edge in the graph, depending on whether edge directions are preserved during graph construction. If an undirected graph is used, $\mathcal{E}$ can be viewed as the symmetrized edge set. This term does not force all selected nodes to be strictly connected, but encourages adjacent nodes to have similar selection probabilities. In matrix form, if $\boldsymbol{\pi}(q) \in \mathbb{R}^{|\mathcal{V}|}$ is the selection probability vector composed of $\pi_e(q)$, and $\mathbf{L}$ is the graph Laplacian matrix, then this term is equivalent to Laplacian smoothing:

$$\Omega_{\text{con}} \propto \boldsymbol{\pi}(q)^\top \mathbf{L}\boldsymbol{\pi}(q). \tag{187}$$

Therefore, it is consistent with the classical assumption of graph signal smoothing: query relevance, as a soft signal on the graph, should maintain a certain degree of continuity within local neighborhoods.

Computational Complexity of the Selector Itself. Let the batch size be B, the number of nodes be $n = |\mathcal{V}|$, the number of edges be $m = |\mathcal{E}|$, and the hidden dimension be d. The cost of computing $W_n \mathbf{h}_e$ is $O(Bnd^2)$, the cost of computing $W_s \mathbf{z}_q$ is $O(Bd^2)$, and the cost of the inner-product logits is $O(Bnd)$. If we cache $W_n \mathbf{h}_e$ in advance, this term can be reduced to $O(Bnd)$. For the contrastive term, the cost of the subgraph pooling in Eq. (183) is $O(Bnd)$, and the cost of the in-batch NCE similarity matrix is $O(B^2 d)$; the size term has cost $O(Bn)$; the connectivity term needs to traverse edges and has cost $O(Bm)$. Therefore, the additional complexity of the selector during training is

$$O\left(Bnd^2 + Bnd + B^2 d + Bm\right), \tag{188}$$

and if the quadratic term of the linear projection is ignored or cached, it can be approximated as

$$O\left(Bnd + B^2 d + Bm\right). \tag{189}$$

During inference, if only the selector logit is used to fuse entity scores, without computing NCE, size, and connectivity regularizers, then the additional cost is mainly $O(Bnd^2 + Bnd)$, or $O(Bnd)$ under caching/lightweight projection.

J Training and Inference Complexity

This appendix separates three notions of cost that are easy to conflate: (i) the asymptotic complexity of the SAGE graph reader, (ii) one-time offline costs for graph construction, indexing, and training, and (iii) the wall-clock and LM-call cost of a fully uncached online query. This distinction is important because the retrieval-engine times in Table 4 measure the retriever-side stage rather than the complete agent pipeline. The best HotpotQA/2WikiMultiHopQA/MuSiQue evaluation runs also used precomputed query-planning/entity-linking artifacts, whereas a fresh online query must execute these stages explicitly.

Unless otherwise stated, the measured costs below come from the same shared server used for our experiments, equipped with $8\times$ NVIDIA A100-SXM4-80GB GPUs. We distinguish values recovered directly from logs, controlled latency measurements, and formula-based estimates. We do not fill in missing historical measurements with estimates from a different run.

J.1 End-to-End Query Pipeline and Accounting Boundary

A fully uncached query follows the pipeline

$$\begin{aligned} &q \to \text{planner extractor} \to \text{planner inferer} \to \text{LLM NER} \\ &\quad \to \text{local entity linking} \to \text{query embedding} \to \text{GFM retrieval/document ranking} \to \text{answer generation}. \end{aligned} \tag{190}$$

The query planner uses two serial LM calls. The extractor identifies explicit entities, candidate aliases, relation clues, hard constraints, and the answer type. The inferer consumes the extractor output and generates at most three pseudo-queries together with confidence scores. The pipeline then applies one additional LM call for NER. Entity linking is performed locally with `colbert-ir/colbertv2.0`; query encoding, graph retrieval, and document ranking are also local computations. Finally, the top retrieved documents are passed to the answer model, which contributes one LM call.

Therefore, in the normal successful uncached path, the online system uses four LM calls per question:

Table 9: LM calls in one successful fully uncached online query. Entity linking and graph retrieval are local and do not invoke a remote LM.

Stage	LM calls / query
Planner extractor	1
Planner inferer	1
LLM NER	1
Local ColBERT entity linking	0
Query embedding + GFM retrieval	0
Answer generation	1
Total	**4**

Cached benchmark evaluation versus uncached deployment. The best HotpotQA, 2WikiMultiHopQA, and MuSiQue Stage-3 evaluations did not invoke the planner and NER from scratch inside Stage 3. Instead, the pipeline first exported a `rewriter.jsonl` file containing the planned explicit entities, pseudo-queries, and linked entities, and Stage 3 reused these artifacts. Consequently, the logged Stage-3 wall-clock contains local query encoding, graph retrieval, document ranking, answer generation, retries, and evaluation, but excludes the two planner calls, the LLM-NER call, and online entity linking. We therefore report cached evaluation time and fully uncached online latency separately.

Scope of Table 4. The values in Table 4 should be interpreted as *retrieval-engine* measurements, not as end-to-end agent latency. In particular, they should not be compared directly with the uncached latency numbers below, which additionally account for query planning, NER, entity linking, and answer generation.

J.2 Notation

For the asymptotic analysis, suppose that the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ contains $n = |\mathcal{V}|$ entity nodes and $m = |\mathcal{E}|$ entity-relation edges. If self-loops are added during GCN propagation, we denote $\tilde{m} = m + n$. Let the hidden dimension be d, the number of propagation layers be L, the batch size be B, and the number of pseudo-queries be M. One original query together with M pseudo-queries therefore requires at most $M + 1$ graph reads.

Let $\mathbf{M} \in \{0, 1\}^{n \times N_D}$ denote the sparse entity–document association matrix, where N_D is the number of documents and $\mathrm{nnz}(\mathbf{M})$ is the number of nonzero entity–document links. Let K_e be the number of top entities used for document projection, K the number of final retrieved documents, and $\bar{f}$ the average number of documents linked to the selected entities.

J.3 Offline Structural Features and Indexing

Let the dimensions of node-, edge-, and graph-level structural features be p_n, p_e, and p_g, respectively. In our implementation these are small constants.

Given an adjacency list, degrees and average neighbor degrees can be computed in $O(n + m)$ time. Clustering coefficients and common-neighbor features require neighbor-set intersections, with cost

$$O\left(n + m + \sum_{(u,v)\in\mathcal{E}} \min\{\deg(u), \deg(v)\}\right). \tag{191}$$

For sparse graphs or graphs with bounded average degree, Eq. (191) is close to $O(n + m)$, while an extremely dense worst case can reach $O(n^3)$. The structural features and the entity–document matrix can be precomputed and cached, requiring

$$O\left(np_n + mp_e + p_g + \mathrm{nnz}(\mathbf{M})\right) \tag{192}$$

space. Because these quantities are query-independent, they can be reused across queries and across repeated reader evaluations of the same candidate graph in the self-evolution loop.

J.4 Forward Complexity of the Structurally Gated GFM

We first consider one forward pass for one query on one graph. A standard GCN layer consists of a node-wise linear transformation and sparse adjacency aggregation. The corresponding cost is

$$C_{\text{plain}} = O\left(nd^2 + \tilde{m}d\right). \tag{193}$$

A structurally gated layer additionally generates a vector gate for each edge:

$$\mathbf{m}_{u\to v} = \gg_{uv} \odot W\mathbf{h}_u, \tag{194}$$

where $\mathbf{h}_u$ is the source-node representation, $W \in \mathbb{R}^{d\times d}$ is the node transformation, $\gg_{uv} \in \mathbb{R}^d$ is the structural gate for edge (u, v), and $\odot$ is element-wise multiplication. Let d_g denote the structural-feature encoding dimension and h_g the hidden width of the gating MLP. When the gate uses source-node structure, target-node structure, edge-pair structure, and a graph-level summary, its generation cost is

$$\begin{aligned} C_{\text{gate}} = O\Big(&np_n d_g + mp_e d_g + p_g d_g \\ &+ m(4d_g h_g + h_g d)\Big). \end{aligned} \tag{195}$$

The first three terms encode node, edge, and graph structural features; the last term is the per-edge gating MLP. If edge-pair features or graph summaries are disabled, their terms can be removed. If d_g and h_g scale with d, gate generation is $O(md^2)$ in the worst case. If the gate is a small constant-width or low-rank module, its practical scaling approaches $O(md)$. A structurally gated layer therefore costs

$$C_{\text{gated}} = O\left(nd^2 + \tilde{m}d + C_{\text{gate}}\right). \tag{196}$$

The implementation can combine a holistic structurally gated branch with a specific prompt branch. Let $\rho_{\text{plain}}, \rho_{\text{gated}} \in \{0, 1\}$ indicate whether the corresponding branches are enabled. For a batch of B queries,

$$C_{\text{enc}}(B) = O\left(BL\left[\rho_{\text{plain}} C_{\text{plain}} + \rho_{\text{gated}} C_{\text{gated}}\right]\right). \tag{197}$$

The factor B appears because the current implementation constructs query-conditioned node inputs and performs graph encoding for each query. Caching query-independent structural features reduces constants but does not change the conservative upper bound in Eq. (197).

For the simplified setting used in the main-text complexity discussion, $B = 1$, $M = 0$, only one branch is active, and the gating network is treated as lightweight. Equation (197) then reduces to

$$O\left(L(md + d^2 n)\right), \tag{198}$$

where md represents sparse edge-level message propagation/gating and $d^2 n$ the node-wise linear projections.

J.5 Entity Scoring, Selector Regularization, and Document Projection

After graph encoding, the entity score is

$$a_e(q) = \mathbf{h}_e^\top \mathbf{z}_q. \tag{199}$$

For a batch of B queries, scoring all n entities costs

$$C_{\text{score}}(B) = O(Bnd). \tag{200}$$

If the query-conditioned subgraph selector in Appendix I is enabled, its inference-stage overhead is

$$C_{\text{sel,infer}}(B) = O(Bnd^2 + Bnd), \tag{201}$$

and can approach $O(Bnd)$ when the node-side projection is cached or a lightweight projection is used. During training, the NCE, size, and connectivity regularizers add

$$C_{\text{sel,train}}(B) = O(Bnd^2 + Bnd + B^2 d + Bm), \tag{202}$$

where $B^2 d$ is the in-batch query–subgraph contrastive term and Bm corresponds to edge-level connectivity smoothing.

Entity scores are projected to documents through the sparse association matrix $\mathbf{M}$. Full sparse matrix multiplication costs

$$C_{\mathrm{doc,full}}(B) = O\left(B \operatorname{nnz}(\mathbf{M})\right). \tag{203}$$

The current `IDFWeightedRanker` follows this sparse-projection form. If only the top K_e entities are projected through inverted lists, the conceptual cost is

$$C_{\mathrm{doc,top}K}(B) = O\left(Bn \log K_e + BK_e\bar{f}\right), \tag{204}$$

where the first term selects top entities and the second accesses their linked documents. Final top-K ranking over all N_D documents costs $O(BN_D \log K)$; ranking only a candidate pool costs $O(BN_{\mathrm{cand}} \log K)$, where $N_{\mathrm{cand}} \ll N_D$.

J.6 Training-Time Complexity

The reader-training cost is dominated by GFM forward propagation, entity-level retrieval loss, the optional selector regularizer, and backpropagation. Let κ_{bw} denote the constant-factor cost of backpropagation relative to the forward computation. For BCE, ranking loss, or ListCE over all entity scores,

$$C_{\mathrm{loss}}(B) = O(Bn). \tag{205}$$

Without the optional selector,

$$C_{\mathrm{train}} = O\left(\kappa_{\mathrm{bw}}\left[C_{\mathrm{enc}}(B) + C_{\mathrm{score}}(B) + C_{\mathrm{loss}}(B)\right]\right). \tag{206}$$

With the selector,

$$C^{\mathrm{sel}}_{\mathrm{train}} = O\left(\kappa_{\mathrm{bw}}\left[C_{\mathrm{enc}}(B) + C_{\mathrm{score}}(B) + C_{\mathrm{loss}}(B) + C_{\mathrm{sel,train}}(B)\right]\right). \tag{207}$$

If document-level supervision is added explicitly, document projection adds either $C_{\mathrm{doc,full}}(B)$ or $C_{\mathrm{doc,top}K}(B)$.

