# Respecting Temporal-Causal Consistency: Entity–Event Knowledge Graphs for Retrieval-Augmented Generation

**Anonymous ACL submission**

## Abstract

Retrieval-augmented generation (RAG) based on large language models often falters on narrative documents with inherent temporal structures. Standard unstructured RAG methods rely solely on embedding-similarity matching and lack any general mechanism to encode or exploit chronological information, while knowledge graph RAG (KG-RAG) frameworks collapse every mention of an entity into a single node, erasing the evolving context that drives many queries. To formalize this challenge and draw the community's attention, we construct ChronoQA, a robust and discriminative QA benchmark that measures temporal, causal, and character consistency understanding in narrative documents (e.g., novels) under the RAG setting. We then introduce *Entity-Event RAG* (E$^2$RAG), a dual-graph framework that keeps separate *entity* and *event* subgraphs *linked by a bipartite mapping*, thereby preserving the temporal and causal facets needed for fine-grained reasoning. Across ChronoQA, our approach outperforms state-of-the-art unstructured and KG-based RAG baselines, with notable gains on causal and character consistency queries. E$^2$RAG therefore offers a practical path to more context-aware retrieval for tasks that require precise answers grounded in chronological information.

## 1 Introduction

Large language models (LLMs) have demonstrated remarkable zero-shot and few-shot capabilities across various NLP tasks. Yet, LLMs remain fundamentally constrained by their fixed context window: once the prompt exceeds a certain length, processing becomes slow and the model's reasoning capability is significantly impaired (Liu et al., 2023; Fiction.live, 2025; Li et al., 2024; Gao et al., 2024; An et al., 2024). To mitigate this limitation, retrieval-augmented generation (RAG) was introduced, coupling the generator with an external document retriever so that fresh, grounded evidence can be injected on demand (Lewis et al., 2020). RAG not only extends an LLM's effective memory but also reduces hallucinations and allows rapid knowledge updates without costly re-training (Shuster et al., 2021; Lewis et al., 2020; Borgeaud et al., 2022; Béchard and Ayala, 2024).

Vanilla unstructured RAG, however, treats each passage in isolation and therefore struggles with reasoning that depends on a document's temporal or causal structure. For example, the following seemingly straightforward question for *Harry Potter* enthusiasts can be challenging for a RAG system to answer reliably:

**Query 1.** *Who was jinxing Harry's broom during his first Quidditch match?*

Because effects in a story typically follow their causes in time, losing chronological ordering also obscures causal links. If a RAG system naively retrieves the seemingly most relevant passage from the earlier part of the story, it will incorrectly conclude that *Snape* was responsible for hexing Harry's broom. This error arises because it overlooks crucial information revealed later—that *Professor Quirrell* was actually causing the interference, while Snape's suspicious actions were attempts to counteract Quirrell's jinx. The key reason is that the document's *chronological context information*, which is crucial for generating answers that are both temporally coherent and causally sound, is not preserved in the vanilla unstructured RAG after being chunked.

To capture richer structure, knowledge-graph (KG) RAG methods such as GraphRAG build an entity graph over the corpus and retrieve communities of related nodes (Edge et al., 2024). Unfortunately, the standard graph-construction pipeline relies on aggressive de-duplication of entity mentions; this collapses distinct temporal or contextual facets of the same character into a single node, discarding

information that is crucial in domains like narrative fiction, where characters evolve continuously. Consider the following query:

**Query 2.** *How would Hermione Granger react if a friend proposed breaking school rules after the troll incident?*

Hermione is introduced as a rule-obsessed know-it-all who reprimands classmates for the smallest infractions (Ch. 9), yet after the troll incident (Ch. 10), she forges a deep friendship with Harry and Ron and is soon helping them brew an illegal potion and sneak past teachers to protect the Stone. If every "Hermione Granger" mention is collapsed into a single KG node, a query will likely retrieve only her predominant rule-abiding persona—the class monitor who scolds Ron for casting *Lumos* in the corridor—while overlooking later chapters in which she calmly steals ingredients and slips past teachers to protect the Stone (Ch. 16). Thus, the aggressive de-duplication of entity mentions in KG-RAG methods can erase her arc from strict disciplinarian to pragmatic rule-breaker, masking precisely the nuance the question probes and leading the system to an outdated, inaccurate answer.

