KMMG-AGENT: A KNOWLEDGE–MEMORY MULTI-GRAPH AGENT FOR MULTI-HOP EGOCENTRIC QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large pretrained language models can store vast amounts of factual knowledge and leverage it to answer user queries. However, they remain limited when handling information beyond their knowledge boundaries. Equipping agents with external knowledge interaction and memory mechanisms has proven to be an effective strategy, enabling them to dynamically acquire explicit knowledge and incorporate it into the reasoning process, thereby improving answer accuracy. Nevertheless, existing memory retrieval and information matching methods often overlook temporal information in memory and lack effective modeling of the connections between dynamic memories and large-scale external knowledge. To address these limitations, we propose the Knowledge Memory Multi-Graph Agent (KMMG-Agent), which preserves repeated memory events as distinct timestamped edges and combines query-specific temporal filtering with intermediate-entity-based chained retrieval to support two-hop question answering over personal memories and Wikipedia knowledge. To systematically evaluate the effectiveness of the proposed framework across different cross-source reasoning scenarios, we further construct a multi-hop question-answering benchmark that integrates memory and external knowledge and covers four distinct reasoning paths. Experimental results show that KMMG-Agent outperforms existing baselines on all three memory-related tasks, achieving approximately 11.8%-33.2% improvements in Exact Match (EM) accuracy. These results demonstrate the effectiveness of KMMG-Agent in temporal memory retrieval and cross-source multi-hop reasoning.

1 INTRODUCTION

In recent years, large pretrained language models have achieved remarkable progress in natural language processing (Brown et al., 2020). Through pretraining on massive corpora, large language models can implicitly encode substantial amounts of factual knowledge in their parameters and leverage such knowledge to answer user queries (Petroni et al., 2019). However, parametric knowledge remains inherently limited in both scale and timeliness. When questions involve personal events, long-tail domain knowledge, or specific external resources, models often struggle to provide accurate answers and may even generate factual hallucinations (Ji et al., 2023). Therefore, supplementing models with missing temporally grounded personal memories and external knowledge has become a key challenge for improving the reliability of question answering.

Agents equipped with external knowledge interaction and memory mechanisms provide an effective solution to this problem. Before generating an answer, an agent can proactively access external knowledge bases or long-term memory and incorporate the retrieved explicit information into the reasoning process. In this way, the model can extend its knowledge beyond the boundaries of its parametric knowledge without retraining, while reducing factual errors and hallucinations through traceable evidence.

As large language model applications expand from static text interaction to complex long-sequence scenarios (Packer et al., 2023; Yu et al., 2023), temporally structured text such as video captions contains rich event information and places higher demands on memory construction, storage, and

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.

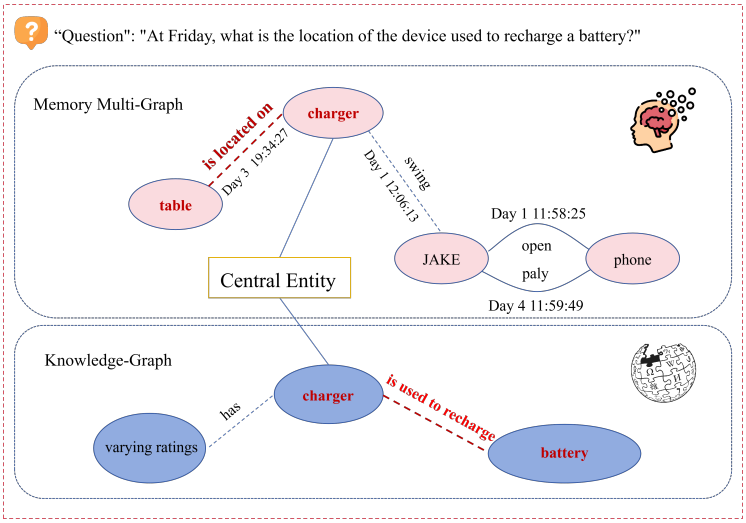

Figure 1: Knowledge Memory Multi-Graph

retrieval (Zhong et al., 2024). Existing approaches typically employ external storage or memory mechanisms to organize information as graph nodes or text fragments for agent access (Zhong et al., 2024; Edge et al., 2024). However, they often lack explicit modeling of temporal attributes, making it difficult to effectively handle complex queries involving event ordering, temporal spans, and dynamic changes (Chu et al., 2024).

To support question answering over temporal memories and external knowledge, we combine event-level memory representation with query-specific temporal filtering and entity-mediated chained retrieval. As illustrated in Figure 1, the memory multi-graph retains repeated interactions between entities as distinct edges with their own relations and timestamps. Temporal filtering restricts memory retrieval to events consistent with the query, while the intermediate entity connects the two retrieval steps, including cross-source transitions between personal memories and Wikipedia knowledge.