The GFM pretraining/fine-tuning computation itself uses cached graph, entity–document, and embedding artifacts and therefore does not require remote LM calls during each optimization step. LM use in raw graph construction is a separate one-time preprocessing cost quantified below.

J.7 Inference-Time Complexity

Let $C_{\mathrm{prep}}(q)$ denote the query-preparation cost before graph reading, including query planning, NER, entity linking, and query encoding. The graph-reader portion for one original query and M pseudo-queries is

$$\begin{aligned} C_{\mathrm{reader}}(q) =& (M+1)\Big[C_{\mathrm{enc}}(1) + C_{\mathrm{score}}(1) + C_{\mathrm{sel,infer}}(1) + C_{\mathrm{doc}}(1)\Big] \\ &+ C_{\mathrm{fuse}}(M, K), \end{aligned} \tag{208}$$

where $C_{\mathrm{doc}}(1)$ is either full sparse projection or top-K_e inverted projection; $C_{\mathrm{sel,infer}}(1) = 0$ when the optional selector is disabled. If every query retains K candidate documents, simple weighted fusion costs

$$C_{\mathrm{fuse}}(M, K) = O\left((M+1)K \log((M+1)K)\right). \tag{209}$$

The complete uncached online cost additionally includes query preparation and answer generation:

$$C_{\mathrm{online}}(q) = C_{\mathrm{prep}}(q) + C_{\mathrm{reader}}(q) + C_{\mathrm{answer}}(q). \tag{210}$$

This decomposition is the one used in the latency audit below. In particular, the small graph-reader latency reported in Table 4 should not be substituted for $C_{\mathrm{online}}(q)$.

J.8 Space Complexity

Offline graph storage requires

$$O\left(n + m + \operatorname{nnz}(\mathbf{M}) + np_n + mp_e + p_g\right). \tag{211}$$

During training, storing layer-wise node activations requires

$$O(BLnd) \tag{212}$$

Table 10: Audited compute for the recoverable epoch-97 GFM pretraining checkpoint. GPU-hours include the resumed segment and its recomputation.

Item	Value
Hardware used by run	3× A100-80GB
Backbone	6-layer GCN
Hidden dimension	256
Training shards	273
Checkpoint parameters	22,233,918
Logged wall-clock compute	≈ 19 h 13 min
Logged GPU-hours	≈ 57.65

space. Fully materializing all edge gates/messages would require $O(md)$ GPU memory. The implementation instead streams edges in chunks. For chunk size c, the peak edge-message memory becomes

$$O(cd), \qquad c \ll m. \tag{213}$$

If `return_gate` is enabled and all edge gates are retained for visualization, memory returns to $O(md)$. Materializing all document scores requires $O(BN_D)$ space, while maintaining only a candidate heap can reduce this to $O(BK)$ or $O(BK_e\bar{f})$.

J.9 Measured Training Compute

Audited pretraining checkpoint. The training server contains eight A100-SXM4-80GB GPUs, but the auditable pretraining run used three A100-80GB GPUs. The corresponding configuration uses a 6-layer GCN backbone with hidden dimension 256, GraphCL pretraining, 128 negative samples, Adam with learning rate 5×10^{-4}, activation checkpointing, and 273 training shards. The epoch-97 checkpoint contains 22,233,918 parameters.

The run is recoverable from two logged segments. The first segment ran for 18 h 26 min 43 s on three GPUs, corresponding to approximately 55.34 GPU-hours. The second resumed from epoch 94 and recomputed through epoch 97, running for 46 min 17 s, or approximately 2.31 GPU-hours. Counting the recomputed work, the total auditable cost to the epoch-97 checkpoint is therefore approximately

$$57.65 \text{ GPU-hours} \tag{214}$$

and 19 h 13 min of aggregate wall-clock training time on three GPUs.

Limitation of the historical fine-tuning record. The historical directory corresponding to the fine-tuning run that produced the best reported HotpotQA/2WikiMultiHopQA/MuSiQue checkpoint is no longer available. Its exact wall-clock time and peak-memory trace therefore cannot be reconstructed. We intentionally do not substitute the duration of a different fine-tuning attempt and present it as the cost of the best model. The 57.65 GPU-hour number above should consequently be interpreted as the auditable pretraining cost, not as a claim about the complete historical end-to-end training budget.

Configuration-matched fine-tuning reference. For scale, a later fine-tuning attempt with the closest recoverable configuration used five A100-80GB GPUs, batch size 1, gradient accumulation 4, 100 batches per epoch, and the same 273 training shards. Its first epoch, including validation and checkpointing, took approximately 3 h 20 min, or 16.68 GPU-hours. That run was terminated during epoch 2 and did not produce the best reported checkpoint. We therefore report this number only as a configuration-level reference and do not extrapolate it into an "actual" fine-tuning cost for the final model.

J.10 Offline LM Cost for Raw Graph Construction

Raw Stage-1 graph construction performs one LLM-NER call and one OpenIE call per document, and one LLM-NER call per QA question. If no processed artifact is cached, the one-time LM-call count is therefore

$$N_{\text{LM,Stage1}} = 2N_{\text{doc}} + N_{\text{question}}. \tag{215}$$

Table 11: One-time Stage-1 LM-call estimate if the 273 training shards are rebuilt from raw data without cached preprocessing. Document counts are raw document instances before cross-shard deduplication.

Dataset	Shards	Document instances	QA questions	Stage-1 LM calls
2WikiMultiHopQA	168	966,907	167,454	2,101,268
HotpotQA	89	864,849	88,447	1,818,145
MuSiQue	16	318,724	15,938	653,386
Total	**273**	**2,150,480**	**271,839**	**4,572,799**

Table 12: Audited per-question token footprint. Planner-inferer input is a lower bound because not all intermediate planner fields are retained. The answer input contains the retrieved top-10 document context.

Dataset	Planner extractor input	Planner inferer input	NER input	Answer input	Answer output
HotpotQA	106	124	82	2,234	105
2WikiMultiHopQA	102	120	78	1,949	87
MuSiQue	107	124	83	2,013	100

For the 273 multi-hop training shards used in the audit, the raw instance counts imply the call budget in Table 11. These are *rebuild estimates*: the audited GFM training run reused processed artifacts and did not regenerate all of these calls.

The 2,150,480 document instances contain substantial duplication across shards. Normalizing and deduplicating by title–content pairs leaves 1,101,879 unique pairs. If such global deduplication is performed before both document-level LM calls, the estimated Stage-1 budget decreases by

$$2(2{,}150{,}480 - 1{,}101{,}879) = 2{,}097{,}202 \tag{216}$$

calls, from 4,572,799 to 2,475,597 calls, a 45.9% reduction. This identifies cross-shard deduplication as a direct optimization for future raw rebuilds.

For reference, rebuilding the three 1,000-question evaluation corpora from raw data would require 19,442 Stage-1 calls for HotpotQA (9,221 documents), 13,238 for 2WikiMultiHopQA (6,119 documents), and 24,312 for MuSiQue (11,656 documents). Again, these are one-time raw-rebuild costs rather than costs repeated for each evaluation.

J.11 Online LM Calls and Token Footprint

Under the successful uncached path in Table 9, 1,000 queries require 4,000 successful LM calls: 1,000 planner-extractor calls, 1,000 planner-inferer calls, 1,000 NER calls, and 1,000 answer-generation calls.

We additionally audited the prompt/response content that remained available from the multi-hop evaluation pipeline. The answer prompts and responses can be reconstructed directly from saved predictions. Planner outputs are only partially preserved, and raw NER outputs are not fully preserved; therefore, planner/NER token counts are lower-bound/proxy estimates rather than provider billing records. Message-framing tokens are not included.

The answer context dominates the recoverable input-token footprint. Across the three datasets, the top-10 answer prompt accounts for approximately 86–88% of the audited input tokens. For roughly 1,000 successful queries, the lower-bound four-call input totals are 2.545M tokens for HotpotQA, 2.248M for 2WikiMultiHopQA, and 2.327M for MuSiQue; the corresponding recoverable output-token proxies are 0.296M, 0.267M, and 0.288M, respectively. We report token volumes instead of a fixed dollar figure because provider pricing can change over time.

Retries in the historical benchmark runs. The Stage-3 answer wrapper allowed up to five attempts and slept for 30 s after an error. The saved logs contain 260 API-error events for HotpotQA, 223 for 2WikiMultiHopQA, and 156 for MuSiQue. With 999, 999, and 1,000 saved answers, respectively, the

Table 13: Measured latency of audited online stages. The graph query-encoding/GFM/document-ranking stage was not separately timed in this controlled audit and is therefore not included in the 13.22 s sum below.

Stage	N	Mean (s)	Median (s)	P95 (s)
Planner (2 serial LM calls)	20	4.72	4.40	12.93
LLM NER	20	1.52	1.30	2.63
Entity linking	20	3.08	2.25	14.30
Answer generation	10	5.55	5.27	10.22

logs imply at least 1,259, 1,222, and 1,156 answer request attempts. The dominant recorded errors were HTTP 429 rate-limit events while the three evaluations were running concurrently under a shared API project. These retries affect wall-clock time, but rejected requests should not automatically be interpreted as successful billed generations.

J.12 Measured Stage-Wise Online Latency

We performed a controlled latency audit using the first 20 saved HotpotQA queries on the same shared server. Query planning was measured on 20 examples (two serial LM calls per example), LLM NER on 20 examples, and local entity linking on 20 examples. Answer generation was measured on the first 10 examples using the original top-10 retrieved-document context. The entity-linking test used an otherwise available A100 GPU and searched over 87,768 HotpotQA graph entities. Because the server was shared and under nontrivial load, these values describe the measured environment rather than an ideal dedicated deployment.

The two planner calls are necessarily serial in the current implementation: the inferer consumes the extractor JSON, so they cannot be issued in parallel without changing the planner. The measured median of the directly audited stages is

$$4.40 + 1.30 + 2.25 + 5.27 = 13.22 \text{ s}. \tag{217}$$

This is *not* the full end-to-end median. The controlled audit did not separately measure Qwen3-Embedding-8B query encoding, GFM forward passes for the primary/pseudo queries, document ranking, or retrieval-result fusion. The only supported statement from these measurements is therefore

$$T_{\text{uncached,online}} > 13.22 \text{ s} \quad \text{(median in the audited environment)}. \tag{218}$$

Entity-linking initialization. Each query supplied a mean of 7.85 mentions to ColBERT (median 7, P95 14), with top-5 linked entities per mention. The first construction of the HotpotQA entity index required 10.42 s to load the ColBERT model and 441.65 s to build the index. When an existing index was reused, model loading took 3.87 s and index reuse 2.82 s. The first searcher query also loads search structures and CUDA components, explaining the 14.30 s P95/cold-query value in Table 13. Index construction is an offline startup cost, and a deployed service should build/reuse and warm the index before serving online queries rather than charge this cost to every request.

J.13 Observed Stage-3 Evaluation Wall-Clock

For completeness, Table 14 reports the wall-clock of the historical best Stage-3 multi-hop evaluations. Each run contains 1,000 questions, uses `n_threads=5` for concurrent answer calls, supplies top-10 documents to the answer model, and reuses precomputed planner/NER/entity-linking results.

These Stage-3 runs include model/data initialization, local query embedding, primary and pseudo-query graph retrieval, document ranking, five-thread answer generation, retry sleeps, and the final evaluator. They exclude the precomputed two-call planner, LLM NER, and online entity-linking stages. Moreover, the three datasets were evaluated concurrently and shared the same API project's rate limit, so these wall-clock values should not be interpreted as isolated single-user latency.

