EventRAG: Enhancing LLM Generation with Event Knowledge Graphs

Zairun Yang[1], **Yilin Wang**[1], **Zhengyan Shi**[2], **Yuan Yao**[1],
Lei Liang[3], **Keyan Ding**[1], **Emine Yilmaz**[2], **Huajun Chen**[1], **Qiang Zhang**[1†],

[1]Zhejiang University, [2]University College London, [3]Ant Group
`{zr.yang, qiang.zhang.cs}@zju.edu.cn`

Abstract

Retrieval-augmented generation (RAG) systems often struggle with narrative-rich documents and event-centric reasoning, particularly when synthesizing information across multiple sources. We present *EventRAG*, a novel framework that enhances text generation through structured event representations. We first construct an *Event Knowledge Graph* by extracting events and merging semantically equivalent nodes across documents, while expanding under-connected relationships. We then employ an iterative retrieval and inference strategy that explicitly captures temporal dependencies and logical relationships across events. Experiments on UltraDomain and MultiHopRAG benchmarks show EventRAG's superiority over baseline RAG systems, with substantial gains in generation effectiveness, logical consistency, and multi-hop reasoning accuracy. Our work advances RAG systems by integrating structured event semantics with iterative inference, particularly benefiting scenarios requiring temporal and logical reasoning across documents.

1 Introduction

Retrieval-augmented generation (RAG) has become a cornerstone for generating knowledge-grounded texts in diverse applications (Edge et al., 2024; Qian et al., 2024; Es et al., 2024). Conventional RAG methods (Lewis et al., 2020) typically enhance the generation process by retrieving relevant passages or documents, thereby improving factual consistency and reducing hallucinations. However, most of these systems process text at the level of documents, paragraphs, or sentences, and thus often neglect the underlying *event structures* that shape real-world narratives. This oversight becomes especially problematic in scenarios where temporal information, logical dependencies, and rich cross-event interactions are critical for accurate and coherent text generation.

Event-Centric Challenges. Prior approaches typically treat documents as flat sentences, obscuring how events evolve or interconnect across multiple texts (Fan et al., 2024). Without explicit event modeling, it is difficult to keep track of shared entities and contexts, leading to inconsistent references and incomplete storylines (Arslan et al., 2024). The absence of event-centric organization also diminishes the capacity to perform *multi-event reasoning*, as no explicit structure is in place to link events that are logically or thematically related (Ravi et al., 2024; Zhao et al., 2024).

Temporal-Aware Limitations. Even when relevant textual fragments are retrieved, conventional methods often fail to capture *temporal dynamics*—such as the order, duration, and shifts in time that separate one event from another. Handling temporal information is indispensable for many tasks (Liu et al., 2025): from reconstructing chronologies and inferring logical sequences to understanding the significance of an event within its temporal context. Without explicit temporal awareness, traditional RAG systems struggle to maintain coherent timelines, overlooking how events change over time and how earlier actions set the stage for subsequent developments.

Multi-Event Reasoning. Most existing RAG models also rely on single-pass retrieval or shallow concatenation of passages. Such approaches rarely support iterative or multi-hop queries and are ill-suited for tasks demanding robust *multi-event chain of reasoning*. When multiple events from different sources need to be inter-connected, whether for logical consistency checks or to resolve ambiguities, flat retrieval methods often lead to fragmented and inconsistent narratives. Moreover, the inability to

---

[†] Corresponding author: qiang.zhang.cs@zju.edu.cn

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 16967–16979
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

distinguish authentic and up-to-date information from multiple sources often leads to hallucinations.

Our Approach: EventRAG. In this work, we introduce *EventRAG*, a novel *event-centric* retrieval-augmented generation framework that explicitly organizes textual information into a structured Event Knowledge Graph (EKG) (Section 2.2). Unlike traditional RAG, our approach first identifies key events and represents them as inter-connected nodes. These nodes are then organized with a *temporal-aware* perspective, enabling the system to capture the chronology and temporal dependencies across events. By merging duplicated or semantically equivalent entities and expanding under-connected nodes, EventRAG provides a comprehensive and non-redundant knowledge backbone.

To enable *multi-event reasoning*, EventRAG utilizes an iterative, Agent-based retrieval and inference strategy (Section 2.3). Rather than relying on a single-pass retrieval, our agent autonomously queries the EKG, refines partial inferences, and corrects inconsistencies in a multi-step process. This mechanism allows the system to chain multiple events together, form logical links, and verify critical information, thereby ensuring logical consistency and reducing hallucinations.

Experimental Validation. We evaluate EventRAG on tasks from *UltraDomain* (Qian et al., 2024) and *MultiHopRAG* (Tang and Yang, 2024) benchmarks, demonstrating consistent improvements over baselines. Our results show that EventRAG excels at tasks involving temporal sensitivity (Section 3.1), multi-step inference, and inter-event logical reasoning (Section 3.2). Ablation studies further indicate that 1) entity fusion significantly enhances the completeness and accuracy of the knowledge graph, and 2) multi-event reasoning is vital for capturing long-range dependencies and ensuring self-consistent generation (Section 3.3).

Contributions. In summary, this paper makes the following contributions:

- We propose a novel *event-centric* retrieval-augmented generation framework, EventRAG, which decomposes text into interconnected events and constructs an Event Knowledge Graph.
- We incorporate *temporal-aware* capabilities into EventRAG, enabling the modeling of event chronologies and time-sensitive relationships that are crucial for coherent and context-rich generation.
- We introduce an agent-based retrieval and generation process that facilitates *multi-event reasoning* with autonomous and self-correcting queries, resulting in improved logical consistency and reduced hallucinations.
- Extensive experiments were conducted to evaluate the effectiveness of EventRAG compared to existing RAG models. The results demonstrated significant improvements over baseline methods, especially in scenarios demanding multi-event, temporal-aware inference.