Existing datasets mainly focus on a single information source and rarely model the interactions between personal memory and external knowledge, making it difficult to evaluate multi-hop reasoning across heterogeneous evidence sources. To address this limitation, we construct Mem-WikiQA, a memory–knowledge question-answering benchmark with four two-hop evidence-source paths: Memory-to-Memory, Memory-to-Wiki, Wiki-to-Memory, and Wiki-to-Wiki. A question may connect two personal events, use a memory-derived entity to retrieve external knowledge, or use a knowledge-derived entity to identify a relevant personal event. The benchmark combines timestamped memories derived from EgoLife captions with Wikipedia knowledge, enabling systematic evaluation of both within-source and cross-source multi-hop reasoning.

The limitations of conventional approaches in temporal information utilization and cross-source reasoning jointly constrain the performance of large language models in realistic and complex question-answering scenarios. To address these two challenges, this work makes contributions at both the dataset and methodological levels. Our main contributions are summarized as follows:

1. We construct Mem-WikiQA, a two-hop question-answering benchmark comprising 2,000 questions derived from EgoLife captions and Wikipedia knowledge. The benchmark covers four evidence-source paths, with 500 questions per path, enabling systematic evaluation of reasoning within and across personal memories and external knowledge.

2. We propose KMMG-Agent, a knowledge–memory retrieval framework that preserves repeated memory events as distinct timestamped edges. By combining query-specific temporal filtering with entity-neighborhood-guided hybrid retrieval, the framework selects evidence for two-hop question answering over personal memories and Wikipedia knowledge.

3. We conduct systematic comparative evaluations of multiple large language models in all four categories and compare our approach with both conventional knowledge-enhanced methods and

memory-oriented baselines. Experimental results show that the proposed agent consistently outperforms baseline methods on all three memory-related tasks, demonstrating its effectiveness.

2 Related Work

2.1 Retrieval Augmented Generation

Retrieval-augmented generation (RAG) complements LLMs with external knowledge through naive, advanced, and modular paradigms (Lewis et al., 2020; Gao et al., 2023). Dense retrieval, joint retriever–generator training, query rewriting, and iterative retrieval improve evidence quality (Karpukhin et al., 2020; Izacard et al., 2022; Ma et al., 2023; Jiang et al., 2023), while retrieval-augmented pretraining and external memory broaden knowledge access (Guu et al., 2020; Borgeaud et al., 2022). Self-RAG and FLARE adaptively retrieve and update evidence to mitigate noise and missing information (Asai et al., 2024; Jiang et al., 2023).

Video RAG uses CLIP-based cross-modal retrieval (Radford et al., 2021), keyframe or clip sampling (Tang et al., 2025), or video-to-text conversion via captions, transcripts, and OCR (Luo et al., 2026; Jeong et al., 2025). However, frame- or fragment-based representations often lack explicit temporal relations, limiting long-range event reasoning.

GraphRAG, G-Retriever, and HippoRAG exploit graph structures for retrieval and multi-hop reasoning (Edge et al., 2024; He et al., 2024; Jimenez Gutierrez et al., 2024), but preserving memory timelines while aligning dynamic memories with external knowledge remains challenging. Our knowledge–memory multi-graph unifies temporal evolution, internal memory relations, and external knowledge associations to support cross-source reasoning.

Our approach preserves repeated caption-derived events as timestamped edges and combines temporal filtering with entity-mediated retrieval over personal memories and Wikipedia knowledge.

2.2 External Memory Mechanisms for Agents

External memory supports persistent information storage and retrieval for long-term agent interactions (Park et al., 2023). MemGPT uses hierarchical storage and on-demand retrieval (Packer et al., 2023), A-Mem autonomously organizes memories (Xu et al., 2025), MemoryBank models forgetting (Zhong et al., 2024), and Voyager maintains reusable skills (Wang et al., 2024). Mem0 extracts salient memories, while Mem0g adds directed relational graphs (Chhikara et al., 2025). LightMem combines a three-tier memory architecture with lightweight filtering and offline updates (Fang et al., 2026), whereas SimpleMem uses structured compression, recursive consolidation, and adaptive retrieval (Liu et al., 2026).

However, modeling complex event relationships over extended periods remains challenging. Our multi-graph organizes caption-derived memories with event, temporal, and relational information and aligns them with an external knowledge graph.

2.3 Multi-Hop Reasoning and Multimodal Datasets

Multi-hop reasoning benchmarks range from text-based HotpotQA and 2WikiMultiHopQA (Yang et al., 2018; Ho et al., 2020) to multimodal WebQA (Chang et al., 2022). OK-VQA emphasizes knowledge-based image understanding (Marino et al., 2019), while M3-VQA links images with external text for multi-hop reasoning (Ma et al., 2026). MVBench and EgoSchema evaluate spatiotemporal understanding and long-form egocentric reasoning, respectively (Li et al., 2024; Mangalam et al., 2023). KnowIT VQA, EgoLife, SeViLA, and NExT-QA address knowledge augmentation, long-term recall, retrieval augmentation, and temporal-causal reasoning, respectively (Garcia et al., 2020; Yang et al., 2025; Yu et al., 2023; Xiao et al., 2021).