J.14 Complexity Interpretation and Comparison

Different retrieval architectures place their cost in different parts of the pipeline. Dense single-step RAG primarily pays for query encoding and nearest- neighbor search over a prebuilt index.

Table 14: Observed Stage-3 evaluation wall-clock. These times are not single-query uncached latencies: planner/NER/entity linking are precomputed, five answer calls can run concurrently, and the runs include rate-limit retries and evaluation overhead.

Dataset	Wall-clock	Wall-clock / question	Throughput (q/s)
HotpotQA	2 h 30 min 45 s	9.05 s	0.110
2WikiMultiHopQA	2 h 00 min 24 s	7.23 s	0.138
MuSiQue	2 h 30 min 40 s	9.04 s	0.111

Iterative RAG adds repeated retrieval and, when generation/reasoning is interleaved, repeated LM calls. GraphRAG systems can shift substantial work to offline graph construction, indexing, or summary generation. SAGE likewise has non-negligible offline graph-construction cost, but once graph artifacts are cached, its core reader uses one primary graph read plus at most a small number of pseudo-query reads, followed by sparse entity-to-document projection.

The practical implication for self-evolving memory is not that the entire SAGE agent has millisecond end-to-end latency. Rather, the *graph-reader feedback path* is a reusable local computation once query preparation and indices are available. This makes it possible to score candidate memory graphs repeatedly without launching a new multi-round agentic search for every graph candidate. At the same time, the measured audit shows that a fully uncached online system is dominated by LM query planning, NER/entity preparation, and answer generation, so the retriever-only times in Table 4 must be interpreted within that narrower scope.

Finally, the paired long-term-memory host experiments in Appendix S are accuracy-oriented interventions. The available logs do not contain a matched host-level end-to-end latency audit for every Native/+SAGE configuration. We therefore do not infer host-level latency improvements from the multi-hop reader timings reported here.

K Implementation Details of Structured Query Planning

K.1 Detailed definitions of the notation and additional information

In $\mathcal{P}_\omega(q) = \left(\mathcal{E}_{\text{exp}}, \mathcal{A}, \mathcal{C}_{\text{rel}}, \mathcal{C}_{\text{hard}}, \tau, \{(\tilde{q}_m, \alpha_m, t_m)\}_{m=1}^{M}\right)$, $\mathcal{E}_{\text{exp}}$ acts as a direct anchor for memory (explicit entities); $\mathcal{A}$ maps the brain's multiple representational habits for the same concept (aliases); $\mathcal{C}_{\text{rel}}$ simulates the relational network in semantic memory; $\mathcal{C}_{\text{hard}}$ serves as the spatiotemporal and logical boundaries of episodic memory (such as hard constraints like time and location); τ presets the cognitive template of the target memory (answer type); and the pseudo-queries $\tilde{q}_m$ with confidence α_m and intent t_m are analogous to the multiple exploratory recalls conducted in the human mind (Simulated Recall).

K.2 Two-stage Planning: Extraction and Inference

Natural-language questions often compress key retrieval cues into implicit relations, such as "the birthplace of the author", "the publication year of the only mystery novel of a certain work", or "the death date of the father". If the original question is sent as a whole to the entity linker, the system can easily hit only surface entities while missing bridge entities or answer-type constraints. Therefore, the query planner is defined as a structured function

$$\mathcal{P}(q) = \left(\mathcal{E}_{\text{exp}}, \mathcal{A}, \mathcal{C}_{\text{rel}}, \mathcal{C}_{\text{hard}}, \tau, \{(\tilde{q}_m, \alpha_m)\}_{m=1}^{M}\right). \tag{219}$$

It consists of two stages: *Extractor* 5 extracts explicit entities, aliases, relation clues, hard constraints, and the answer type; *Inferer* 6 generates at most M retrieval intents based on the extraction results.

Extractor prompt template.

```
You are a retrieval planner for graph-based multi-hop QA.
Question:
{QUESTION}

Extract structured retrieval signals.
Return JSON only with keys:
{
  "explicit_entities": [string],
  "candidate_aliases": {"entity": [alias]},
  "relation_clues": [string],
  "constraints": {},
  "answer_type": "string"
}
Rules: keep entries short, avoid explanations, keep empty fields as [] or {}.
```

Figure 5: Metadata of the case study from HotpotQA.

Inferer prompt template.

```
You are a retrieval planner for graph-based multi-hop QA.
Question:
{QUESTION}

Structured extraction:
{EXTRACTOR_JSON}

Generate at most M retrieval intents that help locate:
- evidence directly supporting the target relation;
- bridge entities required for multi-hop reasoning;
- documents likely to contain the target attribute;
- evidence satisfying temporal, spatial, type, comparison or negation constraints;
- evidence using aliases or alternative mentions.
Return JSON only with keys:
{
  "pseudo_queries": [string],
  "rewriter_confidence": [number]
}
```

Figure 6: Metadata of the case study from HotpotQA.

L Computation Details of Topological Structural Features

L.1 Normalized Structural Graph

Structural features are computed on an undirected, self-loop-free, binarized adjacency matrix $\mathcal{A}_s$:

$$\mathcal{A}_s = \mathbb{I}[(\mathcal{A} + \mathcal{A}^\top) > 0], \qquad \operatorname{diag}(\mathcal{A}_s) = 0. \tag{220}$$

This avoids drastic fluctuations in topological statistics caused by unstable relation-extraction directions. Message propagation can still use the original bidirectional edges or relation-aware graph; structural statistics are only used as gating conditions.

L.2 Node-level Structural Features

For node v, let $\mathcal{N}(v) = \{u : \mathcal{A}_{s,uv} = 1\}$ and $d_v = |\mathcal{N}(v)|$. The node-level features are

$$\phi(v) = \big[\log(1 + d_v), c_v, \kappa_v, \bar{d}_{\mathcal{N}(v)}\big]. \tag{221}$$

The local clustering coefficient is

$$c_v = \begin{cases} \dfrac{2T_v}{d_v(d_v - 1)}, & d_v \geq 2, \\ 0, & d_v < 2, \end{cases} \tag{222}$$

where T_v is the number of undirected edges inside the neighborhood of v; κ_v is the core number; and the average neighbor degree is

$$\bar{d}_{\mathcal{N}(v)} = \begin{cases} \dfrac{1}{d_v} \sum_{u \in \mathcal{N}(v)} d_u, & d_v > 0, \\ 0, & d_v = 0. \end{cases} \tag{223}$$

These quantities respectively characterize node frequency, local clustering, core/peripheral position, and neighborhood density. For RAG memory, they correspond to four common structural risks: over-propagation by high-frequency hubs, redundant diffusion inside clustered regions, ignored peripheral bridge entities, and scale mismatch between sparse and dense regions.

L.3 Edge-pair Structural Features

For an undirected structural edge (u, v), the pairwise features are

$$\psi(u, v) = \big[|d_u - d_v|, \mathrm{CN}(u, v), \mathrm{Jac}(u, v)\big], \tag{224}$$

where

$$\mathrm{CN}(u, v) = |\mathcal{N}(u) \cap \mathcal{N}(v)|, \qquad \mathrm{Jac}(u, v) = \frac{|\mathcal{N}(u) \cap \mathcal{N}(v)|}{|\mathcal{N}(u) \cup \mathcal{N}(v)| + \epsilon}. \tag{225}$$

Degree difference reflects cross-level connections, while common neighbors and Jaccard reflect local community overlap. Based on these features, the gate can distinguish intra-community evidence aggregation edges from cross-community bridge edges.

L.4 Graph-level Summary and Normalization

The graph-level summary concatenates the mean, standard deviation, and density of node features:

$$\mathbf{r}_{\mathcal{G}} = \big[\,\mathrm{mean}_{v\in\mathcal{V}}\,\phi(v); \mathrm{std}_{v\in\mathcal{V}}\,\phi(v); \mathrm{dens}(\mathcal{G})\big], \tag{226}$$

where

$$\mathrm{dens}(\mathcal{G}) = \begin{cases} \dfrac{2m_s}{n(n-1)}, & n \geq 2, \\ 0, & n < 2, \end{cases} \tag{227}$$

$n = |\mathcal{V}|$, and m_s is the number of undirected structural edges. To remove graph-size differences, node and edge features are z-scored within each graph, and the graph-level summary computes global mean and standard deviation over the set of training graphs:

$$\bar{\mathbf{r}}_{\mathcal{G}} = \frac{\mathbf{r}_{\mathcal{G}} - \mu_r}{\sigma_r + \epsilon}. \tag{228}$$

If the standard deviation of a certain dimension is close to zero, we only perform centering to avoid division by an unstable small value.

L.5 Gating Input Encoding

For each message edge $u \to v$, the structural gate reads the source node, target node, pairwise features, and graph-level summary:

$$\bar{\phi}(u) = \mathrm{NormNode}(\phi(u)), \qquad \bar{\psi}(u, v) = \mathrm{NormPair}(\psi(u, v)), \tag{229}$$

$$u_u^{(l)} = E_n^{(l)}(\bar{\phi}(u)), \qquad u_v^{(l)} = E_n^{(l)}(\bar{\phi}(v)), \tag{230}$$

$$v_{uv}^{(l)} = E_p^{(l)}(\bar{\psi}(u, v)), \qquad r_{\mathcal{G}}^{(l)} = E_g^{(l)}(\bar{\mathbf{r}}_{\mathcal{G}}). \tag{231}$$

The encoders E_n, E_p, E_g are all two-layer MLPs. The concatenated gating input is

$$\mathbf{z}_{uv}^{(l)} = [\mathbf{u}_u^{(l)}; \mathbf{u}_v^{(l)}; \mathbf{v}_{uv}^{(l)}; \mathbf{r}_{\mathcal{G}}^{(l)}]. \tag{232}$$

The gate itself is a vector rather than a scalar:

$$\mathbf{g}_{uv}^{(l)} = \mathbf{1} + \delta \tanh\big(\mathrm{MLP}_g^{(l)}(\mathbf{z}_{uv}^{(l)})\big), \qquad \delta = 0.1. \tag{233}$$

The last layer of the gating MLP is initialized to zero, so initially $\mathbf{g}_{uv}^{(l)} = \mathbf{1}$. At the beginning of training, the model does not destroy the original propagation scale; the learned structural bias gradually emerges in a residual manner.

L.6 Message Propagation with Normalized Weights

Let $\tilde{\mathbb{E}}$ be the edge set after adding self-loops. Structural gates are used for non-self-loop edges, and unit gates are used for self-loops. The GCN normalization coefficient is

$$\eta_{uv} = \frac{w_{uv}}{\sqrt{\tilde{d}_u \tilde{d}_v}}, \qquad \tilde{d}_v = \sum_{u:(u,v)\in\tilde{\mathbb{E}}} w_{uv}, \tag{234}$$

where w_{uv} defaults to 1, but can also come from edge weights. The propagation at layer l is

$$\mathbf{m}_{u\to v}^{(l)} = \eta_{uv}\, \mathbf{g}_{uv}^{(l)} \odot W^{(l)} \mathbf{h}_u^{(l-1)}, \tag{235}$$

$$\mathbf{h}_v^{(l)} = \sigma\left(\mathbf{b}^{(l)} + \sum_{u:(u,v)\in\tilde{\mathbb{E}}} \mathbf{m}_{u\to v}^{(l)}\right). \tag{236}$$

The multi-layer wrapper also contains inter-layer residuals: when $l > 1$ and the dimensions are consistent,

$$\mathbf{H}^{(l)} \leftarrow \mathbf{H}^{(l)} + \mathbf{H}^{(l-1)}. \tag{237}$$

This residual and Eq. (233) form a dual stability mechanism: the former stabilizes deep propagation, while the latter stabilizes structural modulation.