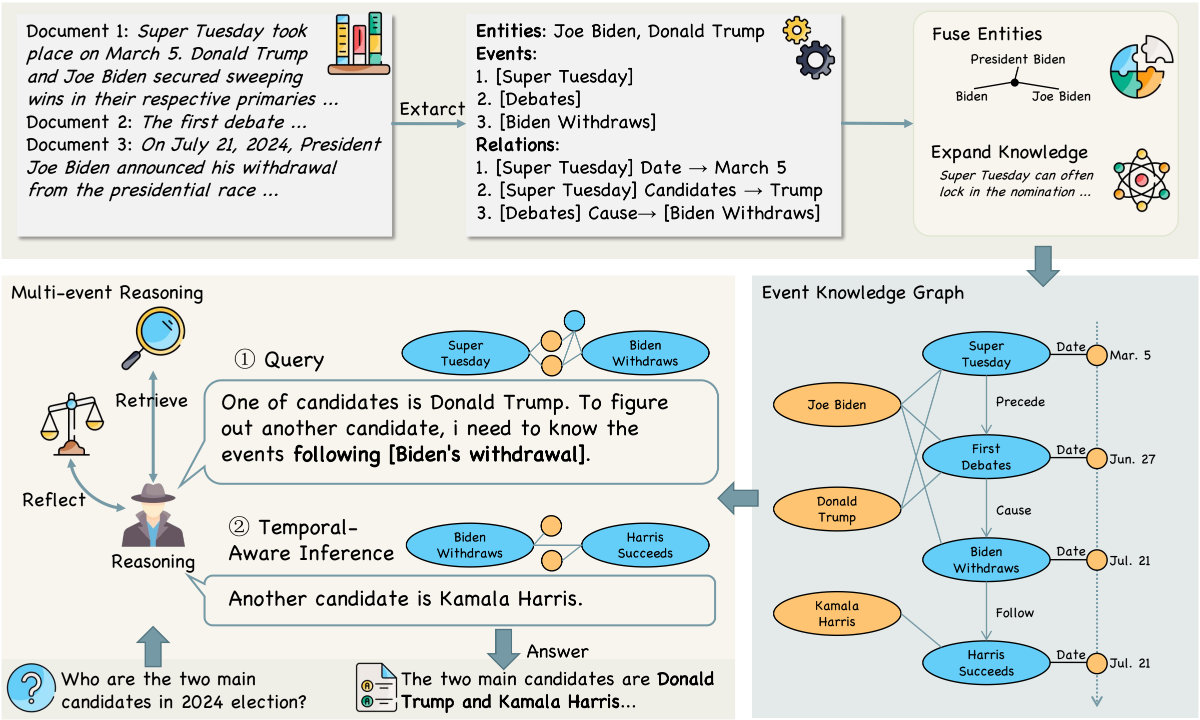
2 EventRAG Framework

2.1 Overall Architecture Design

The EventRAG framework is designed to tackle the challenges associated with narrative-rich documents by emphasizing three critical dimensions: **event-centric**, **temporal-aware**, and **multi-event reasoning**. As shown in Figure 1, the framework consists of two main phases: 1) Event Knowledge Graph Construction and 2) Multi-event Reasoning.

Unlike traditional RAG methods that focus on paragraphs or sentences, we adopt an event-centric approach. In the first phase, EventRAG decomposes text into a series of events, representing key event components—such as participants and temporal markers—as nodes in a knowledge graph, while extracting logical relationships between events as edges. These elements are then organized into an Event Knowledge Graph (EKG). This event-centric representation enables a richer and more structured understanding of the textual information, providing a robust foundation for retrieval, reasoning, and generation tasks. In the second phase, an agent interacts with the EKG to retrieve relevant information, perform multi-step reasoning, and generate outputs based on the structured event knowledge.

By focusing on events, EventRAG provides a more semantically rich and temporally grounded understanding of the text, capturing event interdependencies such as logical relationships and temporal sequences. This approach aims to address complex problems and mitigate hallucinations, while significantly enhancing generation performance in narrative-rich scenarios.



Document 1: *Super Tuesday took place on March 5. Donald Trump and Joe Biden secured sweeping wins in their respective primaries ...*
Document 2: *The first debate ...*
Document 3: *On July 21, 2024, President Joe Biden announced his withdrawal from the presidential race ...*

Entities: Joe Biden, Donald Trump
Events:
1. [Super Tuesday]
2. [Debates]
3. [Biden Withdraws]

Relations:
1. [Super Tuesday] Date → March 5
2. [Super Tuesday] Candidates → Trump
3. [Debates] Cause→ [Biden Withdraws]

Super Tuesday can often lock in the nomination ...

One of candidates is Donald Trump. To figure out another candidate, i need to know the events **following [Biden's withdrawal].**

Another candidate is Kamala Harris.

Who are the two main candidates in 2024 election?

The two main candidates are **Donald Trump and Kamala Harris...**

Figure 1: System architecture of the proposed EVENTRAG framework, containing two main phases: event extraction & graph construction, and agent-based retrieval & generation.

2.2 Event Knowledge Graph Construction

EVENTRAG employs an LLM for comprehensive information extraction from text, focusing on events, entities, and their relationships. The extracted information undergoes dual processing streams: In the first stream, the extracted elements are embedded into dense vector representations and stored in a vector database to facilitate efficient similarity-based retrieval. In the second stream, the extracted information is processed through similarity-based merging to consolidate redundant entries, followed by knowledge expansion which enriches the initial extractions with relevant contextual information. These processed results are then integrated to construct an event-centric knowledge graph that captures the complex interconnections between events, entities, and their relationships.

Fuse Entities. A significant challenge in event extraction arises when entities are repeated across multiple documents or when the same entity is represented through varied expressions (Bugert et al., 2021). To address this, the Entity Fusion mechanism identifies and merges events or entities that are semantically or temporally equivalent but come from different sources. The merging process employs a vector-based similarity-matching approach. Let $\mathcal{V}_i$ and $\mathcal{V}_j$ be vector representations of two newly extracted entries (events, entities, or relationships). The system calculates the similarity between $\mathcal{V}_i$ and existing entries in a vector database. The fusion process can be described as:

$$\mathcal{V}_i \cup \mathcal{V}_j \to \mathcal{V}_f, \quad \text{if similarity}(\mathcal{V}_i, \mathcal{V}_j) > \theta, \tag{1}$$

where $\mathcal{V}_f$ represents the fused entry, and θ is a predefined similarity threshold. We measure the similarity using cosine similarity, which is applied to the dense vector representations obtained from an embedding model. A candidate with the highest similarity score above θ is selected as the merge target. Rather than directly combining the information, the system establishes a "similar" relationship between the original entry and the new one, maintaining provenance while capturing their semantic equivalence. The original entry is then substituted with the new entry in subsequent EKG construction steps, ensuring that all future connections and expansions operate on the merged representation. This approach preserves the original information while reducing redundancy and maintaining the semantic coherence of the knowledge graph.