Existing benchmarks rarely provide explicit connections between temporal memories and external knowledge for stepwise reasoning. We address this gap with an EgoLife-based memory–knowledge multi-hop question-answering dataset.

We extract timestamped memory facts from egocentric video captions, construct Wikipedia triples, and align entities to generate two-hop questions spanning different source combinations. These sam-

Table 1: Comparison of different QA datasets.

Dataset	Memory	Knowledge	Multi-hop	Time
HotpotQA (Yang et al., 2018)	✗	✓	✓	✗
OK-VQA (Marino et al., 2019)	✗	✓	✗	✗
M3-VQA (Ma et al., 2026)	✗	✓	✓	✗
EgoLife (Yang et al., 2025)	✓	✗	✗	✓
Mem-WikiQA	✓	✓	✓	✓

ples evaluate agents' ability to combine personal experience with external knowledge and assess the effectiveness of our knowledge–memory multi-graph.

2.4 Source of Data

EgoLife is a large-scale, multimodal, egocentric daily-life dataset featuring ultra-long contexts (Yang et al., 2025). It provides an immersive record of the daily activities of six participants living together in a real-world villa for seven consecutive days, comprising approximately 300 hours of high-resolution first-person video captured using smart glasses. In terms of annotation, the dataset provides complete speech transcripts as well as dense, continuous video and audio captions at 30-second intervals. These annotations cover multi-person social interactions, complex daily collaboration, and temporal and contextual clues spanning several hours or even multiple days.

3 Dataset

3.1 Construction of the Multi-Hop QA Dataset

Mem-WikiQA evaluates two-hop question answering over personal memories and Wikipedia knowledge. We first use the DeepSeek-chat API to transform descriptions of the daily activities of six participants in the EgoLife video captions into timestamped memory triples. We then aggregate the triples using their entities as anchors and select 80 high-frequency entities according to their occurrence frequencies as candidate central entities. The associated memory triples, original captions, and timestamps are retained as memory-side evidence.

During knowledge-base construction, we retrieve passages from the full Wikipedia corpus using an encoding and retrieval architecture based on ColBERT-v2 and FAISS (Santhanam et al., 2022; Johnson et al., 2019). Each central entity is used as a query to retrieve the 50 most semantically relevant Wikipedia passages. The DeepSeek-chat API is subsequently employed to extract the corresponding knowledge triples. Entities on the memory and knowledge sides are then normalized and aligned to establish connections between the two information sources.

Specifically, we treat each central entity as an intermediate hop and select combinations of triples that can form valid multi-hop reasoning paths. These triples, together with their associated timestamps, are provided to the Gemini-3 API to generate candidate question–answer pairs. Questions involving memory evidence are required to preserve the temporal information of the corresponding events. According to the sources of evidence along the multi-hop paths, the question–answer pairs are divided into four categories. In memory-to-memory questions, both pieces of evidence are derived from memory triples, and the question contains two corresponding time expressions. Memory-to-wiki and wiki-to-memory questions combine memory and knowledge evidence in different orders, with each question containing one memory-related time expression. In wiki-to-wiki questions, both pieces of evidence are derived from knowledge triples, and no temporal information is included.

For each central entity, we generate 20 candidate question–answer pairs in each category, yielding 1,600 candidate samples per category. We then manually screen the samples against the original captions and Wikipedia passages. The quality-control process primarily examines the factual correctness of the evidence, the consistency of the intermediate entity, the completeness of the reasoning chain, and the accuracy of the temporal descriptions. Ultimately, 500 high-quality samples are retained for each category.

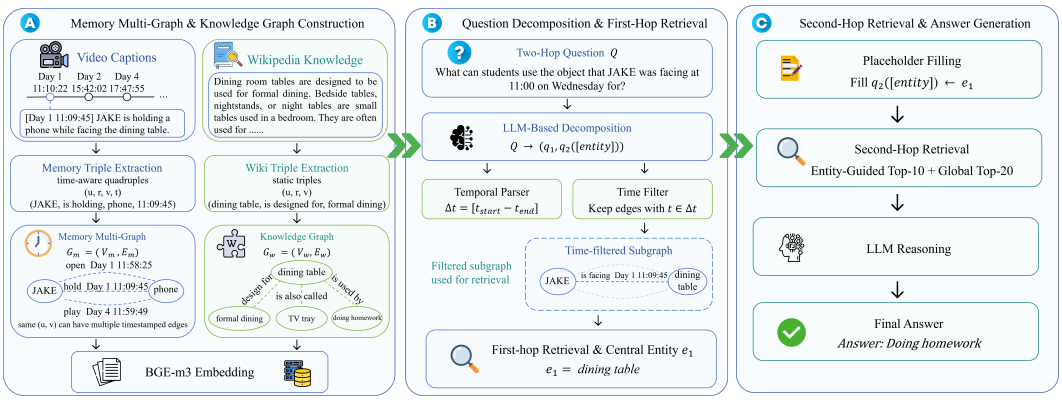

Figure 2: Knowledge Memory Multi-Graph Agent framework

To better reflect how temporal memories are recalled in real-life scenarios, we rewrite the time expressions at four levels of granularity: half-hour, hour, half-day, and full-day, accounting for 50%, 30%, 10%, and 10% of the samples, respectively. We further assume that all questions are asked after the end of Day 7 and employ three types of date expressions—days of the week, specific dates, and relative-day expressions—to increase the diversity of temporal descriptions. Details of the data generation procedure are provided in the appendix A.