L.7 Chunked Gating and GPU Memory Complexity

Explicitly storing all gates requires $O(|\mathbb{E}|d)$ GPU memory. For large graphs, gates are computed by edge chunks:

$$\mathbb{E} = \bigcup_{b=1}^{B_e} \mathbb{E}_b, \qquad |\mathbb{E}_b| \le C_e. \tag{238}$$

Each edge chunk sequentially executes

$$\mathbf{g}_b \to \mathbf{m}_b \to \mathrm{scatter_add}(\mathbf{m}_b), \tag{239}$$

and immediately releases the intermediate gate tensor. Online GPU memory is reduced from $O(|\mathbb{E}|d)$ to $O(C_e d)$, while the time complexity remains linear, $O(|\mathbb{E}|d)$. This is especially important for self-evolving memory, because the same reader needs to repeatedly evaluate candidate graphs produced by different writers.

M Pretraining Objective and Augmented Views

M.1 GraphCL View Construction

The goal of the pretraining stage is to learn cross-graph transferable structural–semantic propagation, rather than fitting specific question-answering labels. Given the original graph view $(\mathcal{G}_0, X_0)$, we construct two augmented views $(\mathcal{G}_1, X_1)$, $(\mathcal{G}_2, X_2)$ and one negative feature view $(\mathcal{G}_0, X^-)$. The augmentation types include edge perturbation, feature masking, node perturbation, and subgraph sampling; let the augmentation operators be $\mathcal{A}_1, \mathcal{A}_2$, then

$$(\mathcal{G}_j, X_j) = \mathcal{A}_j(\mathcal{G}_0, X_0), \qquad j \in \{1, 2\}. \tag{240}$$

If structural gating is enabled, each view precomputes its own node structural features, edge-pair features, and graph-level summary; the negative feature view shares the base graph structure, but its node features are shuffled or replaced.

M.2 Graph-level Contrastive Objective

The encoder outputs four sets of node representations:

$$H_0 = f_\theta(X_0, \mathcal{G}_0), \quad H_1 = f_\theta(X_1, \mathcal{G}_1), \quad H_2 = f_\theta(X_2, \mathcal{G}_2), \quad H^- = f_\theta(X^-, \mathcal{G}_0). \tag{241}$$

The graph readout of each augmented view is

$$c_j = \text{sigmoid}\left(\frac{1}{|\mathcal{V}_j|}\sum_{v\in\mathcal{V}_j} H_{j,v}\right), \qquad j \in \{1,2\}. \tag{242}$$

The bilinear discriminator

$$D(c,h) = h^\top W_D c \tag{243}$$

determines whether the node representation comes from the same graph semantics. The pretraining loss is

$$\mathcal{L}_{\text{GCL}} = \frac{1}{2}\sum_{j=1}^{2}\left[\text{BCE}\left(D(c_j,H_0),\mathbf{1}\right) + \text{BCE}\left(D(c_j,H^-),\mathbf{0}\right)\right]. \tag{244}$$

When edge-level gating is enabled, traditional static structural prompts are neutralized into identity mappings to avoid scale confusion caused by two sets of structural modulations acting simultaneously; the structural bias is mainly carried by the target edge's $\mathbf{g}_{uv}^{(l)}$.

M.3 Feature Alignment Layer

When the input dimensions produced by different graphs or different text encoders are consistent but their distributions have large shifts, the feature alignment layer can be enabled:

$$\text{Align}(x) = \text{Dropout}\left(\text{LayerNorm}\left(\text{PReLU}(W_a x + b_a)\right)\right). \tag{245}$$

W_a is initialized as the identity matrix, and b_a is initialized as zero. Therefore, this layer is initially an approximately identity transformation; after training, it absorbs inter-graph feature-scale differences without changing the core structure of the graph propagator.

N Supervised Fine-tuning Objective

N.1 Entity-level Supervision

For each question q_b, the data provide a supporting-entity mask $y_{b,e} \in \{0,1\}$. The model outputs entity logits $a_{b,e}$. The weighted BCE is defined as

$$\mathcal{L}_{\text{bce}} = \frac{1}{B}\sum_{b=1}^{B}\frac{\sum_e w_{b,e}\,\text{BCEWithLogits}(a_{b,e},y_{b,e})}{\sum_e w_{b,e}+\epsilon}. \tag{246}$$

Positive weights are uniformly normalized within the positive set; if the adversarial temperature T_a is enabled for negative weights, they are computed by applying softmax to the current model scores:

$$w^-_{b,e} = \frac{\exp(a_{b,e}/T_a)}{\sum_{v:y_{b,v}=0}\exp(a_{b,v}/T_a)}, \qquad y_{b,e}=0. \tag{247}$$

If $T_a = 0$, the negative weights degenerate into a uniform distribution. This design makes training focus more on high-scoring hard negatives, rather than being dominated by a large number of obviously irrelevant entities.

N.2 Multi-positive List Cross-Entropy

Using only BCE treats each entity as an independent binary classification problem, lacking the constraint that "supporting entities should collectively rank near the top of the same candidate list". To this end, we introduce a multi-positive list loss. Let

$$p_{b,e} = \frac{\text{sigmoid}(a_{b,e})}{\sum_v \text{sigmoid}(a_{b,v})+\epsilon}. \tag{248}$$

If sample b has at least one supporting entity, the list loss is

$$\mathcal{L}_{\text{list}} = -\frac{1}{|\mathcal{B}_+|}\sum_{b\in\mathcal{B}_+}\frac{1}{|Y_E(q_b)|}\sum_{e\in Y_E(q_b)}\log(p_{b,e}+\epsilon). \tag{249}$$

Samples with empty supporting-entity sets are skipped. The final entity fine-tuning objective is

$$\mathcal{L}_{\text{ent}} = \lambda_{\text{bce}}\mathcal{L}_{\text{bce}} + \lambda_{\text{list}}\mathcal{L}_{\text{list}}, \qquad (\lambda_{\text{bce}},\lambda_{\text{list}}) = (0.3,0.7). \tag{250}$$

N.3 Optional Document-level Supervision

If the training configuration provides a document-level loss, entity logits are first projected into document logits:

$$\tilde{S}_b = a_b^\top \mathbf{M}, \tag{251}$$

and then the same type of BCE or list loss is computed with the supporting-document mask $z_{b,i}$. This term is suitable for tasks where entity annotations are noisy but the document support set is reliable; if it is not enabled, training is entirely driven by the entity-level support set, and document ranking is obtained through projection only during inference or validation.

O Memory Writer Implementation Details

O.1 The Markov Decision Process for Multi-turn Graph Construction

Specifically, the training of our graph constructor is implemented through VeRL's multi-turn GRPO loop. The state machine of the interactor can be abstracted as a finite-horizon MDP:

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, \mathcal{R}, \rho_0, H). \tag{252}$$

Given a sample x, at round t, the state can be written as

$$s_t = (q, \mathcal{G}_t, \mathcal{D}_t^{\text{proc}}, \mathcal{D}_t^{\text{rem}}, \zeta_t), \tag{253}$$

where $\mathcal{G}_t$ is the current partially written graph, $\mathcal{D}_t^{\text{proc}}$ and $\mathcal{D}_t^{\text{rem}}$ denote the processed and remaining documents, respectively, and ζ_t is an interaction control flag, such as whether the process is still in the graph-construction stage or has already switched to the RAG stage. The action is generated by the language model in JSON format:

$$a_t \sim \pi_\theta(\cdot \mid s_t), \tag{254}$$

and is restricted to two types of legal actions:

1. **Triple action**: output a JSON array, where each element is of the form $\{\texttt{subject}, \texttt{relation}, \texttt{object}\}$, representing the set of facts $\mathcal{T}_t$ written in the current round;
2. **Termination action**: after graph construction is completed, output a JSON object carrying the terminal fields required by the reader side, such as `answer`, `recall`, `precision`, `deducible`, and so on.

In implementation, the environment first checks whether the action can be parsed by `json_repair`, and strictly cleans the triples: items with missing keys, empty strings, or non-dictionary entries are all removed. If illegal JSON is output during the graph-construction stage, the interaction terminates immediately and returns zero reward; if legal triples are output, the environment proceeds to the next round and returns a round-level format reward. The corresponding environment transition can be written as

$$s_{t+1} = P(s_t, a_t) = \begin{cases} (q, \mathcal{G}_t \oplus \mathcal{T}_t, \mathcal{D}_t^{\text{proc}} \cup \{d_t\}, \mathcal{D}_t^{\text{rem}} \setminus \{d_t\}, \zeta_{t+1}), & a_t \text{ is legal ,} \\ (q, \mathcal{G}_t, \mathcal{D}_t^{\text{proc}}, \mathcal{D}_t^{\text{rem}}, \text{STOP}), & a_t \text{ is illegal,} \\ (q, \mathcal{G}_t, \mathcal{D}_t^{\text{proc}}, \mathcal{D}_t^{\text{rem}}, \text{RAG}), & a_t \text{ triggers the reading stage.} \end{cases} \tag{255}$$

Iterative and non-iterative writing. Two strategies are supported. In the non-iterative mode, the model reads the entire context $\mathcal{D}$ at once and outputs all triples. In the iterative mode, the environment reads the documents segment by segment in document order, and in each round the model is only allowed to write triples for the current document. After all documents have been processed, the environment then switches to the RAG stage. If $\mathcal{T}_i$ denotes the set of triples output for document d_i, then the final graph constructed in the iterative mode is $\mathcal{G} = \bigoplus_{i=1}^{m} \mathcal{T}_i$, where $\oplus$ denotes edge-set union and node deduplication. We adopt the iterative strategy by default, because it decomposes the long-context problem into a sequence of local writing decisions, significantly reducing the difficulty of performing global planning in advance. At the same time, it also allows the source document of each triple to be precisely recorded, providing explicit source edges for subsequent text-graph retrieval.

Constructing text-graph memory from output triples. To enable the frozen retriever to operate under the **graph-guided text retrieval** setting, the environment does not directly pass the raw triple strings to the retriever. Instead, it first constructs a text graph with document nodes:

$$\mathcal{G} = (\mathcal{V}_e \cup \mathcal{V}_d, \mathcal{E}_{ee} \cup \mathcal{E}_{ed}), \tag{256}$$

where the entity node set $\mathcal{V}_e$ comes from the subjects and objects in the triples, and the document node set $\mathcal{V}_d = \{d_1, \ldots, d_m\}$ corresponds to the original documents in the context. The entity-entity edges are defined as

$$\mathcal{E}_{ee} = \{(u, r, v) \mid (u, r, v) \in \mathcal{T}\}, \tag{257}$$

and the entity-document source edges are defined as

$$\mathcal{E}_{ed} = \{(u, \texttt{source}, d_i), (v, \texttt{source}, d_i) \mid (u, r, v) \in \mathcal{T}_i\}. \tag{258}$$

In the iterative mode, the source edges are explicit, because the environment already knows that the triples in each round come from the current document. In the non-iterative mode, we use a heuristic alignment method based on tokenizer token overlap to map each triple to the most similar document. The significance of this design is that, after separating writing from reading, the graph constructor is only responsible for deciding "what to write into memory"; as for how the reader aggregates entities on the graph and retrieves documents, this is entirely determined by the frozen f_ϕ.

Frozen GFM retrieval environment. When training the graph constructor, the reader f_ϕ is fixed as the already trained GFM retriever. Let the entity set be $\mathcal{V}_e = \{e_1, \ldots, e_n\}$ and the document set be $\mathcal{V}_d = \{d_1, \ldots, d_M\}$. We then construct:

1. the relation-edge index $\mathbf{E}$ with both forward and reverse directions, together with the relation types $\mathbf{r}$;
2. the sparse entity-document matrix $\mathbf{M} \in \{0, 1\}^{n \times M}$, where $M_{ij} = 1$ if entity e_i appears in document d_j;
3. the question-related entity mask $\mathbf{m}_q \in \{0, 1\}^n$, which is obtained preferentially through lexical matching with the question; if lexical matching fails, it falls back to a heuristic seed set ranked by entity degree.