This step is essential for handling multi-document inputs, where entities may appear in multiple contexts, and ensures that the final EKG remains free from contradictions and inconsistencies.

The goal of entity fusion is not only to eliminate redundancy but also to create a more robust and coherent knowledge structure that supports downstream retrieval and multi-event reasoning.

Expand Knowledge. Upon fusing similar entities, the Event Knowledge Graph needs to be further enhanced by expanding under-connected nodes and relations. This step is crucial to maintaining the graph's comprehensiveness, ensuring it captures all pertinent event-related knowledge, including information that may not be fully represented in the input documents.

The Expand Knowledge process involves tracing gaps in the graph by identifying nodes or relationships that are under-explored. These gaps are filled by utilizing additional contextual clues derived from the documents or leveraging the inherent knowledge of the LLM itself. For instance, events sharing the same participants or occurring in similar temporal contexts can often be linked together, further enriching the graph and enhancing its connectivity. This expansion is key to ensuring that the constructed EKG supports deeper inference and allows for more accurate and contextually aware retrieval when queried by the agent. Fundamentally, knowledge expansion represents a front-loading of cognitive processes, incorporating analytical thinking during the graph construction phase to reduce computational overhead during subsequent retrieval and reasoning operations.

2.3 Event-based Retrieval and Generation

After constructing the enriched EKG, an intelligent agent is employed to perform retrieval and generation tasks. This agent is designed to leverage the event-centric structure of the EKG for sophisticated information processing, moving beyond simple keyword-based retrieval. The agent's workflow is characterized by **iterative multi-event reasoning** and **temporal awareness**, enabling it to address complex queries and generate coherent, contextually grounded outputs.

Autonomous EKG Querying. Upon receiving a query q, the agent initiates an **event-centric** retrieval process by automatically decomposing q into components relevant to event elements (*e.g.,* logical relationships, participants, temporal markers). Let $\mathbf{q}$ be the query vector derived from q. The agent formulates a series of targeted queries to the EKG to retrieve relevant events $e_j \in \mathcal{E}$, where $\mathcal{E}$ is the set of events in the EKG. The retrieval process aims to find the event $\mathbf{e}_j$ in the EKG that maximizes the similarity with the query vector:

$$\mathbf{e}_j = \arg\max_{\mathbf{e}_k \in \mathcal{E}} \text{similarity}(\mathbf{q}, \mathbf{e}_k), \tag{2}$$

Where $\mathbf{e}_k$ denotes the vector representation of an event e_k in the knowledge graph, and similarity$(\cdot, \cdot)$ is a function measuring the similarity between two vectors. This querying process is autonomous, allowing the agent to refine its reasoning based on the context gathered in each step. By focusing on retrieving inter-connected events rather than isolated text snippets, this automatic querying mechanism ensures that retrieved information is highly relevant to the query and maintains the **temporal and relational context** of events. This targeted retrieval is a key enabler for effective **multi-event reasoning**.

Temporal-Aware Inference. EVENTRAG introduces temporal-aware capabilities, enabling the system to capture both static event details and the dynamic evolution of events over time. For two events e_j and e_k, if event e_j precedes event e_k, we can represent this temporal relationship as:

$$T(e_j) \preceq T(e_k), \quad \text{if event } e_j \text{ precedes } e_k, \tag{3}$$

During the EKG construction phase, these temporal markers are processed as attributes of event nodes, and temporal relationships are extracted as edges between events. In the query and reasoning phase, these temporal markers and sequential information are interpreted by LLMs to infer the progression of events, dynamic relationships between entities, and the integration of new events to update existing information.

This capability allows the system to effectively handle tasks involving complex temporal dependencies, such as generating narratives that reflect the natural progression of events, interpreting the significance of events within their temporal context, or predicting outcomes of ongoing processes.

Multi-Event Reasoning. Unlike conventional RAG methods that rely on a single-pass retrieval process, EVENTRAG enables multi-event reasoning through iterative querying and multi-hop inference. The agent begins by identifying an initial set of relevant events from the EKG based on the input query. It then iteratively queries the EKG to retrieve related events, leveraging both logical and temporal relationships encoded in the graph. During each iteration, the agent refines its understanding