To rigorously define the challenge of temporal-causal consistency in RAG applications and evaluate how current RAG systems handle the aforementioned narrative-focused challenges, we construct **ChronoQA**, a retrieval-style QA benchmark drawn from nine public-domain narrative documents. Every question targets one of eight fine-grained reasoning facets, ranging from causal and character consistency to symbolism and thematic insight.

To address the aforementioned challenges, we propose an *Entity–Event KG* composed of (i) an entity subgraph; (ii) an event subgraph; and (iii) a *bipartite edge set* that maps each entity mention to the events in which it participates, thereby preserving those nuances. Instead of collapsing duplicates, we first extract both entities *and* their associated events, then link every event to the *specific* mention of each participating entity; because we never merge mentions that arise in different parts of the story, each entity node carries its own *context-specific description*. The resulting graph retains multiple, context-specific representations of entities while still exposing the relational structure needed for retrieval. We term the resulting RAG framework, which retrieves over the proposed entity–event KG, *Entity–Event RAG* (E$^2$RAG). At retrieval time, beyond the vanilla pipeline, we introduce an augmentation operation that calls the backbone LLM only once to inject richer context. Empirically, this single-call augmentation lets E$^2$RAG achieve significant gains on fine-grained narrative-reasoning queries, outperforming state-of-the-art unstructured and KG-based RAG baselines and pointing to a practical path toward more context-aware retrieval for LLMs.

The contributions of this work are as follows:

- *We release ChronoQA, the first open, passage-grounded benchmark that stresses temporal and causal reasoning over book-length narratives under a RAG setting.*
- *We propose* E$^2$RAG*, a dual-graph retrieval framework that keeps every entity mention distinct and grounded in its associated event, thereby preserving the causal and temporal consistency of evolving characters and plot.*
- *Extensive experiments demonstrate that variants of* E$^2$RAG *outperform state-of-the-art unstructured and KG-based RAG baselines on ChronoQA.*

## 2 Related Work

### 2.1 Retrieval-Augmented Generation (RAG)

Although today's large language models can store vast amounts of world knowledge, that knowledge is *static*, *unverifiable*, and *expensive* to refresh. Lewis et al. (2020) demonstrated that attaching a live retriever to a generator lets the model pull *up-to-date evidence on demand*, achieving higher accuracy than closed-book BART while returning the very passages that support each claim. Subsequent analyses show that retrieval-augmented language models markedly cut hallucination rates—especially on fact-heavy tasks—because retrieved text provides an external "ground truth" that the decoder can copy or paraphrase instead of guessing (Lv et al., 2024; Sree Mala et al., 2025). Gao et al. (2023) go a step further: their HyDE method lets the LLM hallucinate a "hypothetical" answer, embeds it, and uses that vector to query the index, delivering *precise zero-shot dense retrieval* without relevance labels—an idea we later adapt in Section 4. RAG also sidesteps model-update costs: keeping the index current is far cheaper and faster than re-training or fine-tuning multi-billion-parameter networks, a point underscored by adaptive systems such as Self-RAG and Speculative RAG that retrieve only when necessary and verify drafts to stay both ef-

ficient and factual (Lv et al., 2024; Lewis et al., 2020). Finally, long-context methods like LongRAG show that retrieval can extend an LLM's effective memory without quadratic attention overhead, enabling faithful reasoning over book-length evidence while keeping latency low (Jiang et al., 2024). In short, RAG equips LLMs with a dynamic, interpretable, and cost-effective memory that tackles three core limitations: knowledge staleness, hallucination, and limited context length, challenges that purely parametric models still struggle with.