After encoding the question as a vector $\mathbf{q}$ and the relation names as a matrix $\mathbf{R}$, the frozen GFM forward pass computes the entity relevance scores:

$$\mathbf{s}_e = f_\phi(\mathcal{G}, \mathbf{q}, \mathbf{m}_q; \phi) \in \mathbb{R}^n. \tag{259}$$

The entity scores are then projected into document scores. Let $\mathbf{M}_{\text{Top-}k}(\mathbf{s}_e)$ denote the masking operation that retains only the top-K entity scores, and let $\mathbf{w}_{\text{idf}}$ denote the inverse-frequency weights defined according to the document frequency of each entity. The four document-scoring modes can be written uniformly as

$$\tilde{\mathbf{s}}_e = \begin{cases} \mathbf{s}_e, & \texttt{raw}, \\ \mathbf{M}_{\text{Top-}k}(\mathbf{s}_e), & \texttt{topk}, \\ \mathbf{w}_{\text{idf}} \odot \mathbf{s}_e, & \texttt{idf}, \\ \mathbf{w}_{\text{idf}} \odot \mathbf{M}_{\text{Top-}k}(\mathbf{s}_e), & \texttt{idf_topk}, \end{cases} \tag{260}$$

and the document scores are obtained by

$$\mathbf{s}_d = \mathbf{M}^\top \tilde{\mathbf{s}}_e. \tag{261}$$

We then take Top-$k(\mathbf{s}_d)$ as the retrieval result. In actual use, we also enable `init_entities_weight`, that is, during the GFM forward pass, a $1/f(e)$ weight is applied to high-frequency entities to suppress the dominance of entities connected to too many documents in the retrieval results.

P Additional Detailed Experimental Results

The results of retrieval performance on multi-hop QA benchmarks are in Table 15.

Table 15: Results of retrieval performance on multi-hop QA benchmarks. We report document-level Recall (%) at top-2 and top-5. Best results are in **bold** and runner-ups are underlined. The darker the cell, the better.

Dataset	HotpotQA		MuSiQue		2WikiMultiHopQA		Avg. Rank
Method	R@2	R@5	R@2	R@5	R@2	R@5	
BM25 (▷ *SIGIR'94*)	55.4	72.2	32.3	41.2	51.8	61.9	18.2
Contriever (▷ *TMLR'22*)	57.2	75.5	34.8	46.6	46.6	57.5	17.2
GTR (▷ *EMNLP'22*)	59.4	73.3	37.4	49.1	60.2	67.9	13.8
ColBERTv2 (▷ *NAACL'22*)	64.7	79.3	37.9	49.2	59.2	68.2	11.5
RAPTOR (▷ *ICLR'24*)	58.1	71.2	35.7	45.3	46.3	53.8	17.7
Proposition (▷ *EMNLP'24*)	58.7	71.1	37.6	49.3	56.4	63.1	14.8
GraphRAG (▷ *arXiv'24*)	58.3	76.6	35.4	49.3	61.6	77.3	12.2
G-Retriever (▷ *NeurIPS'24*)	53.3	65.5	38.8	45.1	60.8	67.8	15.7
LightRAG (▷ *arXiv'24*)	38.8	54.7	24.8	34.7	45.1	59.1	20.5
HippoRAG (▷ *NeurIPS'24*)	60.1	78.5	41.2	53.2	68.4	87.0	8.7
HippoRAG 2 (▷ *ICML'25*)	80.5	88.1	47.0	56.7	**88.9**	90.1	2.4
SubgraphRAG (▷ *ICLR'25*)	61.5	73.0	42.1	49.3	70.7	85.5	9.7
PropRAG (▷ *EMNLP'25*)	**81.9**	88.0	**47.7**	**57.9**	87.9	90.1	**1.9**
GFM-RAG (▷ *NeurIPS'25*)	75.6	**89.6**	43.5	57.6	79.1	92.4	2.9
FLARE (▷ *EMNLP'23*)	73.1	81.3	44.3	55.1	67.1	73.1	6.5
Adaptive-RAG (▷ *NAACL'24*)	61.0	76.4	35.1	44.7	44.7	61.4	16.6
BM25 + IRCoT (▷ *ACL'23*)	65.6	79.0	34.2	44.7	61.2	75.6	12.4
Contriever + IRCoT (▷ *ACL'23*)	65.9	81.6	39.1	52.2	51.6	63.8	10.7
ColBERTv2 + IRCoT (▷ *ACL'23*)	67.9	82.0	41.7	53.7	64.1	74.4	7.2
HippoRAG + IRCoT (▷ *ACL'23*)	67.0	83.0	45.3	57.6	75.8	**93.9**	3.6
SAGE (ours)	65.1	77.6	43.2	53.1	83.6	88.6	7.0

Table 16: Performance of representative baselines on the original AmazonQA full-test protocol. BLEU-1/2/3/4 are denoted as B-1/2/3/4, and R denotes ROUGE. Best results are in **bold** and runner-ups are underlined. **Only rows marked with 0-shot are our zero-shot transfer results; baseline rows and trained variants are not marked as zero-shot.**

Zero-shot setting applies only to Ours rows marked with 0-shot on AmazonQA.

Method	B-1	B-2	B-3	B-4	R
Heuristic baselines from the original AmazonQA protocol					
Random Sentence (▷ *IJCAI'19*)	78.56	63.95	44.37	29.87	49.12
Top-1 using IR (▷ *IJCAI'19*)	89.49	74.80	56.76	43.52	61.48
Top-1 Using BLEU (▷ *IJCAI'19*)	**92.74**	**78.43**	**60.91**	**48.08**	**62.68**
Top-1 Helpfulness (▷ *IJCAI'19*)	20.66	19.78	16.39	12.54	37.01
Top-1 Wilson Score (▷ *IJCAI'19*)	20.74	19.84	16.44	12.58	37.26
Neural baseline from the original AmazonQA protocol					
R-Net (▷ *IJCAI'19*)	47.04	40.32	31.48	23.92	40.22
Human answers under the original AmazonQA protocol					
Amazon User Community (▷ *IJCAI'19*)	80.88	68.86	54.36	42.01	62.18
Expert (Spans) (▷ *IJCAI'19*)	68.33	57.79	44.61	34.43	51.09
Expert (Descriptive) (▷ *IJCAI'19*)	53.67	46.56	37.81	30.76	53.31
Our method					
Ours (0-shot)[0-shot]	61.84[0-shot]	49.36[0-shot]	37.82[0-shot]	28.41[0-shot]	46.73[0-shot]
Ours (trained)	74.92	61.58	47.63	35.86	54.92
Ours +1 round	82.76	68.91	52.74	39.68	58.83

The results on AmazonQA are in Table 16.

The HaluMem results are shown in Table 17.

[1] Supermemory: https://github.com/supermemoryai/supermemory.
[2] Memobase: https://github.com/memodb-io/memobase.

Table 17: Results on HaluMem-Medium. We report memory extraction, memory updating, and memory question-answering metrics. R denotes Recall, W-R denotes Weighted Recall, T-P denotes Target Memory Precision, Acc. denotes Memory Accuracy, FMR denotes False Memory Resistance, F1 denotes Memory Extraction F1-score, C denotes Correct Rate, H denotes Hallucination Rate, and O denotes Omission Rate. For R, W-R, T-P, Acc., FMR, F1, and C, higher is better; for H and O, lower is better. Best numerical results are in **bold**, and runner-ups are underlined; ties at the runner-up value are all underlined. Missing entries are denoted by "–".

Dataset	Memory Extraction						Memory Updating			Memory QA		
Method	R↑	W-R↑	T-P↑	Acc.↑	FMR↑	F1↑	C↑	H↓	O↓	C↑	H↓	O↓
Other memory-system baselines												
Supermemory[1]	41.53	64.76	90.32	60.83	51.77	56.90	16.37	1.15	82.47	54.07	22.24	23.69
Mem0 (▷ *arXiv'25*)	42.91	65.03	86.26	60.86	56.80	57.31	25.50	0.45	74.02	53.02	19.17	27.81
Zep (▷ *arXiv'25*)	–	–	–	–	–	–	47.28	0.42	52.31	55.47	21.92	22.62
*Matched baseline–*SAGE *comparisons*												
MemOS (▷ *arXiv'25*)	74.07	84.81	86.25	59.55	44.94	79.70	62.11	0.42	37.48	67.23	15.17	17.59
MemOS + SAGE	78.46	87.86	89.66	63.88	73.26	83.69	70.25	0.42	29.33	69.65	13.07	**17.27**
MemOS + SAGE (+1 round)	**80.31**	**87.93**	91.02	66.41	74.09	**85.33**	**71.80**	**0.39**	**27.81**	**70.61**	**11.48**	17.91
Memobase[2]	14.55	25.88	**92.24**	32.29	80.78	25.13	5.20	0.55	94.25	35.33	29.97	34.71
Memobase + SAGE	54.79	74.75	90.12	66.72	82.73	68.15	34.16	0.51	65.33	50.43	21.97	27.60
Memobase + SAGE (+1 round)	58.37	75.30	89.85	**68.19**	**84.23**	70.77	33.28	0.46	66.26	55.84	20.40	23.76

Table 18: Path interpretations for multi-hop reasoning, where r^{-1} denotes the inverse of original relation.

Question	Which man who presented the *Australia 2022 FIFA World Cup bid* was born on *October 22, 1930*?
Answer	Frank Lowy
Sup. Doc.	["Frank Lowy", "Australia 2022 FIFA World Cup bid"]
Paths	1: (the bid for the 2022 fifa world cup, was one of the representatives of^{-1}, frank lowy) → (frank lowy, equivalent, sir frank p lowy) → (sir frank p lowy, was born on, 22 october 1930) 2: (22 october 1930, was born on^{-1}, sir frank p lowy) → (sir frank p lowy, equivalent, frank lowy) → (frank lowy, was one of the representatives of, the bid for the 2022 fifa world cup)

P.1 Path Interpretations

We provide path interpretations of SAGE for multi-hop reasoning in Table 18. The importance of each path to the final prediction can be measured by the partial derivative of the prediction score with respect to the triples at each reasoning layer. The top-k path interpretations are then obtained by selecting the top-k longest paths with beam search.

As shown in Table 18, SAGE successfully identifies the answer by connecting two key constraints in the question: the person who presented the Australia 2022 FIFA World Cup bid and the person born on October 22, 1930. Specifically, the first path starts from the entity "the bid for the 2022 FIFA World Cup" and follows the inverse relation of "was one of the representatives of" to reach "Frank Lowy". Then, through an entity-equivalence relation, it links "Frank Lowy" to "Sir Frank P. Lowy", whose birth date is "22 October 1930". The second path verifies the reasoning in the reverse direction by starting from the birth date and tracing back to the representative of the World Cup bid. These paths demonstrate that SAGE can effectively align different surface forms of the same entity and integrate multiple question constraints within a single-step retrieval process, showing its ability to perform interpretable multi-hop reasoning.

Q Dataset Details

Table 19 summarizes the details of each dataset.