To further demonstrate the comprehensiveness of the proposed dataset, we compare it with existing mainstream benchmarks across multiple dimensions. The detailed comparison is presented in Table 1.

4 METHOD

To better address cross-source multi-hop reasoning, we propose the Knowledge Memory Multi-Graph Agent framework, whose overall workflow is illustrated in Figure 2. The proposed method consists of two main components: (1) the construction of the memory multi-graph and the Wikipedia knowledge graph, and (2) knowledge–memory retrieval and answer generation.

4.1 CONSTRUCTION OF THE MEMORY MULTI-GRAPH AND KNOWLEDGE GRAPH

We construct a memory multi-graph from caption-derived memory triples by representing each event instance as a separate timestamped edge. Repeated interactions between the same entities at different times are retained as distinct temporal facts, preserving both their relational content and temporal information. This event-level representation provides the basis for subsequent query-specific temporal filtering and two-hop evidence retrieval.

Let V_m denote the set of entity nodes extracted from the dense captions of egocentric videos, and let R_m denote the set of relation types. We define the memory multi-graph as

$$G_m = (V_m, E_m), \tag{1}$$

where E_m is a multiset of directed temporal edges. Each edge $e \in E_m$ is formally defined as a quadruple:

$$e = (u, r, v, t), \tag{2}$$

where $u \in V_m$ is the head entity, $v \in V_m$ is the tail entity, $r \in R_m$ denotes a specific action or relation, and $t \in \mathcal{T}$ represents the precise timestamp at which the corresponding event occurred. Therefore, multiple events involving the same pair of entities can be represented by distinct temporal edges, allowing the memory multi-graph to preserve both relational and temporal information.

In parallel, we organize the extracted knowledge triples into an external Wikipedia knowledge graph, which provides external factual knowledge for subsequent reasoning tasks. The Wikipedia knowl-

edge graph is defined as a standard graph:

$$G_w = (V_w, E_w), \tag{3}$$

where V_w denotes the set of Wikipedia entities and

$$E_w \subseteq V_w \times R_w \times V_w \tag{4}$$

is the set of static knowledge triples without temporal attributes. Each knowledge edge $e_w \in E_w$ is represented as

$$e_w = (u, r, v), \tag{5}$$

where $u, v \in V_w$ denote the head and tail entities, respectively, and $r \in R_w$ denotes the semantic relation between them.

After constructing the graph structures, we further encode the triples in the graphs into vector representations to enable efficient semantic retrieval in subsequent stages. Specifically, we convert each "head entity–relation–tail entity" structure into a complete natural-language sentence. We then employ the pretrained BGE-M3 model to encode these sentences. Leveraging the strong deep semantic representation capability of BGE-M3, the structured triples are mapped into high-dimensional dense vectors enriched with contextual information. This process not only bridges the semantic gap between structured graph data and unstructured text, but also provides a solid foundation for accurate knowledge retrieval and information recall in downstream tasks.

4.2 Knowledge-Memory Retrieval and Answer Generation

To address cross-source multi-hop reasoning, we adopt a stepwise framework based on question decomposition and chained retrieval.

Question Decomposition. Given a multi-hop question Q, an LLM decomposes it into two dependent subquestions:

$$(q_1, q_2) = Decompose(Q), \tag{6}$$

where q_1 retrieves the intermediate entity and q_2 contains a placeholder `[entity]` representing the answer to q_1. For memory-related subquestions, the temporal information in Q is explicitly preserved to support time-aware retrieval.

First-Hop Retrieval. If q_1 contains a temporal expression, it is normalized into an interval

$$\Delta t = [t_{\text{start}}, t_{\text{end}}] \subseteq \mathcal{T}. \tag{7}$$

A hard temporal filtering operator then extracts the corresponding memory subgraph:

$$G_m^{\Delta t} = \Phi(G_m, \Delta t) = (V_m^{\Delta t}, E_m^{\Delta t}), \tag{8}$$

where

$$E_m^{\Delta t} = \{(u, r, v, t) \in E_m \mid t_{\text{start}} \leq t \leq t_{\text{end}}\}. \tag{9}$$

The subquestion and candidate graph entries are encoded using BGE-M3, ranked by cosine similarity, and the Top-30 entries are provided to the LLM as evidence. The model is constrained to output only the intermediate entity e_1. For subquestions without temporal constraints, retrieval is performed directly over the complete graph.