### 2.2 Knowledge Graph based RAG

GraphRAG (Edge et al., 2024) explicitly builds an entity-level knowledge graph, runs community detection, and then retrieves and summarizes the most query-relevant communities, thereby improving the relevance of the top-$k$ chunks with respect to a query. Follow-up studies highlight three practical drawbacks of this design: (i) the multi-pass entity/relation extraction and community-summary generation make preprocessing costly in both tokens and compute, (ii) traversing and summarizing the graph at inference time adds 2–3× higher end-to-end latency, and (iii) the graph index and its summaries grow super-linearly with corpus size, complicating incremental updates and ballooning memory usage (Wang et al., 2025; Chen et al., 2025; Peng et al., 2024).

LightRAG (Guo et al., 2024) tackles these limitations by folding relational signals into a standard dense index and introducing a dual-level, coarse-to-fine retriever that first selects cluster representatives and then expands to their ego networks; this removes explicit graph traversal, supports incremental index patches, and reduces indexing token cost by $\approx 60\%$ while roughly halving median query latency, all without hurting answer quality on UltraDomain, QFS and other multi-hop QA benchmarks (Guo et al., 2024; Chen et al., 2025).

### 2.3 Long-context narrative benchmarks

Recent work has begun to probe whether LLMs can reason over book-length inputs without truncation. XL$^2$BENCH (Ni et al., 2024), LOONG (Wang et al., 2024), and LONGGENBENCH (Liu et al., 2024) extend QA or generation tasks to 100 K-token contexts, while the very recent FICTION.LIVEBENCH (Fiction.live, 2025) packages full user-written stories into the prompt and asks multi-step comprehension questions that require tracking characters and foreshadowing across tens of thousands of tokens. Because every benchmark above *gives the model the entire story up front*, they measure intrinsic long-context reasoning rather than the retrieval quality.

## 3 New Benchmark: ChronoQA

Although the datasets mentioned above for long context reasoning tasks can be adapted to benchmarking the RAG methods, there are very limited resources that can be used to evaluate a RAG method specifically on the temporal-causal consistency question: **(i)** Most existing datasets focus on general long context tasks, but have limited focus on these challenging reasoning tasks related to temporal-causal consistency. For example, a rare existing task requires reasoning over a character's time-specific state (e.g., "late-story Hermione" versus "early-story Hermione"). **(ii)** Existing benchmarks typically provide *no* passage-level evidence for their ground-truth answers. Different from the long context tasks, which only evaluate the model answers, evaluating the retrieval snippets is also a common metric when benchmarking the RAG systems. However, without the passage-level evidence, extra effort may be required to determine the retrieval stage's correctness. **(iii)** Although some datasets (e.g., Fiction.live (2025)) also focus tasks based on narrative documents, they rely on an entirely private evaluation framework with no publicly available data. These existing issues make it difficult for researchers to verify reported results or to probe where and why current RAG systems fail.

To test whether different RAG systems can reason over a document's *temporal* and *causal consistency*, we introduce **ChronoQA**—a QA benchmark built from nine narrative works spanning novels, musical scripts, and children's stories.[1] For each story we automatically generate questions that probe eight fine-grained reasoning facets:

- *Causal Consistency* – cause and effect, logical sequences, or explanations of how events unfolded.
- *Character & Behavioral Consistency* – character motivations, development, or psychology.
- *Setting, Environment & Atmosphere* – physical locations, time periods, or mood/atmosphere
- *Symbolism, Imagery & Motifs* – symbolic elements, recurring imagery, or metaphorical representations.

[1] Most of the underlying literary works are in the public domain in the United States. The e-book files were obtained from Project Gutenberg and are redistributed under the *Project Gutenberg-tm License* that accompanies each file.

• *Thematic, Philosophical & Moral* – deeper meanings, philosophical ideas, or ethical implications
• *Narrative & Plot Structure* – story organization, plot devices, or narrative techniques.
• *Social, Cultural & Political* – societal contexts, cultural elements, or political dimensions.
• *Emotional & Psychological* – emotional responses, psychological states, or mental processes.