General and Multi-hop QA. We first evaluate SAGE on a set of general open-domain and multi-hop QA benchmarks that stress different aspects of retrieval-augmented reasoning. NQ-Open is derived from Natural Questions and is widely used as a standard open-domain short-answer QA

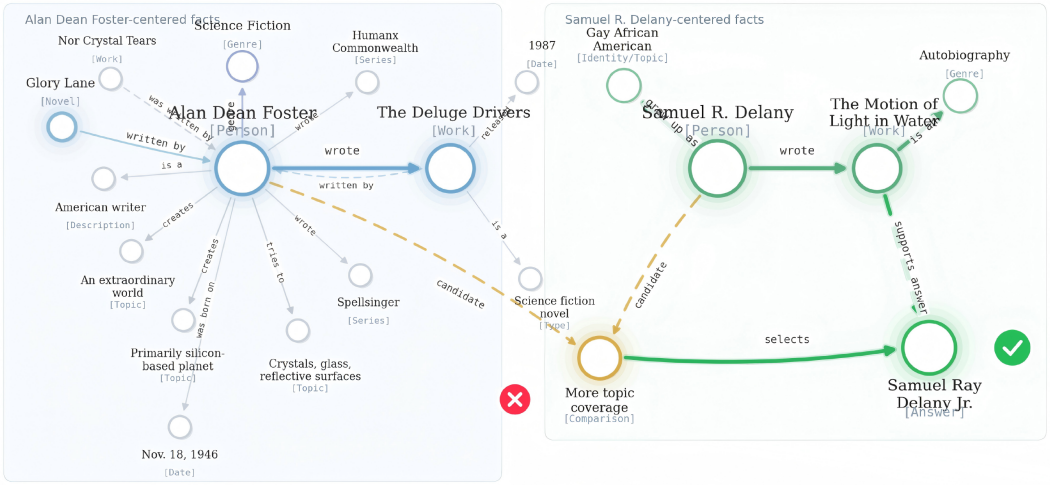


Figure 7: Visualization of the retrieved results.

Table 19: Dataset statistics and evaluation scenarios. We evaluate SAGE on three complementary categories: general and multi-hop QA, practical e-commerce review QA, and long-term agent memory. "Train/Dev/Test" denotes the standard split when available. For benchmark-only datasets without a conventional supervised training split, we report the total number of evaluation instances or benchmark scale.

Category	Dataset	Scale / Split	Evidence Source	Task Type	Key Capabilities	Main Metrics
General / Multi-hop QA	NQ-Open (▷)	79,168 / 8,757 / 3,610	English Wikipedia	Open-domain short-answer QA	Factual retrieval; entity-level knowledge access; open-domain answer generation	EM / F1 / Acc.; Recall@k
	PopQA (▷)	14,267 QA pairs	Wikidata triples + Wikipedia page-view popularity	Entity-centric open-domain QA	Long-tail factual recall; parametric vs. non-parametric memory; retrieval under entity popularity shift	Acc. / EM; long-tail breakdown
	HotpotQA (▷)	90,447 / 7,405 / 7,405	Wikipedia paragraphs; 10-paragraph distractor setting	Explainable 2-hop QA	Bridge-entity recovery; comparison reasoning; sentence-level supporting facts	Answer EM/F1; Support EM/F1; Joint EM/F1
	2WikiMultiHopQA (▷)	167,454 / 12,576 / 12,576	Wikipedia + Wikidata; 10 passages per instance	2–4 hop multi-hop QA	Reasoning-path recovery; comparison, bridge, and bridge-comparison reasoning	Answer EM/F1; Evidence / path recall
	MuSiQue (▷)	19,938 / 2,417 / 2,459 (24,814 total)	Composed single-hop QA over textual passages	2–4 hop connected multi-hop QA	Connected reasoning; shortcut-resistant evidence aggregation; multi-hop compositionality	Answer EM/F1; Support / evidence recall
E-commerce Review QA	AmazonQA (▷)	923K questions; 3.6M answers; 14M reviews; 156K products	Amazon product reviews, questions, answers, and product metadata	Review-based QA with answerability annotation	Noisy review retrieval; answerable / unanswerable detection; evidence synthesis from user-generated reviews	BLEU / ROUGE; answerability Acc./F1; groundedness
Long-term Agent Memory	LongMemEval (▷)	500 eval. instances per file; S: ∼115K tokens / 30–40 sessions; M: ∼1.5M tokens / ∼500 sessions; Oracle: evidence sessions only	Long multi-session human–AI chat histories	Long-term interactive memory QA	Information extraction; multi-session reasoning; temporal reasoning; knowledge update; abstention	Overall Acc.; category-wise Acc.; context tokens; latency
	HaluMem (▷)	Medium: 20 users, 30,073 dialogue rounds, ∼160K tokens/user, 14,948 memory points, 3,467 QA pairs; Long: 53,516 rounds, ∼1M tokens/user	Synthetic long-term human–AI interaction histories with memory points and multi-type questions	Operation-level memory hallucination benchmark	Memory extraction; memory updating; memory QA; hallucination, omission, and conflict propagation across memory operations	Extraction R/P/F1; Updating C/H/O; QA C/H/O

benchmark; it evaluates whether a system can retrieve and ground factual answers from a large Wikipedia-scale corpus. PopQA complements NQ by focusing on entity-centric factual questions whose subjects span different popularity levels, making it particularly useful for testing whether a retrieval or memory system can recover long-tail factual knowledge rather than relying only on parametric memorization. HotpotQA contains Wikipedia-based multi-hop questions with sentence-level supporting facts, allowing us to evaluate not only answer correctness but also whether the system can recover bridge evidence and produce interpretable reasoning chains. 2WikiMultiHopQA further stresses structured multi-hop reasoning by combining Wikipedia text with Wikidata-derived relations and providing evidence paths for 2–4 hop questions. Finally, MuSiQue is designed to reduce shortcut reasoning by composing connected single-hop questions into 2–4 hop questions,

making it a strong testbed for evaluating whether SAGE can retrieve and integrate multiple pieces of evidence in a genuinely compositional manner.

E-commerce Review-based QA. We use `AmazonQA` to evaluate SAGE in a practical, noisy, user-generated e-commerce setting. Unlike Wikipedia-style QA benchmarks, `AmazonQA` consists of real product questions, community answers, product reviews, and product metadata, and includes answerability annotations indicating whether a question can be answered from available reviews. This makes it a suitable benchmark for testing whether a memory system can identify useful evidence from noisy review collections, distinguish answerable from unanswerable questions, and synthesize grounded answers from multiple user-generated snippets. From the perspective of self-evolving memory, AmazonQA is especially valuable because the system must learn which review facts, product attributes, and user opinions are worth indexing for future retrieval, rather than simply matching a question to a clean encyclopedic passage.

Long-term Agent Memory. To move beyond conventional RAG evaluation, we further evaluate SAGE on long-term agent memory benchmarks. `LongMemEval` is designed to assess the long-term memory abilities of chat assistants over extended multi-session interaction histories. It covers five core memory abilities: information extraction, multi-session reasoning, temporal reasoning, knowledge updates, and abstention. This benchmark directly tests whether SAGE can retrieve sparse but relevant memory traces from long histories, combine evidence across sessions, respect temporal order, and update previously stored information when new interactions supersede old memories. We use `HaluMem` as a complementary benchmark for evaluating hallucination in memory systems. Rather than only measuring end-to-end QA accuracy, HaluMem decomposes memory evaluation into memory extraction, memory updating, and memory question answering, thereby revealing at which operational stage hallucinations, omissions, or conflicts arise. This is particularly important for our setting because errors introduced during graph construction or memory updating may propagate to graph-guided retrieval and final answer generation.

Evaluation Rationale. Together, these datasets form a progressively broader evaluation suite. NQ and PopQA test factual open-domain retrieval; HotpotQA, 2WikiMultiHopQA, and MuSiQue test multi-hop evidence composition; AmazonQA evaluates noisy real-world review memory in an e-commerce domain; LongMemEval tests long-horizon interactive memory; and HaluMem diagnoses operation-level hallucinations in memory systems. This combination allows us to evaluate SAGE not merely as a retrieval-augmented QA pipeline, but as a self-evolving memory system that must decide what to store, how to organize stored information, how to retrieve it under different query conditions, and how to update or suppress unreliable memories over time.

R Baselines and Metrics

Baselines. We evaluate SAGE against state-of-the-art baselines, including their combined variants, which are grouped into **four** categories:

- ***Base LLM:*** `GPT-4o-mini` Hurst et al. (2024).
- ***Single-step RAGs:*** including `BM25` Robertson & Walker (1994), `Contriever` Izacard et al. (2021), `GTR` Ni et al. (2022), `ColBERTv2` Santhanam et al. (2022), `RAPTOR` Sarthi et al. (2024), and `Proposition` Chen et al. (2024).
- ***Graph-enhanced RAGs:*** including `GraphRAG` Edge et al. (2024), `G-Retriever` He et al. (2024), `LightRAG` Guo et al. (2024), `HippoRAG` Gutiérrez et al. (2024), `HippoRAG 2` Gutiérrez et al. (2025), `SubgraphRAG` Li et al. (2024), `PropRAG` Wang & Han (2025), and the closely related `GFM-RAG` Luo et al. (2025).
- ***Multi-step RAGs:*** `IRCoT` Trivedi et al. (2023), `FLARE` Jiang et al. (2023), and `Adaptive-RAG` Jeong et al. (2024).

In particular, `IRCoT` Trivedi et al. (2023) is a general multi-step reasoning framework that can be integrated with non-iterative retrievers, allowing both single-step RAG and graph-based methods to conduct multi-hop reasoning through interleaved retrieval and generation.

Table 20: Descriptive latency measurements. All entries are milliseconds. Retrieve-plus-answer excludes the offline evaluator.

Host/task	Stage	Native	SAGE
Memobase HaluMem	Retrieval mean	253.1	297.9
	Retrieval p95	393.5	523.2
Memobase LongMemEval	Retrieval mean	221.8	314.5
	Retrieval p95	287.1	376.0
	Retrieve+answer mean	2,192.3	2,795.7
MemOS LongMemEval	Search mean	–	70,385.9
	Search+answer mean	–	72,110.8

Metrics. To evaluate retrieval quality, we report Recall@2 and Recall@5 for both retrieved entities and documents, denoted as $R@2/5_E$ and $R@2/5_D$, respectively. For end-to-end QA evaluation, we use standard metrics, including Exact Match (EM), F1 score, Precision (P), and Recall (R), in the main experiments to comprehensively measure answer correctness and coverage.

S Integrate SAGE into Multiple Memory Systems

S.1 Host implementations

Zep. The native context budget is 16,384 tokens. Extraction uses GPT-4o-mini; answer generation and judging use GPT-4o with temperature zero.

Memobase. We use Memobase v0.0.42. Both arms share ingestion, durable memory state, profiles, queries, reference times, local Qwen3-Embedding-8B embeddings, and answer configuration. Extraction and summarization use GPT-4o-mini; answers and judges request GPT-4o at temperature zero.

MemOS. Native embeddings and the SAGE branch use all-mpnet-base-v2. The graph index is prebuilt: normal MemOS writes do not synchronize new memories into that index online. LongMemEval answers and judges use GPT-4o-mini; HaluMem answers and judges use GPT-4o. The MemReader extraction temperature is 0.6, distinct from the zero-temperature answer/judge settings.

S.2 Deployment cost

T Matched Retrieval Protocol for Restricted-Corpus Transfer

T.1 Evaluation unit and restricted corpora

The NQ and PopQA measurements use 1,000 queries each and restricted document corpora. NQ contains 964 documents and 17,035 passages after 256-token segmentation; PopQA contains 1,793 documents and 8,722 passages.

Each retriever produces ten documents. The retrieved documents are segmented into passages and reranked using the same MPNet passage reranker. Five passages are supplied to the common answer generator. All 1,000 queries in each dataset were evaluated without dropping failed or incomplete cases in the reported matched rerun.

T.2 Retrieval metric

Passage R@k is a question-level hit rate: a query succeeds if at least one selected passage contains a normalized gold-answer string. Gold answers are used only for scoring, not for retrieval or reranking. This metric differs from fractional recall of annotated support documents. Retrieving a support page does not ensure that an answer-bearing passage appears in the final context. Conversely, a valid answer-bearing passage may come from a document that is not one of the annotated support pages.

The current paper uses the passage-level metric consistently and does not compare it numerically with an earlier document-title-based measurement. QA EM/F1 is reported alongside retrieval hit rate to measure the final answer effect under the same context-construction protocol.