Second-Hop Query: Placeholder Filling, Entity-Guided Retrieval, Final Answer Generation. The predicted first-hop entity e_1 is substituted into the `[entity]` placeholder in the second subquestion:

$$\hat{q}_2 = \text{Replace}\,(q_2, \texttt{[entity]}, e_1)\,. \tag{10}$$

For the second hop, candidate facts are filtered according to the temporal constraints of the instantiated query $\hat{q}_2$, using the same procedure as in the first hop. Temporal filtering is omitted when $\hat{q}_2$ contains no temporal constraints. Let $\mathcal{C}_2$ denote the resulting candidate set. We encode $\hat{q}_2$ and each fact in $\mathcal{C}_2$ using BGE-M3 and compute their query–fact cosine similarity scores.

We use an entity-neighborhood-guided hybrid retrieval strategy with two branches. The entity-neighborhood branch uses the first-hop answer entity e_1 as an anchor and constructs $\mathcal{C}_2(e_1) \subseteq \mathcal{C}_2$,

Table 2: Different Methods Performance on Multi-Hop Question-Answering Tasks

Method	Model	Memory-to-Memory	Memory-to-Wiki	Wiki-to-Memory	Wiki-to-Wiki
KMMG(Ours)	Qwen3-32B	52.4%	64.0%	46.8%	43.6%
	GPT-5.4	**53.8%**	**70.4%**	**49.4%**	**49.2%**
RAG	Qwen3-32B	14.4%	22.8%	6.0%	31.4%
	GPT-5.4	13.2%	29.6%	8.6%	36.6%
LightMem	Qwen3-32B	7.8%	24.0%	5.6%	31.4%
	GPT-5.4	4.8%	28.0%	5.8%	36.6%
SimpleMem	Qwen3-32B	12.8%	25.8%	9.6%	31.4%
	GPT-5.4	14.8%	36.2%	13.6%	36.6%
Mem0g	Qwen3-32B	19.2%	23.2%	15.8%	31.4%
	GPT-5.4	20.6%	32.4%	17.8%	36.6%
GraphRAG	Qwen3-32B	18.0%	52.2%	20.0%	43.6%
	GPT-5.4	19.2%	56.0%	21.6%	**49.2%**

containing only facts whose head or tail entity matches e_1. It selects the Top-K facts from $\mathcal{C}_2(e_1)$ based on query–fact cosine similarity. The global semantic branch selects the Top-L facts from $\mathcal{C}_2$ using the same similarity scores, without imposing an entity-matching constraint. Both branches therefore respect the same temporal constraints, when present, while only the entity-neighborhood branch restricts retrieval to facts directly connected to e_1.

The retrieval outputs from both branches are combined to form the second-hop evidence, which is provided to the LLM together with $\hat{q}_2$ to generate the final answer to the original multi-hop question. This strategy combines explicit entity connectivity with global semantic relevance for evidence selection. Its effectiveness is evaluated in the ablation study.

5 Experiments

We evaluate KMMG-Agent on Mem-WikiQA and investigate the following three research questions: (1) How does KMMG-Agent compare with retrieval-augmented and memory-oriented baselines across the four evidence-source paths and two backbone models? (2) How do different configurations of memory representation, temporal filtering, and second-hop retrieval affect question-answering performance? (3) How does the runtime of KMMG-Agent compare with that of the baselines under the Qwen3-32B setting?

5.1 Experimental Setup

Baseline Methods. Our comparative experiments cover two mainstream technical paradigms. First, to highlight the advantages of our approach in memory modeling, we benchmark KMMG-Agent against representative memory-based methods, including LightMem, SimpleMem, and Mem0. Second, we include conventional RAG and GraphRAG as baselines to systematically compare the proposed multi-graph architecture with traditional text-based retrieval and single-graph retrieval in terms of knowledge recall and question-answering accuracy.

Backbone Models. To evaluate KMMG-Agent across different backbone models, we conduct experiments using Qwen3-32B and GPT-5.4. Both models are evaluated on the same four evidence-source paths.

Implementation Details. We conduct experiments following the complete pipeline of memory multi-graph and Wikipedia knowledge graph construction, placeholder filling, entity-neighborhood-guided retrieval, and final answer generation. We introduce subquestion decomposition into all baseline methods and adopt a consistent two-hop retrieval and reasoning pipeline across methods. For the default second-hop configuration, we set $K = 10$ for the entity-neighborhood branch and $L = 20$ for the global semantic branch. The former retrieves facts whose head or tail entity matches the first-hop predicted entity, while the latter ranks the candidate set without this entity constraint. The retrieval

Table 3: Ablation Study of Memory Representation, Temporal Filtering, and Retrieval Strategy

Setting	Model	Memory-to-Memory	Memory-to-Wiki	Wiki-to-Memory
Graph w/o Time	Qwen3-32B	18.0%	52.2%	20.0%
	GPT-5.4	19.2%	56.0%	21.6%
Text Retrieval w/ Time	Qwen3-32B	49.6%	64.4%	47.4%
	GPT-5.4	50.4%	69.2%	48.0%
Multi-Graph w/o Time	Qwen3-32B	23.8%	54.4%	23.0%
	GPT-5.4	25.6%	58.6%	24.4%
Multi-Graph w/ Time	Qwen3-32B	52.4%	64.0%	46.8%
	GPT-5.4	**53.8%**	**70.4%**	**49.4%**

Table 4: Runtime comparison of different methods.