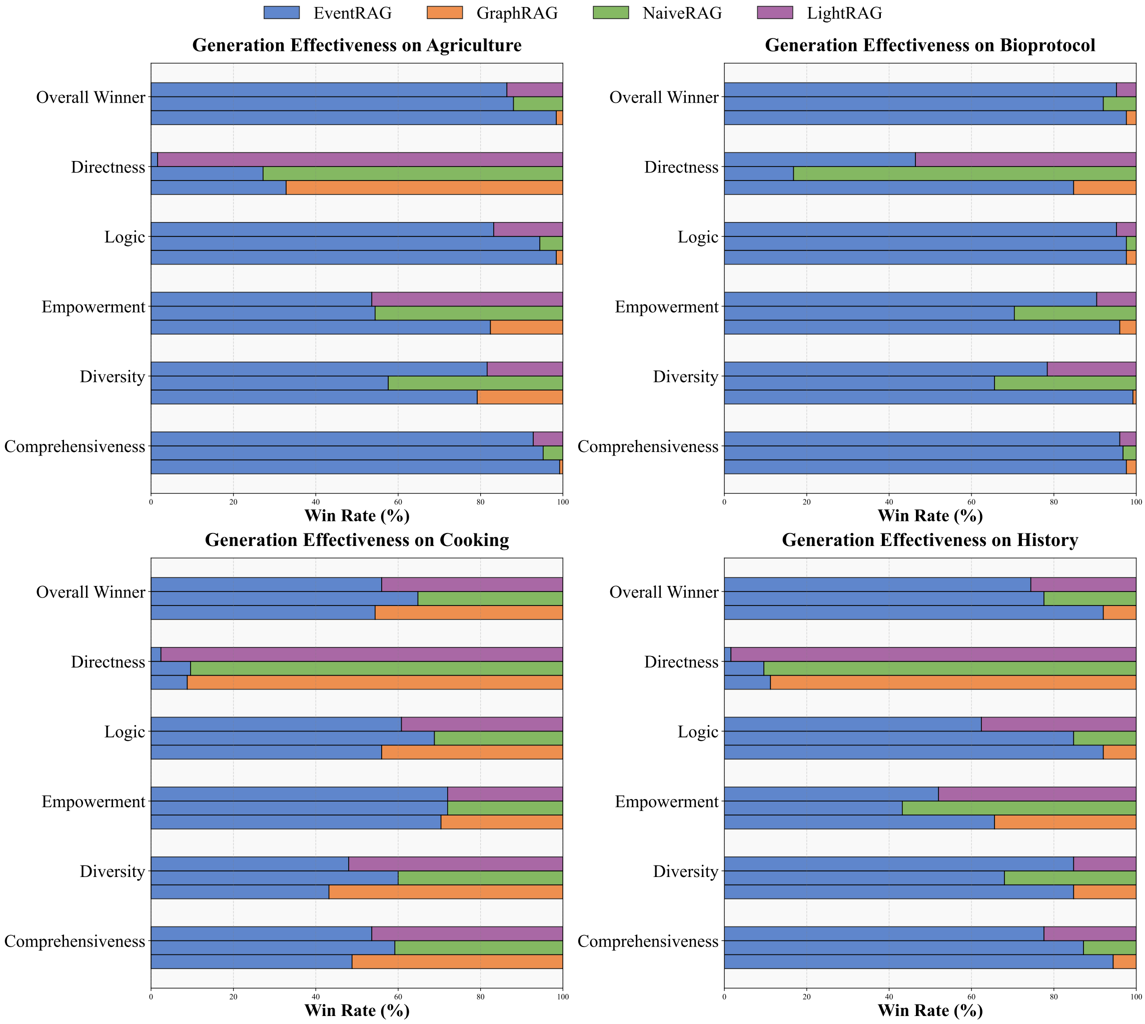


Figure 2: Win rates (%) of baselines v.s. EventRAG across four datasets and six evaluation dimensions.

by chaining together events that are logically or temporally connected, constructing intermediate reasoning paths. These paths are validated against the graph structure to ensure consistency, and the agent dynamically updates its focus based on the newly discovered connections.

The agent utilizes relationships within the EKG to guide its reasoning, ensuring that the generated content captures the nuanced inter-dependencies across events. This mechanism enables EventRAG to handle complex tasks that involve interlinking multiple events from diverse sources, producing outputs that are logically consistent and contextually grounded.

Reflection & Self-Correction. To ensure the robustness and reliability of its reasoning, the agent incorporates a reflection and self-correction mechanism. During the inference process, the agent periodically evaluates its own conclusions for coherence and accuracy, particularly in the context of multi-event reasoning. If inconsistencies or ambiguities are detected, the agent revisits relevant event nodes, re-examines temporal relationships, and potentially explores additional parts of the EKG to refine its understanding. This self-correction loop is vital for mitigating potential errors and hallucinations, especially when dealing with complex queries that require chaining multiple events and maintaining temporal consistency. Finally, after completing the reasoning process and undergoing reflection, the agent generates a final output. This generation process is guided by the EKG, ensuring that the generated text is contextually grounded in the identified events and their relationships. Through iterative refinement of its inferences, the agent guarantees that the final output is logically sound and factually accurate within the event-centric framework.

3 Experimental Evaluation

We conduct extensive experiments to assess the performance of our proposed EventRAG framework.

Query Type	Method	Answer Relevancy	Answer Correctness	Semantic Similarity
Inference Query	NaiveRAG	0.6840	0.2951	0.9077
	GraphRAG	0.7046	0.7580	0.9412
	LightRAG	0.7252	0.8174	0.9158
	EventRAG	**0.8083**	**0.8409**	**0.9637**
Comparison Query	NaiveRAG	0.2115	0.1921	0.7341
	GraphRAG	0.5180	0.4362	**0.8019**
	LightRAG	0.6573	0.5592	0.7913
	EventRAG	**0.6955**	**0.8164**	0.7800
Null Query	NaiveRAG	0.2849	0.2892	0.8237
	GraphRAG	0.3159	0.2430	0.9720
	LightRAG	0.3673	0.2916	0.9518
	EventRAG	**0.3761**	**0.3271**	**0.9751**
Temporal Query	NaiveRAG	0.0624	0.1864	0.7456
	GraphRAG	0.0740	0.4549	0.7754
	LightRAG	0.1829	0.6170	0.7359
	EventRAG	**0.2763**	**0.8786**	**0.7760**
Average	NaiveRAG	0.3107	0.2407	0.8028
	GraphRAG	0.4031	0.4730	0.8726
	LightRAG	0.4832	0.5713	0.8487
	EventRAG	**0.5391**	**0.7158**	**0.8737**

Table 1: Inference capability results on the MultiHopRAG dataset. Best scores are in **bold**.

Our evaluation comprises three major experiments: 1) a *generation effectiveness* study examining how EventRAG handles narrative-rich documents; 2) an *inference capability* assessment focusing on temporal and multi-event reasoning; and 3) an *ablation analysis* investigating the contribution of key components. In all experiments, we compare EventRAG against three baselines: *NaiveRAG*, *GraphRAG* (Edge et al., 2024), and *LightRAG* (Guo et al., 2024).

Implementation Details. In our experiments, we utilize the `milvus` vector database for vector data management and access. For all LLM-based operations, we default to using `gpt-4o-2024-08-06` (OpenAI, 2024). To ensure consistency, the chunk size is set to `1200` across all datasets. Additionally, the gleaning parameter is fixed at `1` for both `GraphRAG` and `LightRAG`. More details can be found in Appendix A.2.