T.3 Meaning of no-target-adaptation(0-shot) transfer

The writer and reader were trained on HotpotQA, 2WikiMultiHopQA, and MuSiQue, with no NQ- or PopQA-specific gradient updates, tuning, or evolution rounds for the reported transfer setting. This is target-task/label transfer, not a claim of novel Wikipedia content or complete absence of knowledge in generic pretrained encoders and LLMs.

U Diagnostic Analysis of Memory Failures

This appendix analyzes *where* long-term-memory errors arise rather than treating every incorrect answer as a retrieval failure. The audit is based on the earlier standalone SAGE memory runs and is reported separately from the paired host-integration experiments in Appendix S. The audited sets contain 110 LongMemEval examples and 763 HaluMem-Medium examples. Its purpose is diagnostic: it localizes missing evidence, ranking failures, and post-retrieval reasoning failures, but it is not used as the primary system-level comparison for the host-integration results.

U.1 Diagnostic Protocol and Scope

Sequential failure attribution. For an incorrect question, we assign the first applicable stage in the following order:

1. **Writer omission / source-anchor failure.** At least one annotated gold memory is absent from the stored corpus, or a gold document has no usable entity in the `document2entities` mapping and therefore cannot be anchored into the KG.
2. **Graph reachability failure.** The gold memories are written and anchorable, but at least one required gold document lies outside the two-hop *undirected* KG neighborhood of the exact-query seed entities.
3. **Top-k ranking failure.** All required gold documents are structurally reachable under the above proxy, but at least one required document is absent from the reader's top-10 retrieved context.
4. **Answer-generation failure.** All annotated gold documents are present in the top-10 context, but the final answer is still scored incorrect.

Because the stages are applied sequentially, they are mutually exclusive within each audited question set.

What the diagnostic does *not* establish. The two-hop neighborhood is a structural proxy, not a reconstruction of the actual GFM computation path. A globally initialized or deeper reader can in principle use evidence outside this proxy, and the presence of a two-hop path does not prove that the GFM actually propagated information along that path. Similarly, the source-anchor test detects a missing or unusable anchor; it does not verify that an existing anchor is semantically the correct entity, because the benchmarks do not provide gold entity-level anchoring annotations. Finally, the last stage localizes an error to the post-retrieval portion of the pipeline, but it can include answer selection, temporal arithmetic, aggregation, polarity conversion, or evaluator mismatch rather than a single unique causal mechanism.

U.2 Gold-Memory and Source-Anchor Coverage

Both audited datasets have 100% gold-memory coverage: the annotated evidence is present in the stored corpus for every audited question. The category-specific source-anchor audit is shown in Table 21.

The four anchor-incomplete documents in the HaluMem temporal-cue proxy occur only in negative/unknown questions for which the model correctly returns an unknown answer. They therefore do

Table 21: Gold-document writing and source-anchor coverage in the diagnostic audit. "Usable KG anchor" means that the gold document has at least one entity in the `document2entities` mapping. HaluMem's temporal row is a cue-based proxy rather than an official benchmark category.

Dataset / type	Gold documents	Written in corpus	Usable KG anchor
LongMemEval KU	36	36	36
LongMemEval TR	61	61	61
LongMemEval MS	82	82	82
HaluMem KU	62	62	62
HaluMem temporal-cue proxy	810	810	806
HaluMem MS	93	93	93

Table 22: Overall failure-stage attribution for the 110-question LongMemEval audit.

Outcome / stage	Count	Share
Correct	38	34.55%
Writer/source-anchor	0	0.00%
Graph reachability	27	24.55%
Top-10 ranking	27	24.55%
Answer generation	18	16.36%

not contribute a writer/source-anchor failure among the incorrect temporal-proxy examples. This result should not be read as proof that the writer never creates a semantically wrong anchor; the proxy only tests whether an anchor is available.

U.3 LongMemEval: Overall and Category-Level Attribution

Across all 110 audited LongMemEval questions, 38 are correct, 27 are assigned to graph-reachability failure, 27 to top-10 ranking failure, and 18 to answer-generation failure. No question is assigned to writer omission under the above gold-memory-coverage test. Thus, the errors are distributed across structure, ranking, and post-retrieval reasoning rather than being explained by one single stage.

LongMemEval provides official question types including Knowledge Update (KU), Temporal Reasoning (TR), and Multi-Session Reasoning (MS). These three subsets contain 18, 29, and 32 audited questions, respectively. The category-level decomposition is reported in Table 23. Percentages in the last four columns use the *incorrect examples of that category* as the denominator.

U.3.1 Knowledge Update: Missing the Latest State or Selecting a Stale One

For LongMemEval KU, 7/13 errors (53.85%) are top-10 ranking failures, 4/13 (30.77%) are structural-reachability failures, and 2/13 (15.38%) remain after all annotated evidence has entered the answer context. The dominant pattern is therefore not that the update was never written; rather, the latest state is inaccessible, is ranked below an older state, or is available but not treated as superseding the stale fact.

Case KU-1: latest fact written but structurally unreachable. For `longmemeval_oracle_train_159`, the question is "How many issues of National Geographic have I finished reading?" The gold answer is *five*, whereas the model answers *three issues*. An earlier session saying that three issues had been completed is ranked first. A later session explicitly stating that the fifth issue has been finished is present in memory, but it does not enter the top-10 and lies outside the two-hop neighborhood of the exact-query seed. The failure chain is therefore

latest fact written → latest session structurally unreachable → old session retrieved → stale answer.

Case KU-2: latest fact reachable but ranked out. For `longmemeval_oracle_train_142`, the question asks "How many MCU films did I watch in the last 3 months?" The gold answer is 5 and the model answers 4. The older four-film session is ranked eighth. The later session updating the

Table 23: LongMemEval category-level error attribution. Each entry in the failure columns is count (percentage of the incorrect examples in that row).

Type	Total	Correct	Errors	Writer/anchor	Graph	Top-k	Generation
KU	18	5	13	0 (0.00%)	4 (30.77%)	**7 (53.85%)**	2 (15.38%)
TR	29	7	22	0 (0.00%)	3 (13.64%)	9 (40.91%)	**10 (45.45%)**
MS	32	4	28	0 (0.00%)	**13 (46.43%)**	11 (39.29%)	4 (14.29%)

count to five is written and structurally reachable, but it does not enter the top-10. This isolates a reader-ranking failure from a memory-writing failure.

Case KU-3: both versions retrieved, but the answerer selects the obsolete value. For `longmemeval_oracle_train_137`, the question asks "How many stars do I need to reach the gold level on my Starbucks Rewards app?" The updated value, 120 stars, is ranked first; the older 125-star memory is ranked third. The model nevertheless answers 125. Because the newer fact is both retrieved and ranked above the stale fact, this is a post-retrieval update-resolution failure: the answerer does not use the update relation to invalidate the obsolete value.

U.3.2 Temporal Reasoning: Evidence Omission versus Temporal Computation

LongMemEval TR exhibits a different profile. Among its 22 errors, 9 (40.91%) are ranking failures and 10 (45.45%) occur after all annotated evidence is already in the top-10. A finer audit identifies 12 errors with temporal-evidence omission, 7 with duration/time-difference mistakes, and 3 with ordering or temporal-qualifier mistakes. Therefore, improving retrieval alone cannot resolve the full TR error set.

Case TR-1: both temporal endpoints reachable but omitted from top-10. For `longmemeval_oracle_train_6`, the question is "How many days did it take for me to find a house I loved after starting to work with Rachel?" The required sessions record February 15 as the start of working with Rachel and March 1 as the date on which a liked house was found. Both sessions are structurally reachable, but neither appears in the top-10, so the answerer cannot reconstruct the interval and returns unknown instead of the gold 14 days.

Case TR-2: both dates retrieved, but duration arithmetic is wrong. For `longmemeval_oracle_train_27`, the question asks for the number of days between buying the Adidas running shoes and the Converse shoelaces breaking. The January 10 purchase is ranked first and the January 24 shoelace event is ranked tenth. Both temporal endpoints are therefore available, but the model answers 25 days instead of the correct 14 days. This is a post-retrieval duration-arithmetic failure.

Case TR-3: both dates retrieved, but event ordering is reversed. For `longmemeval_oracle_train_32`, the question is "Which device did I set up first, the smart thermostat or the new router?" The router event is dated January 15 and ranked first; the thermostat event is dated February 10 and ranked third. Despite retrieving both dates, the model answers that the thermostat came first. The evidence acquisition succeeds, but the answerer mishandles the temporal ordering implied by *first*.

U.3.3 Multi-Session Reasoning: Incomplete Support Coverage and Aggregation

MS is the LongMemEval category most strongly associated with graph/evidence coverage. Graph reachability and top-10 ranking jointly account for 24/28 errors (85.72%). The main retrieval-side problem is that evidence distributed across several sessions is not simultaneously available; when all sessions are available, aggregation, category boundaries, and arithmetic can still fail.

Case MS-1: only part of a distributed weekly pattern is retrieved. For `longmemeval_oracle_train_94`, the question is "How many fitness classes do I attend in a typical week?" The gold answer is 5. The required evidence is distributed over several

Table 24: HaluMem structural-coverage diagnostics.

Metric	Result
Gold memory coverage	100.00%
All gold two-hop reachable	98.30%
At least one gold in top-10	21.76%
All gold in top-10	6.42%
Top-k failure conditioned on reachability	93.47%

Table 25: HaluMem category-level diagnostic attribution. Percentages in the failure columns use incorrect examples in the corresponding row as the denominator. The temporal-cue proxy overlaps other categories. The last row is a manual evaluator-sensitivity view rather than a replacement of the automated score.

Type	Total	Correct	Errors	Writer/anchor	Graph	Top-k	Generation
KU	43	4	39	0 (0.00%)	0 (0.00%)	**36 (92.31%)**	3 (7.69%)
Temporal-cue proxy	387	94	293	0 (0.00%)	4 (1.37%)	**271 (92.49%)**	18 (6.14%)
MS, automated	39	0	39	0 (0.00%)	0 (0.00%)	**38 (97.44%)**	1 (2.56%)
MS, manual sensitivity	39	1	38	0 (0.00%)	0 (0.00%)	**38 (100.00%)**	0 (0.00%)

sessions, including Monday BodyPump, Saturday Hip Hop Abs, Sunday yoga, and other weekly classes. Only the Monday BodyPump session enters the top-10; one key document lies outside the two-hop proxy and other required sessions are also absent from the reader context. Because "a typical week" is not itself a single explicit entity, the reader does not recover a stable bridge that connects all of the distributed supporting sessions.

Case MS-2: all annotated evidence retrieved, but aggregation is incorrect. For `longmemeval_oracle_train_114`, the answer requires combining 17 poems (rank 7), 5 short stories (rank 5), and one writing-challenge piece (rank 1). All supporting evidence is present, so the correct total is $17 + 5 + 1 = 23$. The model answers 24 and incorrectly expands the number of short stories. This case isolates multi-session aggregation, category boundaries, and arithmetic from retrieval coverage.

U.4 HaluMem: Category-Level Attribution

HaluMem-Medium shows a more concentrated retrieval bottleneck than LongMemEval. Across the 763-example audit, gold-memory coverage is 100%. Moreover, 98.30% of questions have all annotated gold memories inside the two-hop structural proxy. In contrast, only 21.76% have at least one gold memory in the reader top-10, and only 6.42% have *all* gold memories in the top-10. Conditioned on structural reachability, the resulting top-k ranking-failure rate is 93.47%.

HaluMem provides an official *Dynamic Update* category, which we denote KU, and an official *Multi-hop Inference* category, which we denote MS. It does *not* provide an official TR category. For diagnostic purposes only, we construct a temporal-cue proxy consisting of questions containing expressions such as *latest*, *current*, *as of*, *before*, *after*, *by*, or *how long*. This proxy contains 387 examples and can overlap both KU and MS; consequently, the rows of Table 25 must not be summed.