Method	Memory-to-Memory	Memory-to-Wiki	Wiki-to-Memory	Average
KMMG(Ours)	**168.55**	**120.09**	**117.44**	**135.36**
RAG	200.45	153.71	154.86	169.67
LightMem	224.74	141.69	157.58	174.67
SimpleMem	173.61	154.06	160.25	162.64
Mem0g	174.26	132.49	135.13	147.29
GraphRAG	170.44	121.74	123.54	138.57

Runtime is reported in seconds using Qwen3-32B. Bold indicates the lowest runtime.

outputs are combined and provided with the instantiated second subquestion to the language model for final answer generation.

Evaluation Metric. We evaluate model performance using Exact Match (EM). Since the large language models are instructed to output only the answer entity, EM provides a strict and unambiguous measure of whether a model can identify the unique correct knowledge entity in complex two-hop reasoning scenarios.

Table 2 reports the performance of KMMG-Agent and the comparison methods with Qwen3-32B and GPT-5.4. KMMG-Agent achieves the highest EM among the methods listed in this table on Memory-to-Memory, Memory-to-Wiki, and Wiki-to-Memory under both backbones. The improvements over the strongest competing method in each of these settings range from 11.8 to 33.2 percentage points. For the Wiki-to-Wiki task, existing memory methods do not explicitly model external knowledge sources such as Wikipedia, and their retrieval process is therefore essentially equivalent to conventional RAG. As a result, their EM scores on this task are also consistent with those of the RAG baseline.

Compared with conventional RAG, KMMG-Agent jointly models personal memories and Wikipedia knowledge in a knowledge–memory multi-graph and applies temporal filtering, enabling more accurate cross-source and time-aware multi-hop retrieval than flat text retrieval. Compared with GraphRAG, KMMG-Agent supports multiple relations and timestamps between the same entities and filters candidate subgraphs according to query time constraints, reducing outdated or temporally conflicting evidence. Compared with LightMem and SimpleMem, KMMG-Agent preserves fine-grained facts as timestamped memory edges rather than relying on compression or summarization, while explicitly modeling memory evolution through its multi-graph structure and temporal filtering. Compared with Mem0g, KMMG-Agent incorporates explicit temporal consistency in addition to semantic relevance, reducing interference from temporally mismatched memories and improving cross-source multi-hop reasoning.

These comparisons evaluate the complete framework rather than isolate the effects of individual components. KMMG-Agent combines event-level memory preservation, temporal filtering, and chained retrieval. We therefore examine additional configurations in Table 3 to characterize the performance differences associated with these design choices.

5.2 Ablation Study

Multi-Graph & Temporal Filtering: Table 3 compares memory representations, temporal filtering, and retrieval configurations. Without temporal filtering, the multi-graph configuration improves EM by 2.2–6.4 percentage points over the standard graph across the three memory-related tasks and both backbones. Adding temporal filtering to the multi-graph yields larger improvements of 9.6–28.6 percentage points. Compared with the *Text Retrieval w/ Time* setting, the full method improves Memory-to-Memory by 2.8 and 3.4 percentage points with Qwen3-32B and GPT-5.4, respectively. On the two cross-source tasks, it improves EM by 1.2–1.4 percentage points with GPT-5.4, while obtaining slightly lower scores with Qwen3-32B, by 0.4–0.6 percentage points. One possible explanation is that Qwen3-32B has difficulty integrating natural-language statements verbalized from structured facts across heterogeneous sources. Such difficulties may limit its ability to exploit entity-linked evidence when hybrid retrieval changes the composition of the retrieved context. Overall, these results support the benefits of the multi-graph representation and temporal filtering in the evaluated settings.

Reasoning Efficiency: Table 4 compares the total runtime required to answer 50 questions on each of the three memory-related tasks using Qwen3-32B. KMMG-Agent achieves the lowest total runtime on all three tasks and the lowest average runtime across tasks, requiring 135.36 s to complete 50 questions on average, compared with 169.67 s for RAG and 138.57 s for GraphRAG. This corresponds to approximately 2.71 s per question for KMMG-Agent, versus 3.39 s for RAG and 2.77 s for GraphRAG. The efficiency advantage of KMMG-Agent is largely attributable to its temporal filtering mechanism, which narrows the candidate document set according to temporal constraints before retrieval, thereby reducing the amount of content that must be processed during subsequent retrieval and evidence selection.

Embedding Model & Top-K: We further evaluate the effects of the embedding model and retrieval size by comparing BGE-M3 with Qwen3-4B and varying Top-K among 10, 30, and 50. The results show that performance is relatively insensitive to encoder scale, while Top-K provides a favorable balance between evidence coverage, retrieval noise, and computational cost. Therefore, BGE-M3 and Top-K=30 are adopted as the default settings, with detailed results reported in the appendix B.