3.1 Generation Effectiveness

Setup. To evaluate the effectiveness of our event-centric approach, we utilize three datasets from the UltraDomain (Qian et al., 2024)—*agriculture*, *cooking*, and *history*—along with a real-world laboratory scenario dataset, *bioprotocol* (ODonoghue et al.). These datasets contain multi-document inputs that require both event understanding and temporal awareness for coherent text generation. Following the evaluation methodology of GraphRAG (Edge et al., 2024), we employ six metrics reflecting both content coverage and coherence:

i) **Comprehensiveness**: Measures the completeness of the generated text in covering salient events and entities. ii) **Diversity**: Assesses lexical and structural variety, ensuring the generation is not repetitive. iii) **Empowerment**: Evaluates how effectively the system illuminates event details that assist the user in understanding or acting upon the information. iv) **Logic**: Gauges the logical flow of the narrative and consistency of event relationships. v) **Directness**: Rates the clarity and succinctness of the generated content. vi) **Overall Winner**: A holistic comparison that indicates the preferred system in pairwise evaluations. More details about evaluation can be found in Appendix A.3.

Results and Analysis. Figure 2 summarizes the results. EventRAG outperforms all baselines across most metrics, with particular improvements in *Comprehensiveness* and *Logic*. These results highlight the effectiveness of the EKG in capturing and relating events across documents. The high performance in *Logic* validates our temporal-aware inference's ability to maintain coherent event relationships. Furthermore, EventRAG consistently provides richer event descriptions and more inspiring content, demonstrating the advantages of our event-centric approach to text generation.

3.2 Inference Capability

Setup. In the second experiment, we focus on multi-hop reasoning and complex inference tasks

Dataset	Method	Comprehensiveness	Diversity	Empowerment	Logic	Directness	Overall Winner
Agriculture	1. EventRAG	**95.2**	57.6	**54.4**	**94.4**	27.2	**88.0**
	NaiveRAG	4.8	42.4	45.6	5.6	72.8	12.0
	2. w/o EK	84.8	48.8	**54.4**	92.8	26.4	86.4
	NaiveRAG	15.2	51.2	45.6	7.2	**73.6**	13.6
	3. w/o MB	72.0	**60.8**	47.2	84.0	62.4	82.4
	NaiveRAG	28.0	39.2	52.8	16.0	37.6	17.6
Bioprotocol	1. EventRAG	**96.8**	**65.6**	**70.4**	**97.6**	16.8	**92.0**
	NaiveRAG	3.2	34.4	29.6	2.4	**83.2**	8.0
	2. w/o EK	74.4	44.8	54.4	92.0	20.0	78.4
	NaiveRAG	25.6	55.2	45.6	8.0	80.0	21.6
	3. w/o MB	91.2	54.4	60.8	63.2	44.8	80.0
	NaiveRAG	8.8	45.6	39.2	36.8	55.2	20.0
Cooking	1. EventRAG	**59.2**	**60.0**	**72.0**	68.8	9.6	**64.8**
	NaiveRAG	40.8	40.0	28.0	31.2	**90.4**	35.2
	2. w/o EK	52.0	45.6	66.4	**77.6**	46.4	63.2
	NaiveRAG	48.0	54.4	33.6	22.4	53.6	36.8
	3. w/o MB	53.6	44.0	31.2	59.2	45.6	50.4
	NaiveRAG	46.4	56.0	68.8	40.8	54.4	49.6
History	1. EventRAG	**87.2**	68.0	43.2	**84.8**	9.6	**77.6**
	NaiveRAG	12.8	32.0	56.8	15.2	**90.4**	22.4
	2. w/o EK	84.0	24.0	52.8	80.0	10.4	69.6
	NaiveRAG	16.0	**76.0**	47.2	20.0	89.6	30.4
	3. w/o MB	85.6	64.0	**79.2**	78.4	65.6	72.8
	NaiveRAG	14.4	36.0	20.8	21.6	34.4	27.2

Table 2: Ablation study results (Win rate in couple). EK=Expand Knowledge, MB=Multi-event Reasoning.

using the *MultiHopRAG* dataset (Tang and Yang, 2024). This dataset contains queries requiring different types of reasoning:

i) **Inference Query**: Demands logical deduction spanning multiple events. ii) **Comparison Query**: Entails comparing event attributes. iii) **Null Query**: Tests the system's ability to correctly identify when no conclusive answer is available. iv) **Temporal Query**: Evaluates understanding of temporal relations among events.

To measure performance, we adopt the **Answer Relevancy**, **Answer Correctness**, and **Semantic Similarity** metrics as defined in *RAGAS* (Es et al., 2024). These metrics jointly capture how precise, contextually fitting, and semantically aligned the system's answers are relative to ground truth.

Results and Analysis. As shown in Table 1, EventRAG delivers superior results for all query types. Specifically, *Answer Correctness* improves by an absolute average margin of 14% over the strongest baseline (*LightRAG*), illustrating the effectiveness of incorporating explicit event structures during retrieval and reasoning. On *Comparison Queries* and *Temporal Queries*, the *multi-event reasoning* mechanism in EventRAG leverages logical chains and cross-event links from the EKG, significantly boosting the *Answer Relevancy* and *Answer Correctness* metric. For *Null Queries*, EventRAG demonstrates fewer incorrect or overconfident answers, attributed to its reflection and self-correction steps that enable it to detect insufficient or contradictory evidence in the EKG.

3.3 Ablation Analysis

Setup. An ablation study was conducted to analyze the impact of two key components: *Expand Knowledge* and *Multi-event Reasoning*.

We remove each component from EventRAG individually and compare the resulting variants against *NaiveRAG* on the same generation task as in Experiment 1. We report the same six metrics to maintain consistency.

Results and Analysis. Table 2 presents the results of the ablation study. Removing *Expand Knowledge* leads to a substantial drop in *Comprehensiveness* and *Diversity*, as this component plays a critical role in capturing additional event details and connecting isolated entities. Disabling *Multi-event Reasoning* diminishes both *Logic* and *Overall Winner* scores, indicating that multi-event inference is essential for understanding complex logical relationships between events and constructing logically consistent responses. However, even in ablated settings, these variants still outperform *NaiveRAG* on most metrics, demonstrating the inherent advantages of event-centric representation in narrative-rich scenarios.

3.4 Case Study

To demonstrate the effectiveness of EventRAG, we present a specific case example from *history* dataset in Figure 3, which includes responses to the same question from both GraphRAG and our framework.

How did public opinions evolve through multiple waves of evacuation during the war?