The questions from each category require solid understanding grounded on the progression of the story and cause-and-effect of the elements involved in the plot. Our **ChronoQA** closes the aforementioned gap with:

• **Focus on temporal and causal consistency** — each query hinges on a character's *specific temporal facet* (e.g. "after the troll incident"), so retrieving an early-story snippet fails.
• **Passage-level supervision** — every answer is paired with exact start/end byte offsets, making verification straightforward and reliable.
• **Unrestricted accessibility** — most narrative is drawn from Project Gutenberg, putting the full texts in the public domain. Unlike fully private benchmarks, This guarantees that anyone can obtain, inspect, and redistribute the benchmark data without restrictions.

Thus, ChronoQA tests whether a RAG system can (1) fetch the right snippet when the full story cannot fit into context and (2) reason over evolving entity states throughout the story.

**Dataset statistics.** The final release contains *19* stories and *1028* question–answer pairs. The actual number of question–answer pairs for each story and categories can be found in Table 1 and Table 4. Other details can be found in Appendix A.

Table 1: Questions per reasoning category.

| Category | # Questions |
|---|---|
| Character Consistency | 392 |
| Causal Consistency | 278 |
| Social, Cultural & Political | 74 |
| Thematic, Philosophical & Moral | 72 |
| Symbolism, Imagery & Motifs | 68 |
| Setting, Environment & Atmosphere | 57 |
| Narrative & Plot Structure | 44 |
| Emotional & Psychological | 43 |
| **Total** | **1028** |

**Generation pipeline.** Because every story is short enough to fit within the oracle's context window, we do not chunk the text. Instead we use a *two-stage* pipeline:

• **Q–A candidates generation:** The entire story is provided to multiple oracle models (GPT-o1-pro (Jaech et al., 2024), GPT-o3 (OpenAI, 2025), Claude-3.7-Sonnet (Anthropic, 2025), and Claude-Opus-4.1 (Anthropic)) together with a category list and an instruction to propose diverse question–answer pairs per category (prompt template in Appendix H.1).[2] For each proposed pair, the generating model's answer is retained as provisional ground truth.

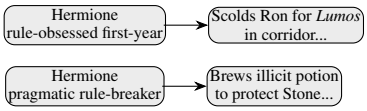


Figure 1: Illustration of two temporally distinct Hermione mentions and their associated events.

• **Verification, Filtering and de-duplication:** For each ground truth, the oracle model is required to output the starting sentence and ending sentence of the excerpt where the answer can be inferred. We discard candidate questions that are duplicates or for which the excerpt extracted does not substantiate the answer.

## 4 Methodology

Standard KG-RAG preprocessing merges every mention of entities extracted from a chunked document (e.g., *"Hermione Granger"*, *"Professor Quirrell"*, or the *"Sorcerer's Stone"*) into a single node—obliterating the time-specific information. Such nuance is needed to answer questions such as Query 2. We therefore keep **each** mention—a concrete, context-specific instance of an entity as it appears in a single chunk of text (e.g., *"Hermione"* in Chapter 9 versus *"Miss Granger"* in Chapter 16) — distinct and tether it to the *event snippet* in which it occurs. Intuitively, the structure is as shown in Figure 1. The left column holds *entity mentions* and their contextual descriptions; the right column holds *events* and their descriptions. Treating the entity as two disjoint vertex sets ensures temporal facets never collapse. Unlike entities, repeated mentions of the same event are deduplicated into a single event node that keeps an occurrence list; entity mentions remain distinct.

**Formal definition. Event schema.** We model each event as $e = \langle \text{trigger}, \text{desc}, \tau \rangle$, where *trigger* is the lexical anchor, *desc* is a one–two sentence description produced by the extractor, and $\tau$ stores narrative position. In our implementation $\tau$ = `source_id` (the chunk id); optional offsets

[2] We include multiple oracles because the strongest publicly available model changes over time; this reduces reduces sensitivity to any single model's quirks and biases.