6 Conclusion

To address insufficient use of temporal information and weak memory–knowledge connections, we introduce Mem-WikiQA, a two-hop question-answering benchmark built from EgoLife captions and Wikipedia knowledge covering four evidence-source paths, and propose KMMG-Agent. The framework preserves repeated events as distinct timestamped edges and combines query-specific temporal filtering with intermediate-entity-based chained retrieval. Its second-hop dual-branch hybrid retrieval integrates an entity-neighborhood branch for evidence directly linked to the intermediate entity and a global semantic branch for semantically relevant evidence without entity-matching constraints. Both branches respect the same temporal constraints, when present, supporting evidence selection for within-source and cross-source reasoning. Future work will expand the benchmark, enhance fine-grained temporal reasoning, and explore long-form video understanding and broader memory–knowledge question answering.

AI Assistance Disclosure

Large language models were used to assist with English translation and language polishing. For the construction of Mem-WikiQA, DeepSeek-chat was used to extract timestamped memory triples from EgoLife captions and knowledge triples from Wikipedia passages, and Gemini-3 was used to generate candidate two-hop question–answer pairs. The generated samples were manually screened against the original captions and Wikipedia passages, as described in Section 3.1.

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.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 2206–2240, 2022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16495–16504, 2022.

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.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. Timebench: A comprehensive evaluation of temporal reasoning abilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1204–1228, 2024.

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

Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, Huajun Chen, and Ningyu Zhang. Lightmem: Lightweight and efficient memory-augmented generation. In *International Conference on Learning Representations*, 2026.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

Noa Garcia, Mayu Otani, Chenhui Chu, and Yuta Nakashima. Knowit vqa: Answering knowledge-based questions about videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 10826–10834, 2020.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 3929–3938, 2020.

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. In *Advances in Neural Information Processing Systems*, volume 37, 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.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*, 1(2):4, 2022.

Soyeong Jeong, Kangsan Kim, Jinheon Baek, and Sung Ju Hwang. Videorag: Retrieval-augmented generation over video corpus (2025). *URL https://arxiv. org/abs/2501.05874*, 2025.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

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.

Bernal Jimenez Gutierrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. Hipporag: Neurobiologically inspired long-term memory for large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 59532–59569, 2024.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

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

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

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024.

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

Yongdong Luo, Xiawu Zheng, Guilin Li, Shukang Yin, Haojia Lin, Chaoyou Fu, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, et al. Video-rag: Visually-aligned retrieval-augmented long video comprehension. *Advances in Neural Information Processing Systems*, 38:168008–168033, 2026.

Jiatong Ma, Longteng Guo, Yuchen Liu, Zijia Zhao, Dongze Hao, Xuanxu Lin, and Jing Liu. M^3-VQA: A benchmark for multimodal, multi-entity, multi-hop visual question answering. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40633–40669, 2026.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5303–5315, 2023.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46212–46244, 2023.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3195–3204, 2019.

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

Joon Sung Park, Joseph C. O'Brien, Carrie J. 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.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 2463–2473, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29118–29128, 2025.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*, 2024.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9777–9786, 2021.

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*, 2025.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin, Marco Cominelli, Zhongang Cai, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, Francesco Gringoli, Lei Yang, Bo Li, and Ziwei Liu. Egolife: Towards egocentric life assistant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28885–28900, 2025.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. 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.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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.

A Data Generation Procedure

To better illustrate the data generation process, we present the complete pipeline in the figure 3.