EventRAG: **Initial Responses to Evacuation** When the first wave began on 1 September 1939, there was strong government and public coordination... 1.5 million people, mostly children and vulnerable individuals, relocated... **Subsequent Waves and Issues** Later waves faced more difficulties... Initial enthusiasm declined, lowering participation rates... **Media and Public Perception** Media, including The Times, influenced opinions by highlighting challenges and successes... **Adaptation and Learning from Experiences** Authorities introduced better billeting and hostels... publications reassured the public... **Conclusion** Public sentiment evolved from early support to disillusionment, then resilience and appreciation...

GraphRAG: **Introduction** WWII evacuations saw shifting public opinions as the government executed mass relocations... **Government and Institutional Efforts** The Ministry of Health coordinated evacuation, shaping public trust... **Personal and Emotional Impacts** Accounts from Joyce Fry, Roy Simon, Walter Hurst showed fear, adaptation, and resilience... **Community and Support Systems** Groups like WVS provided welfare and morale support... **Conclusion** Public perception evolved with improved systems... Fear and criticism gave way to resilience and appreciation...

LLM Evaluation: ... Overall Winner: Answer 1 (EventRAG), Explanation: Overall, Answer 1 is the winner due to its thoroughness in covering various facets of the topic, the diversity of perspectives included, and the logical, detailed narrative that provides a deep understanding of the evacuation processes during World War II. While Answer 2 is commendably direct and concise, it lacks the depth and detail that make Answer 1 more informative and enlightening.

Figure 3: Case study: comparison between EventRAG and the baseline method GraphRAG.

As is shown in Figure 3, EventRAG demonstrates notable advantages over GraphRAG in analyzing the evolution of public opinions during wartime evacuation waves. The comparative analysis reveals that EventRAG excels in several key aspects: 1) more structured temporal organization with clear chronological progression, 2) superior contextual depth supported by specific dates and statistics, 3) more nuanced capture of opinion dynamics from initial enthusiasm to eventual resilience, and 4) stronger integration of primary sources. These findings highlight EventRAG's effectiveness in analyzing complex historical phenomena where public sentiment evolves over time, particularly through its ability to maintain narrative coherence while incorporating detailed evidence.

4 Related Work

Advances in Retrieval-Augmented Generation. Recent developments in retrieval-augmented generation have seen significant advancements in reasoning capabilities and knowledge integration. While previous works, such as SelfRAG (Asai et al., 2023) and Auto-RAG (Yu et al., 2024), have made strides in enhancing reasoning through automatic data synthesis, and Search-o1 (Li et al., 2025) has demonstrated the potential of integrating retrieval into inference processes, these approaches primarily operate on flat text representations. DeepRAG's (Guan et al., 2025) introduction of the Markov decision process with self-consistency represents another notable advancement in the field. However, these systems generally lack the ability to capture and reason about complex event structures and their interconnections, limiting their effectiveness in scenarios requiring temporal and logical understanding. In contrast, EventRAG's event-centric approach explicitly models these crucial relationships, enabling more nuanced and accurate generation across complex, multi-document scenarios.

Information Extraction for RAG Current Information Extraction (IE) approaches often rely on specialized models and annotated datasets for specific extraction tasks (Zhang et al., 2019), which makes them less suitable for RAG scenarios where flexibility and generalization across diverse data sources are required. CollabKG (Wei et al., 2024) unifies multiple IE subtasks—named entity recognition, relation extraction, and event extraction—to support both KG and EKG. Their approach introduces human-machine collaboration mechanisms with LLMs serving as assistants, reducing costs while improving performance. Similarly, OneKE (Luo et al., 2024) offers a dockerized schema-guided system employing LLM agents for knowledge extraction across diverse domains. For temporal relation extraction specifically, Wang et al. (Wang et al., 2023) address faithfulness challenges through counterfactual analysis to mitigate training biases and improve uncertainty estimation. While these approaches demonstrate the evolving capabilities of IE systems, recent research (Zhu et al., 2024) indicates that advanced LLMs can now perform extraction tasks effectively without fine-tuning. EventRAG builds upon this insight by leveraging pre-trained LLMs for EKG construction without task-specific training, prioritizing generality and deployment flexibility over specialized extraction architectures.

Graph-Based Language Models. The application of graph-based representations in language models has emerged as a promising direction for capturing structured information (Chen, 2024). Previous work has explored various integration strategies, including using Graph Neural Networks as prefix processors for graph data (Tang et al., 2024), employing language models to enhance node embeddings (Xie et al., 2023; Liu et al., 2023), and developing fusion techniques for seamless LLM-graph interaction (Li et al., 2023; Brannon et al., 2023). While these approaches demonstrate the value of structured representations, they typically focus on static relationship modeling rather than dynamic event understanding. EventRAG bridges this gap by introducing a specialized Event Knowledge Graph that not only captures traditional entity relationships but also incorporates temporal dynamics and logical chains.

5 Conclusions

In this paper, we presented EventRAG, a novel event-centric framework for Retrieval-Augmented Generation that explicitly organizes textual information into a structured Event Knowledge Graph and employs an agent-based iterative retrieval strategy for enhanced multi-event reasoning. EventRAG incorporates temporal-aware capabilities to capture chronological relationships and dependencies across events, enabling more coherent and context-rich generation. Through its innovative approach to event modeling and multi-step inference, the framework reduces hallucinations and improves logical consistency in generated content. Experimental results demonstrate EventRAG's superior performance, particularly in tasks requiring temporal sensitivity and complex reasoning.

Limitations

A primary limitation of EventRAG lies in its computational efficiency during the knowledge graph construction phase. The framework heavily relies on multiple LLM calls for event extraction, relationship identification, and knowledge expansion. For each document, the system requires separate LLM inferences to extract entities, determine event relationships, fuse entities, and expand knowledge with contextual information. This multi-step extraction process, while comprehensive in capturing event-centric knowledge, results in considerable computational overhead and slower processing times, especially when handling large document collections. The frequent LLM invocations during graph construction may pose challenges for real-time applications or scenarios where computational resources are limited.

Ethical Considerations

While EventRAG demonstrates promising capabilities in event-centric reasoning, we must acknowledge potential risks in its real-world applications. The system's ability to construct comprehensive event timelines and infer relationships between events could lead to premature or oversimplified conclusions when applied to ongoing real-world situations, such as developing news stories or evolving social movements. There is a risk that users might overly rely on the system's temporal reasoning to predict future events or make critical decisions, without considering the full complexity of real-world dynamics and human factors. To address these concerns, we recommend that EventRAG be used as a supplementary tool rather than a primary decision-making system, particularly in sensitive contexts.