U.4.1 Knowledge Update: Stale Memories Outrank the Latest State

HaluMem KU is overwhelmingly a top-10 coverage problem: 36/39 errors (92.31%) occur after the relevant memories are written and structurally reachable but before they enter the reader context. In 6/43 KU questions, the retrieved context also contains a detectable stale value, illustrating the competition between old and updated states.

Case KU-4: the latest health update is omitted while an older state is retrieved. For `halumem_medium_train_271`, the question asks "What is the latest health status of Johnson Joseph as of December 20, 2035?" The gold answer is *Poor*, whereas the model answers *Highly*

Positive. One gold memory states that Johnson Joseph's physical health status changed from "Excellent" to "Poor", and another supporting memory records chronic fatigue. Both are written and structurally reachable, but neither enters the top-10. Rank 4 instead contains an older update from "Good" to "Excellent", while higher-ranked context also contains unrelated records about other people and topics. This is a concrete stale-state ranking failure:

$$\text{latest state exists} \rightarrow \text{latest evidence ranked out} \rightarrow \text{stale state retrieved} \rightarrow \text{outdated answer}.$$

Case KU-5: the current update is retrieved but the answerer still abstains. For `halumem_medium_train_308`, the question asks "What is Johnson Joseph's current monthly income as of July 18, 2039?" The gold answer is $8,000, and rank 2 explicitly states that the monthly income was updated from $7,000 to $8,000. The model nevertheless answers unknown. This is an answer-generation omission rather than a retrieval failure.

U.4.2 Temporal-Cue Proxy: Time-Conditioned Retrieval and Polarity

The temporal-cue proxy contains 293 incorrect examples. Of these, 271 (92.49%) are top-10 ranking failures, 4 (1.37%) fail the structural reachability proxy, and 18 (6.14%) remain incorrect after annotated evidence is in the reader context. Because this subset is cue-derived and overlaps other HaluMem categories, these numbers are used only for diagnostic behavior, not as an official HaluMem TR score.

Case TR-4: all memories at the relevant date are reachable but ranked out. For `halumem_medium_train_113`, the question asks "What kind of pet ownership practices might Martin adopt after March 23, 2032?" The gold answer is conservation-focused and ethical practices. Three gold memories at the specified date mention conservation and ethical considerations, conservation-focused pet ownership, and responsible ownership and conservation efforts. All three are structurally reachable, but none enters the top-10; the model consequently reports insufficient information. The failure is time-conditioned reader coverage rather than missing timestamps in memory.

Case TR-5: temporal/causal evidence is retrieved but the polarity conversion fails. For `halumem_medium_train_550`, the question asks "Did Sarah decide to continue consuming sugary sodas after the seminar on Aug 13, 2027?" A rank-3 memory explicitly states that the seminar led to a decision to "avoid sugary sodas entirely." The model still answers unknown. Here the evidence is available, but the answerer fails to combine the temporal condition *after the seminar*, the causal relation between the seminar and the decision, and the polarity conversion

$$\text{avoid entirely} \Rightarrow \text{did not continue}.$$

U.4.3 Multi-Session Reasoning: Complete Support Coverage Is the Main Bottleneck

The official HaluMem Multi-hop Inference subset contains 39 examples. Under the automated diagnostic, 38 are top-10 ranking failures and one is labeled as an answer-generation failure. Manual inspection of that one generation case, `halumem_medium_train_2394`, shows that the model answer contains the annotated advice "expanding the studio's creative team", but the longer response falls below the local token-overlap threshold. Moving only this case to correct yields the sensitivity row in Table 25: all 38 remaining errors are top-10 coverage failures. We keep the automated result as the primary record and use the manual row only to expose evaluator sensitivity.

Case MS-3: the motivation is retrieved but the concrete strategy is missing. For `halumem_medium_train_207`, the question asks "How did Johnson's exploratory motivation influence his stress management strategies?" The answer requires combining three memories: (i) his desire to explore new skills through a part-time role; (ii) the exploratory motivation that led him to test new approaches; and (iii) the concrete stress-management strategies of mindfulness and scheduled downtime. The reader retrieves only the second memory (rank 4); the first and third do not enter the top-10. The system therefore recovers the motivation but not the endpoint that the question asks for.

Case MS-4: all supporting memories are omitted and the model falls back to a plausible generic statement. For `halumem_medium_train_180`, the question asks "What role did community support play in Johnson Joseph's career progress?" The gold answer is "Increased guest traffic and referrals." Three supporting memories describe local ties producing referrals and guest numbers, a new partnership increasing guest traffic, and Johnson's appreciation of community support. All three are structurally reachable but none appears in the top-10. The model answers that community support provided collaborative opportunities that improved his career prospects. The response is plausible in general, but it does not recover the specific effect supported by memory. This example illustrates how incomplete evidence coverage can be followed by a fluent but underspecified answer based on generic prior knowledge.

U.5 Cross-Category Interpretation

The audits reveal different dominant failure mechanisms across memory abilities.

Knowledge Update. The latest memory is generally written, but the updated session can be structurally inaccessible or ranked below an older state. Even when both versions enter the context, the answerer can still select the stale value or ignore qualifiers such as *latest*, *current*, and *as of*. This indicates that robust long-term memory requires not only retrieval coverage but also explicit handling of version/state validity.

Temporal Reasoning. TR contains both retrieval-side and reasoning-side errors. Retrieval can omit one or both temporal endpoints, while successful retrieval can still be followed by incorrect duration arithmetic, relative-time normalization, event ordering, or handling of qualifiers such as *first*, *before*, and *after*. Therefore, increasing top-k coverage alone is insufficient for complete temporal-memory reasoning.

Multi-Session Reasoning. MS is most sensitive to simultaneous support coverage. Relevant facts are distributed across sessions, and a missing bridge or a single ranked-out support can break the complete evidence set. LongMemEval additionally shows that even complete retrieval does not eliminate aggregation, deduplication, category-boundary, and arithmetic errors.

Overall, LongMemEval exhibits a mixed bottleneck across graph reachability, reader ranking, and post-retrieval reasoning. HaluMem is more strongly concentrated at reader ranking/coverage: the annotated memories are nearly always written and structurally reachable, yet the complete evidence set is rarely present in the top-10.

U.6 Relation to the Paired Host-Integration Experiments

The dominant HaluMem ranking failures motivate our paired host experiments: strong memory systems already provide mature extraction, persistence, updating, and answer-generation machinery, while SAGE specifically targets structure-aware memory access. Replacing the host's native memory-access/retrieval backend with SAGE therefore tests whether improving this layer can add value without requiring the rest of the host memory stack to be replaced; the exact intervention boundary is described in Appendix S.

However, the diagnostic audit above and the host-integration experiments are different evaluations. We therefore do *not* claim that the earlier failure percentages causally explain every Native→+SAGE improvement. Establishing that stronger causal link would require question-level provenance for the discordant pairs: which new evidence entered the SAGE context, whether it was semantically required, whether native evidence was displaced under the matched context budget, whether a stale fact was replaced by the current fact, and whether the answer actually used the added evidence. We consequently use the present audit to identify failure modes and motivate the intervention, while reserving performance attribution for the within-host matched comparisons.

Limitations

SAGE treats graph memory as a dynamic substrate for writing, reading, and self-evolution, but its effectiveness still depends on the quality of entity extraction, relation writing, source anchoring,

Table 26: End-to-end training ablations and selector diagnostics. Panel (a) reports EM and F1 (%) separately on three multi-hop QA benchmarks. The one-shot and equal-compute controls isolate the effects of additional optimization and writer-reader coupling, respectively. Panel (b) further reports selector diagnostics. Best results are in **bold** and runner-ups are underlined; darker cells indicate better performance.

Configuration	HotpotQA		MuSiQue		2Wiki	
	EM	F1	EM	F1	EM	F1
Full end-to-end model						
Full SAGE, +2 rounds	**56.1**	**80.8**	**34.2**	**53.1**	**74.0**	**80.4**
Component ablations						
w/o reader-aware writer training	52.4	76.7	30.4	48.3	68.1	76.8
w/o structured query planning	54.0	79.2	31.7	50.1	71.4	78.2
w/o global soft addressing	52.4	76.5	29.8	47.7	68.7	75.1
w/o structural gating	53.3	77.1	30.2	49.5	70.8	77.7
w/o context–schema decomposition, parameter-matched	53.2	77.1	30.7	49.3	70.7	77.4
w/o GFM pretraining, same architecture	53.4	77.0	30.6	49.9	70.3	77.1
w/o selector regularization	55.7	80.7	33.5	52.3	73.0	79.8
w/o query-conditioned selector	55.0	79.4	33.0	51.1	72.6	79.9
Optimization and coupling controls						
w/o self-evolution, one-shot	51.7	72.4	28.4	47.3	64.1	75.7
w/o closed-loop coupling, equal compute	54.6	78.6	31.3	49.2	70.1	77.3

(b) Selector diagnostics

Variant	Avg. QA F1	Avg. R@5	Activated nodes	Subgraph components
Full selector + all regularizers	**71.4**	**73.1**	18.6%	1.7
w/o NCE	70.9	72.6	26.2%	2.0
w/o size regularization	71.3	73.0	34.7%	1.5
w/o connectivity regularization	71.0	72.7	18.9%	3.8
w/o all selector regularizers	70.8	72.5	31.8%	3.4
w/o selector	70.3	72.0	100.0%	–

and reader feedback. Our experiments show promising results across multi-hop QA, open-domain retrieval, review-based QA, and long-term agent-memory benchmarks, but the current system still leaves room for improvement on memory updating, high-coverage extraction, and hallucination control in more realistic deployments.

Broader Impact

This work may have positive societal impact by improving the reliability and grounding of long-horizon language agents. A structure-aware and self-evolving memory system can help agents recover evidence chains from fragmented cues, maintain more consistent long-term interactions, and reduce unsupported answers in applications such as knowledge assistance, research support, customer support, and review-based question answering. At the same time, long-term agent memory raises important risks. If deployed on personal or sensitive interaction histories, such systems may store private information, infer user preferences, preserve outdated or incorrect memories, or enable profiling and surveillance. Incorrect graph writes or retrieval failures may also lead to confidently grounded but wrong answers. Practical deployments should therefore use consent-based data collection, data minimization, access control, deletion and forgetting mechanisms, provenance tracking, auditing, and human oversight for high-stakes use cases.

COMPUTE RESOURCES

All experiments were run on a server equipped with 8 NVIDIA A100 GPUs. The main computational cost of SAGE comes from graph-memory construction, GFM-based graph propagation, selector regularization, and entity-to-document projection. Appendix J analyzes the training and inference complexity in terms of the number of graph nodes n, edges m, hidden dimension d, propagation layers L, batch size B, pseudo-queries M, and entity-document links. In our implementation, structural features and entity-document indices can be precomputed and cached, while edge-level gates are computed in chunks to reduce peak GPU memory from $O(|E|d)$ to $O(C_e d)$ for chunk size C_e. The dominant inference cost is one or a small number of query-conditioned graph propagations followed by sparse document projection, making the reader suitable for repeated evaluation inside the self-evolving writer-reader loop.

LICENSES AND EXISTING ASSETS

This paper uses existing public benchmarks and baselines, including NQ-Open, PopQA, HotpotQA, 2WikiMultiHopQA, MuSiQue, AmazonQA, LongMemEval, HaluMem, BM25, Contriever, GTR, ColBERTv2, RAPTOR, GraphRAG, G-Retriever, LightRAG, HippoRAG, HippoRAG 2, SubgraphRAG, PropRAG, GFM-RAG, IRCoT, FLARE, and Adaptive-RAG. We cite the original papers or repositories for these assets and use them only for research evaluation under their stated licenses and terms of use. We do not redistribute modified versions of the datasets beyond the preprocessing scripts and instructions needed for reproducibility. The released code is intended for research use and includes documentation for environment setup, data preparation, training, and evaluation.

AI USE STATEMENT

Generative AI tools are only used for English grammar and format checking. No scientific claims, results, analyses, citations, or other substantive manuscript content were changed.