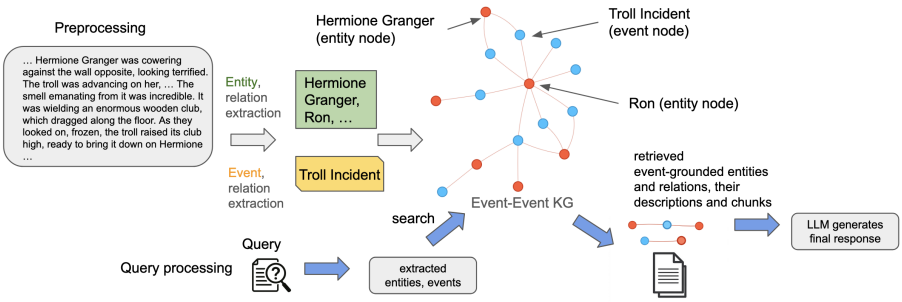


Figure 2: Overall architecture of the E$^2$RAG framework.

(e.g., byte or sentence indices) are kept in metadata.

Let $\mathcal{G}_{\text{ent}} = (V_{\text{ent}}, E_{\text{ent}})$ be the directed graph of entity mentions and $\mathcal{G}_{\text{evt}} = (V_{\text{evt}}, E_{\text{evt}})$ the graph of events. For convenience we write $\tau(e)$ and $\text{chunk}(v)$ for the chunk id; we use $\text{desc}(e)$ and $\text{name}(v)$ for the event description and the entity surface form. The two are connected with an edge set

$$\mathcal{B} = \big\{(v, e) \,\big|\, v \in V_{\text{ent}},\ e \in V_{\text{evt}},\ \text{chunk}(v) = \tau(e), \text{name}(v) \subseteq \text{desc}(e)\big\}. \quad (1)$$

An edge $(v, e) \in \mathcal{B}$ is created when the entity mention and the event come from the *same* chunk ($\text{chunk}(v) = \tau(e)$) and the (normalized) entity surface form appears in the event description. The bipartite graph refers to the inter-layer subgraph induced by $\mathcal{B}$, i.e., $(V_{\text{ent}}^{\mathcal{B}} \cup V_{\text{evt}}^{\mathcal{B}}, \mathcal{B})$ with $V_{\text{ent}}^{\mathcal{B}} = \{v \in V_{\text{ent}} : \exists e\,(v, e) \in \mathcal{B}\}$ and $V_{\text{evt}}^{\mathcal{B}} = \{e \in V_{\text{evt}} : \exists v\,(v, e) \in \mathcal{B}\}$. Details and examples appear in Appendix C.

**Event-event edges.** We maintain typed edges between event nodes when the text in a chunk explicitly or implicitly states a relation (e.g., PRECEDES, CAUSES, COUNTERACTS, REVEALS, PARALLELS). These edges are assigned by the extractor using cue phrases and local semantics and are preserved under event deduplication; each link stores the cue (if any) as metadata. The narrative position $\tau$ is recorded for each event but is not used to generate edges directly; it preserves ordering for later use. We denote the set of typed event-event edges by

$$E_{\text{link}} = \{(e, \rho, e') \mid e, e' \in V_{\text{evt}},\ \rho \in \mathcal{R}_{\text{link}}\}.$$

Here $\mathcal{R}_{\text{link}}$ is an *extensible* label set produced by the extractor (e.g., PRECEDES, CAUSES, COUNTERACTS, REVEALS, PARALLELS); additional fine-grained labels may be emitted and are stored verbatim.

**Temporal modeling and preservation.** Our entity–event KG preserves time through three complementary signals.

*(i) Narrative order.* Every node stores its document position (chapter, chunk id, and byte or sentence span). This yields an intrinsic ordering of mentions across the story without an absolute clock.

*(ii) Typed event-event edges.* When a chunk states a temporal or causal link, the extractor adds a directed edge with a label drawn from an extensible schema (e.g., PRECEDES, FOLLOWS, DURING, CAUSES, COUNTERACTS, REVEALS, PARALLELS); additional fine-grained labels may be emitted and are stored verbatim. Typed event–event edges are assigned by an LLM that uses local cue phrases and semantic context; the raw cue (if any) is kept in metadata. These edges retain relative chronology and causal direction after chunking.