Acknowledgment

This work is funded by National Natural Science Foundation of China (U23A20496, 62302433), Zhejiang Provincial Key Research and Development Project of China (2024C01135), Zhejiang Provincial Natural Science Foundation of China (LQ24F020007) and the Ningbo Natural Science Foundation (2024J020). This work was supported by AntGroup.

References

Muhammad Arslan, Saba Munawar, and Christophe Cruz. 2024. Exploring business events using multi-source rag. *Procedia Computer Science*, 246:4534–4540.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Self-reflective retrieval augmented generation. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

William Brannon, Wonjune Kang, Suyash Fulay, Hang Jiang, Brandon Roy, Deb Roy, and Jad Kabbara. 2023. Congrat: Self-supervised contrastive pretraining for joint graph and text embeddings. *arXiv preprint arXiv:2305.14321*.

Michael Bugert, Nils Reimers, and Iryna Gurevych. 2021. Generalizing cross-document event coreference resolution across multiple corpora. *Computational Linguistics*, 47(3):575–614.

Huajun Chen. 2024. Large knowledge model: Perspectives and challenges. *Data Intelligence*, 6(3):587–620.

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

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

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501.

Xinyan Guan, Jiali Zeng, Fandong Meng, Chunlei Xin, Yaojie Lu, Hongyu Lin, Xianpei Han, Le Sun, and Jie Zhou. 2025. Deeprag: Thinking to retrieval step by step for large language models. *arXiv preprint arXiv:2502.01142*.

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

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

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025. Search-o1: Agentic search-enhanced large reasoning models. *arXiv preprint arXiv:2501.05366*.

Yichuan Li, Kaize Ding, and Kyumin Lee. 2023. Grenade: Graph-centric language model for self-supervised representation learning on text-attributed graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2745–2757.

Hao Liu, Jiarui Feng, Lecheng Kong, Ningyue Liang, Dacheng Tao, Yixin Chen, and Muhan Zhang. 2023. One for all: Towards training one graph model for all classification tasks. *arXiv preprint arXiv:2310.00149*.

Qingwen Liu, Zejun Li, Zhihao Fan, Cunxiang Yin, Yancheng He, Jing Cai, Jinhua Fu, and Zhongyu Wei. 2025. Bridging the domain gap in grounded situation recognition via unifying event extraction across modalities. *Data Intelligence*, 7(1):143–162.

Yujie Luo, Xiangyuan Ru, Kangwei Liu, Lin Yuan, Mengshu Sun, Ningyu Zhang, Lei Liang, Zhiqiang Zhang, Jun Zhou, Lanning Wei, et al. 2024. Oneke: A dockerized schema-guided llm agent-based knowledge extraction system. *arXiv preprint arXiv:2412.20005*.

Odhran ODonoghue, Aleksandar Shtedritski, John Ginger, Ralph Abboud, Ali Essam Ghareeb, and Samuel G Rodriques. Bioplanner: Automatic evaluation of llms on protocol planning in biology. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

OpenAI. 2024. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`. Accessed: 2025-01-15.

Hongjin Qian, Peitian Zhang, Zheng Liu, Kelong Mao, and Zhicheng Dou. 2024. Memorag: Moving towards next-gen rag via memory-inspired knowledge discovery. *arXiv preprint arXiv:2409.05591*.

Rahul Ravi, Gouri Ginde, and Jon Rokne. 2024. Pragyan-connecting the dots in tweets. In *International Conference on Advances in Social Networks Analysis and Mining*, pages 338–354. Springer.

Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Lixin Su, Suqi Cheng, Dawei Yin, and Chao Huang. 2024. Graphgpt: Graph instruction tuning for large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 491–500.

Yixuan Tang and Yi Yang. 2024. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries. *arXiv preprint arXiv:2401.15391*.

Haoyu Wang, Hongming Zhang, Yuqian Deng, Jacob Gardner, Dan Roth, and Muhao Chen. 2023. Extracting or guessing? improving faithfulness of event temporal relation extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 541–553.

Xiang Wei, Yufeng Chen, Ning Cheng, Xingyu Cui, Jinan Xu, and Wenjuan Han. 2024. Collabkg: A learnable human-machine-cooperative information extraction toolkit for (event) knowledge graph construction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3490–3506.

Han Xie, Da Zheng, Jun Ma, Houyu Zhang, Vassilis N Ioannidis, Xiang Song, Qing Ping, Sheng Wang, Carl Yang, Yi Xu, et al. 2023. Graph-aware language model pre-training on a large graph corpus can help multiple graph applications. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5270–5281.

Tian Yu, Shaolei Zhang, and Yang Feng. 2024. Auto-rag: Autonomous retrieval-augmented generation for large language models. *arXiv preprint arXiv:2411.19443*.

Tongtao Zhang, Heng Ji, and Avirup Sil. 2019. Joint entity and event extraction with generative adversarial imitation learning. *Data Intelligence*, 1(2):99–120.

Siyun Zhao, Yuqing Yang, Zilong Wang, Zhiyuan He, Luna K Qiu, and Lili Qiu. 2024. Retrieval augmented generation (rag) and beyond: A comprehensive survey on how to make your llms use external data more wisely. *arXiv preprint arXiv:2409.14924*.

Yuqi Zhu, Xiaohan Wang, Jing Chen, Shuofei Qiao, Yixin Ou, Yunzhi Yao, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2024. Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities. *World Wide Web*, 27(5):58.

A Appendix

In this section, we elaborate on the methodologies and experimental settings employed in the EventRAG framework. We provide comprehensive details about our implementation, covering entity extraction, relationship identification, and the integration of large language models (LLMs) in our pipeline.

A.1 Experimental Data Details

Statistics	Total Documents	Total Tokens
Agriculture	12	2,242,790
Bioprotocol	100	61,251
Cooking	14	2,371,289
History	26	5,572,313
MultiHop	138	418,674

Table 3: Statistical information of the datasets.