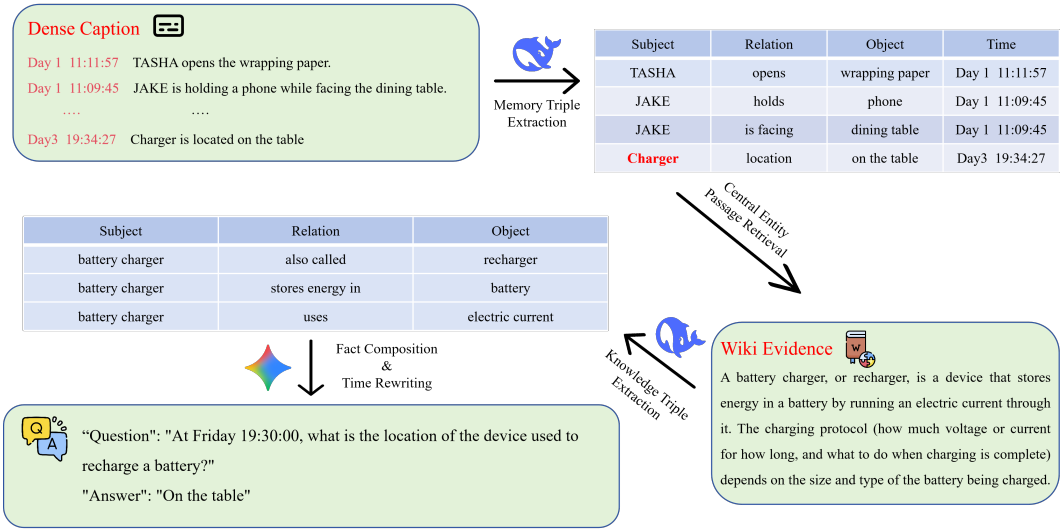

Figure 3: Construction Pipeline of the Memory–Wikipedia Two-Hop QA Dataset

B Embedding Model & Top-K Ablation Study

Embedding Model: To assess the impact of embedding models on question-answering performance, we conduct ablation experiments using BGE-M3 and Qwen3-4B as encoders. The detailed experimental results are presented in the table 5. The results show that the two encoders achieve comparable Exact Match (EM) scores across the four types of two-hop tasks, suggesting that the multi-graph structure, temporal filtering, and chained retrieval reduce the system's dependence on embedding model size. Balancing performance and computational cost, we adopt the more lightweight and efficient BGE-M3 as our default encoder.

Table 5: Ablation study of embedding models across four two-hop QA tasks. Results are reported as Exact Match (EM). The best result on each task is highlighted in bold.

Embedding Model	Backbone Model	Memory-to-Memory	Memory-to-Wiki	Wiki-to-Memory	Wiki-to-Wiki
BGE-M3	Qwen3-32B	52.4%	64.0%	46.8%	43.6%
	GPT-5.4	53.8%	**70.4%**	49.4%	**49.2%**
Qwen3-4B	Qwen3-32B	52.8%	63.4%	47.2%	43.4%
	GPT-5.4	**54.2%**	68.8%	**49.8%**	48.4%

Top-K: To assess the impact of the number of retrieved candidates on performance, we conduct ablation experiments with Top-K set to 10, 30, and 50. The detailed experimental results are presented in the table 6. The results show that increasing K generally improves multi-hop evidence coverage, whereas retrieving too many candidates can introduce redundant information and degrade performance on some tasks. Among the tested settings, K=30 offers the best balance between accuracy, context noise, and inference cost, and is therefore adopted as the default setting for the main experiments.

Table 6: Ablation study of Top-K across four two-hop QA tasks. Results are reported as Exact Match (EM). The best result for each Backbone Model on each task is highlighted in bold. We use $K = 30$ as the default setting.

Backbone Model	Top-K	Memory-to-Memory	Memory-to-Wiki	Wiki-to-Memory	Wiki-to-Wiki
Qwen3-32B	10	40.4%	61.8%	42.4%	38.2%
	30	52.4%	64.0%	46.8%	**43.6%**
	50	**53.2%**	**66.8%**	**48.0%**	38.6%
GPT-5.4	10	43.2%	64.4%	44.8%	46.6%
	30	53.8%	70.4%	49.4%	**49.2%**
	50	**54.6%**	**71.2%**	**50.2%**	45.0%

C Prompt Templates

This section presents the prompt templates used in our Memory-to-Wiki framework. Placeholders enclosed in braces are replaced with instance-specific inputs at runtime.

C.1 Query Decomposition

Prompt for query decomposition

Task: Decompose this question into two sub−questions. The first sub−question should identify the entity that satisfies the relation at the specified time in the question. Only return the two sub−questions and do not provide any explanations.
Original Question: {question_en}

C.2 First Subquestion Answer Generation

Prompt for answer generation

Task: Answers the question with memory.
Reference Knowledge: {results_memory}
Target Question: {first subquestion}

Constraints:
1. Source Grounding: Strictly use ONLY the provided "Memory Triples". If no answer is contained within the triples, return "None".
2. Selection Logic: Return ALL entities that match the Query. Do NOT prioritize by score.
3. Relevance Rule: An entity is considered valid if it correctly answers the question based on the relation in the triples.
4. Strict Formatting: Return ONLY the entity names. Do NOT include scores, punctuation, explanations, or any additional text.
5. multi−answer: If multiple entities satisfy the query, return all of them, each on a new line.
6. Zero Reasoning: Do not explain your thought process. Do not justify the selection.

C.3 Rewrite Second Subquestion

Prompt for rewrite second subquestion

Task: Rewrite a question by replacing a specific entity.
Original Question: "{second subquestion}"
Replacement Entity: "{first subquestion response}"

Instruction: Replace the main subject/entity in the Original Question with the Replacement Entity.
Constraints:
1. Maintain the original sentence structure and grammatical correctness.
2. Output ONLY the rewritten question.
3. Do not include any introductory text, quotes, or explanations.

C.4 Answer Generation

Prompt for answer generation

Task: Answers the question with knowledge.
Reference Knowledge: {results_wiki}
Target Question: {rewritten question}

Constraints:
1. Selection Strategy: Identify Triples that directly address the question.
2. Cross−check Sources: "entity_filtered" triples are grounded to the first−hop entity; " global_semantic" triples can correct entity extraction, aliases, or missed matches; " both" means both retrieval routes agree. Compare both sources and choose the answer best supported by the target question and triple relation. Do not blindly prefer either source.
3. Conflict Handling: If the two sources conflict, prefer triples whose head/tail and relation jointly match the Target Question.
4. Output Format: Return ONLY the raw text. No punctuation, no quotes, no introductory filler.