*(iii) Events deduplication.* Multiple textual mentions of the same event are deduplicated to a single event node that stores all its occurrences (positions and spans). In contrast, entity mentions are not merged: each entity node is a context-aware mention linked by a bipartite edge to the event instance it participates in. This preserves temporally distinct facets of a character while still consolidating the underlying event.

**What temporal information is captured?** We capture (a) *explicit* temporal expressions such as "the next morning" or "at dinner" in node or edge metadata; (b) *relative* relations such as before, after, and during via typed edges between events; and (c) *implicit* order through each node's narrative position. We do not normalize to absolute time; the combination of typed relations and narrative order is sufficient for the narrative QA that we target.

**Preprocessing.** E$^2$RAG consists of the following preprocessing steps to build knowledge graphs[3].

[3]Pseudocode can be found as Algs. 1 and 2 in the appendix.

*1. Chunking.* The document is split into chunks $\{c_\ell\}_{\ell=1}^{L}$ and stored in a key–value store.
*2. Creating entity and event sets.* Each chunk $c_\ell$ is fed twice to an LLM to extract *entities* $V_{\text{ent}}^{(\ell)}$ and *events* $V_{\text{evt}}^{(\ell)}$ respectively. For every entity and event extracted, we also ask the LLM for a one-sentence description. Note that in entity's case, this description will be *context-aware* so the same entity extracted from different document chunks will have different descriptions (e.g. "Hermione—rule-obsessed first-year" versus "Hermione—pragmatic rule-breaker"). This description is stored in the node's `description` field and is kept *distinct* for every mention. The extractions form the node sets $V_{\text{ent}} = \bigcup_\ell V_{\text{ent}}^{(\ell)}$ and $V_{\text{evt}} = \bigcup_\ell V_{\text{evt}}^{(\ell)}$. Co-mentions inside the same chunk yield the directed edge sets $E_{\text{ent}}$ and $E_{\text{evt}}$ exactly as in LightRAG.
*3. Link entities to events ($\mathcal{B}$).* For every entity $v \in V_{\text{ent}}^{(\ell)}$ we scan the events $e \in V_{\text{evt}}^{(\ell)}$; if the entity surface form $\text{name}(v)$ appears (exact or normalized) in $\text{desc}(e)$, we add $(v, e)$ to $\mathcal{B}$ (Eq. 1). Because both $v$ and $e$ originate from the same chunk $\ell$, this step enforces the chunk gating $\text{chunk}(v) = \tau(e)$ used in the formal definition.
*4. Embed and index.* Each node's name and description are embedded with embedding model $f(\cdot)$ and written to two vector stores—one for entities, one for events—while the full graph $\mathcal{G}_{\text{ent}} \cup \mathcal{G}_{\text{evt}} \cup \mathcal{B}$ is saved for hop-limited traversals.

Figure 2 shows the overall pipeline of E$^2$RAG and Figure 4 shows a subgraph of the Entity-Event KG of *Harry Potter and the Sorcerer's Stone* and an event node's meta data.

**Query-time retrieval.** Given a query $q$, the retrieval mechanism conducts the following steps.
*1. Cue extraction.* A pair of entity and event phrase sets are first extracted by an LLM extractor $g(\cdot)$, namely $(S_{\text{ent}}, S_{\text{evt}}) = g(q)$
*2. Embedding generation.* The $(S_{\text{ent}}, S_{\text{evt}})$ will be fed into the embedding model and generate a embedding $z$ used for retrieval, e.g., $z = f(g(q))$.
*3. Seed nodes retrieval.* With the embedding $z$, the next step is to retrieve seed nodes $V_q \subseteq V_{\text{ent}} \cup V_{\text{evt}}$ from the vector stores.
*4. One-hop expansion.* We take one step over the chunk-gated bipartite edges $\mathcal{B}$ and over the typed event-event edges $E_{\text{link}}$:

$$\begin{aligned} V_q^+ = V_q &\cup \{\, e \in V_{\text{evt}} \mid \exists v \in V_q : (v, e) \in \mathcal{B} \,\} \\ &\cup \{\, v \in V_{\text{ent}} \mid \exists e \in V_q : (v, e) \in \mathcal{B} \,\} \\ &\cup \{\, e' \in V_{\text{evt}} \mid \exists e \in V_q, \\ &\rho \in \mathcal{R}_{\text{link}} : (e, \rho, e') \in E_{\text{link}} \,\}. \end{aligned} \tag{2}$$

**How temporal information affects retrieval.** After seed selection, we take one hop over the bipartite mapping and over typed event–event edges, which brings in events that the text marks as temporally or causally linked to the seeds (e.g., REVEALS). This ensures that if an early event is retrieved, later clarifying events are also considered.