A.1.1 Dataset Characteristics

Our evaluation encompasses five diverse datasets spanning different domains and complexity levels. As shown in Table 3, these datasets vary significantly in both document count and token volume. The History dataset, despite having only 26 documents, contains the highest token count (5,572,313), indicating longer, more detailed documents. In contrast, the Bioprotocol dataset has 100 documents but relatively fewer tokens (61,251), suggesting more concise, procedure-focused content.

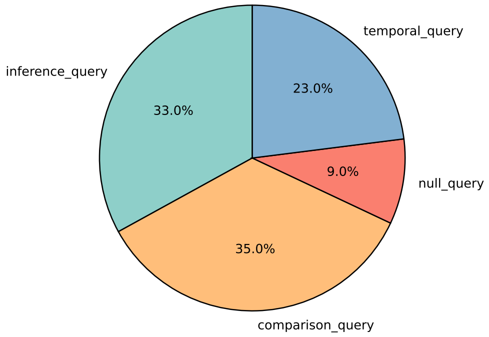


Figure 4: Question type distribution in MultiHop-RAG dataset.

A.1.2 MultiHop-RAG Dataset Analysis

The MultiHop-RAG dataset presents a unique evaluation opportunity due to its diverse question types and complex reasoning requirements. As illustrated in Figure 4, the dataset comprises four distinct question categories. To ensure a balanced and computationally manageable evaluation while maintaining statistical significance, we selected the first 100 questions from the dataset as our evaluation set.

```
Given the following description of a
    dataset:
{total_description}
Please identify 5 potential users who
    would engage withthis dataset. For
    each user, list 5 tasks they
    wouldperform with this dataset.Then,
    for each (user, task) combination,
    generate 5questions that require a
    high-level understanding of theentire
     dataset.
Output the results in the following
    structure:
- User 1: [user description]
   - Task 1: [task description]
      - Question 1:
      - Question 2:
      - Question 3:
      - Question 4:
      - Question 5:
   - Task 2: [task description]
   ...
   - Task 5: [task description]
- User 2: [user description]
...
- User 5: [user description]
...
```

Figure 5: Prompts for generating evaluation questions.

A.2 Implementation Details

A.2.1 LLM Configuration

For our implementation, we utilized the following LLM configurations:

- **LLM**: gpt-4o-2024-08-06 for event extraction, EKG construction, query generation
- **Embedding Model**: text-embedding-3-small for embedding
- **Max Tokens**: 8192 for embedding
- **Temperature**: 0.1 for deterministic outputs
- **Top-p**: 0.95

A.3 Generation Effectiveness Evaluation

A.3.1 Prompts for Query Generation

In the *Generation Effectiveness* experiment, we leveraged LLM to generate 125 questions for each dataset, guided by the prompt illustrated in Figure 5. This approach is inspired by LightRAG, which aids in identifying user roles and their specific goals for engaging with the dataset. It outlines five distinct users, details five key tasks each user would perform, and generates five high-level evaluation questions for each (user, task) pair.

A.3.2 Prompts for LLM Evaluation

The evaluation prompt is illustrated in Figure 6. It introduces a comprehensive evaluation framework for comparing two answers to the same question. Its purpose is to guide the LLM through the process of selecting the better answer for each criterion, followed by an overall assessment.

A.4 Inference Capability Evaluation Metrics

The **Answer Relevancy** metric measures how relevant the generated answer is to the input question. It is computed by generating N artificial questions based on the response and calculating the cosine similarity between the embeddings of these questions and the original input. The final score is the average of these cosine similarity values:

$$\text{Relevancy} = \frac{1}{N}\sum_{i=1}^{N} \text{cosine similarity}(E_{g_i}, E_o) \tag{4}$$

where E_{g_i} is the embedding of the ith generated question, E_o is the embedding of the user input, and N is the number of generated questions.

Answer Correctness combines factual correctness and semantic similarity. Factual correctness is assessed using the F1 score based on True Positives (TP), False Positives (FP), and False Negatives (FN):

$$\text{F1 Score} = \frac{|\text{TP}|}{|\text{TP}| + 0.5 \times (|\text{FP}| + |\text{FN}|)} \tag{5}$$

Semantic similarity between the answer and ground truth is calculated using cosine similarity. The final score is a weighted average of factual correctness and semantic similarity.

The **Semantic Similarity** metric evaluates the semantic alignment between the generated answer and the ground truth. It is computed by embedding both the answer and the ground truth, and then calculating the cosine similarity between their vector representations:

$$\text{cosine similarity}(A, B) = \frac{A \cdot B}{\|A\|\|B\|} \tag{6}$$

where A and B are the embeddings of the answer and the ground truth, respectively. A higher score indicates better semantic alignment.

```
---Role---
You are an expert tasked with evaluating
    two answers to the same question
    based on three criteria: **
    Comprehensiveness**, **Diversity**,
    **Empowerment**, **Directness** and
    **Logic**.

You will evaluate two answers to the same
     question based on five criteria: **
    Comprehensiveness**, **Diversity**,
    **Empowerment**, **Directness** and
    **Logic**.
- **Comprehensiveness**: How much detail
    does the answer provide to cover all
    aspects and details of the question?
- **Diversity**: How varied and rich is
    the answer in providing different
    perspectives and insights on the
    question?
- **Empowerment**: How well does the
    answer help the reader understand and
     make informed judgments about the
    topic?
- **Directness**: How specifically and
    clearly does the answer address the
    question?
- **Logic**: How logical and coherent is
    the answer in its structure and
    reasoning?
For each criterion, choose the better
    answer (either Answer 1 or Answer 2)
    and explain why. Then, select an
    overall winner based on these 5
    categories.
Here is the question:
{query}
Here are the two answers:
**Answer 1:**
{answer1}
**Answer 2:**
{answer2}
Evaluate both answers using the five
    criteria listed above and provide
    detailed explanations for each
    criterion.
Output your evaluation in the following
    JSON format:
{ { "Comprehensiveness": { "Winner": "[
    Answer 1 or Answer 2]", "Explanation
    ": "[Provide explanation here]" }, "
    Empowerment": { "Winner": "[Answer 1
    or Answer 2]", "Explanation": "[
    Provide explanation here]" }, "
    Overall Winner": { "Winner": "[Answer
     1 or Answer 2]", "Explanation": "[
    Summarize why this answer is the
    overall winner based on the five
    criteria]" } } }
```

Figure 6: Prompts for generation effectiveness evaluation.