*5. Similarity ranking and score propagation.*

All passages are embedded offline; at query time we compute similarities only for nodes in $V_q^+$. Event scores are propagated to linked entities via $\mathcal{B}$ and to one-hop neighboring events via typed event–event edges. An entity that participates in a high-scoring event is up-weighted even if the entity mention alone is only moderately similar to the query. Likewise, an event connected by a temporal edge to a high-scoring event inherits part of that score. This re-ranking preserves temporally consistent evidence: entity mentions remain tethered to the time-appropriate events, and later disambiguating events can outrank earlier but misleading ones. The hop count is a tunable hyperparameter; we use one hop in all experiments.

*6. Context assembly.* We collect (i) the raw passages behind the top-$k$ nodes; (ii) a linearized dump of the subgraph $\mathcal{G}_{\text{sub}} = (V_q^+, (E_{\text{ent}} \cup E_{\text{evt}} \cup \mathcal{B} \cup E_{\text{link}}) \cap (V_q^+ \times V_q^+))$; and (iii) the metadata (descriptions, labels, etc.) associated with every selected node and edge. This enriched, structured context is fed to the backbone LLM, allowing it to answer questions that demand fine-grained temporal and causal reasoning using the truly relevant document chunks.

Together, these two routines turn the intuition from Figure 1 into a fully operational retrieval pipeline, with $\mathcal{B}$ acting as the critical "glue" that preserves evolving entity states across time.

**Hypothetical response coupling.** One potential limitation of E$^2$RAG is that the retrieval effectiveness heavily depends on how much information can be extracted from the query. If the query lacks details, the extraction might not be able to capture enough information, preventing the core mechanism from being fully utilized. To make the *entity–event* approach more effective, we incorporate the core idea in HyDE (Gao et al., 2023). HyDE first asks the backbone LLM to draft a *hypothetical response* without the document, then merges the embedding of the hypothetical response with the original query before performing similarity-based retrieval. Even though the content might be factu-

ally off, the inclusion of the hypothetical response provides a richer context for similarity matching.

We introduce four variants of hypothetical response mechanism in E$^2$RAG for retrieving more accurate and comprehensive information. To clearly present them, we use $h$ denote hypothetical response and $[q; h]$ as text-level concatenation of the original query $q$ and hypothetical response $h$.

*1. Combined extraction (Comb. extraction).* Instead of providing the query $q$ to the extractor, this approach gives $[q; h]$ as input to the extractor, and generates embedding as $z = f(g[q; h])$.
*2. Hypothetical extraction (Hyp. extraction)*. Similarly, this variant replaces the original query $q$ with the hypothetical answer $h$ to the extractor, and embeds the output, i.e., $z = f(g(h))$.
*3. Combined embedding (Comb. embedding).* This method omits the extractor, and directly embeds the concatenated hypothetical answer and the original query, i.e., $f([q; h])$.
*4. Hypothetical embedding (Hyp. embedding).* Similar to the above one in terms of removing the extractor step, we directly generate embeddings with the hypothetical documents, i.e., $z = f(h)$.

We experiment with these four variants to investigate the importance of two key factors: incorporating the original query and employing the extraction step. Specifically, we examine how the presence or absence of the original query and the extraction mechanism affect retrieval effectiveness. After generating embeddings using each approach, the following steps are identical to the query-time retrieval steps 3 to 6 introduced earlier.

## 5 Experiments

**Baselines.** We compare the five variants of E$^2$RAG (four hypothetical response variants, one without hypothetical response) against the three modes of LightRAG (Guo et al., 2024) (local, global, hybrid), three modes of GraphRAG (Edge et al., 2024) (local, global, drift), RQ-RAG (Chan et al., 2024), vanilla HyDE (Gao et al., 2023) as well as vanilla RAG. In addition, we also give the hybrid mode of LightRAG the same four variants with the hypothetical response (thirteen in total excluding all the variants of our own methods). All methods use GPT-4o-mini (Hurst et al., 2024) as the backbone LLM for preprocessing, inference, or both, and text-embedding-3-small (OpenAI, 2024) as the embedding model $f(\cdot)$.

Table 2: Overall average scores

| Rank | Mode | Avg Score | Total |
|---|---|---|---|
| 1 | **E$^2$RAG** (comb. extraction) | 6.4000 | 38 581 |
| 2 | **E$^2$RAG** (hyp. embedding) | 6.3500 | 38 425 |
| 3 | **E$^2$RAG** (comb. embedding) | 6.3200 | 38 041 |
| 4 | **E$^2$RAG** (hyp. extraction) | 6.3100 | 38 016 |
| 5 | LightRAG hybrid | 6.2500 | 36 740 |
| 6 | GraphRAG local | 6.2300 | 36 559 |
| 7 | **E$^2$RAG** (vanilla) | 6.1300 | 36 239 |
| 8 | GraphRAG drift | 6.1100 | 35 797 |
| 9 | LightRAG local | 5.9800 | 35 415 |
| 10 | LightRAG global | 5.8200 | 34 347 |
| 11 | GraphRAG global | 5.7500 | 34 219 |
| 12 | vanilla HyDE | 5.5500 | 33 376 |
| 13 | vanilla RAG | 5.3900 | 31 919 |
| 14 | LightRAG hybrid (comb. embedding) | 5.3700 | 31 683 |
| 15 | LightRAG hybrid (comb. extraction) | 5.3700 | 31 235 |
| 16 | LightRAG hybrid (hyp. embedding) | 5.3400 | 31 699 |
| 17 | LightRAG hybrid (hyp. extraction) | 5.3300 | 31 043 |
| 18 | RQ-RAG | 3.3200 | 19 239 |

**Evaluation.** For each query, we give an LLM judge the query, the ground truth and the responses of the variants of E$^2$RAG and all baselines. Each answer is graded independently by $J = 7$ LLM judges (Claude-3.7-Sonnet, Claude-opus-4.1, gemini-2.5-pro (Google), Qwen-max (Alibaba Cloud), GPT-4o (OpenAI, 2024) and GPT-4.1-mini (OpenAI, 2025); check Appendix N for the specific checkpoints used) on a 1–10 Likert scale using the rubric in Appendix H.2. The overall quality score for a system is the mean of those ratings across all samples and judges, computed as $\text{Score} = \frac{1}{J}\sum_{j=1}^{J}\frac{1}{N}\sum_{i=1}^{N} s_{ij}$, where $N$ is the number of question–answer pairs in the test set; $J$ is the number of LLM judges; and $s_{ij} \in \{1, \ldots, 10\}$ is the score assigned by judge $j$ to sample $i$. We report the (*mean score*) in Table 2; per-judge scores are provided in Appendix K.

**Results.** In the overall ranking, the top-3 modes are all hypothetical response variants of E$^2$RAG, while E$^2$RAG (vanilla) loses to GraphRAG local and LightRAG hybrid mode. This corroborates that hypothetical responses play a major role in improving retrieval quality. In particular, the superior performance of the combined extraction variant, followed by the combined embedding variant, confirms that both incorporating the original query and employing the extraction step further enhance the retrieval effectiveness. In contrast, all hypothetical response variants of LightRAG hybrid mode performed poorly. We speculate that because E$^2$RAG does not employ de-duplication, it synergizes particularly well with the hypothetical response, achieving top performance regardless of the specific variant used. Furthermore, the hypothetical response provides a richer context and an abundant number of entity and event candidates. This is beneficial to the matching when there are

Table 3: Preprocessing runtimes for the three KG RAG systems

| System | Stage | Time (s) |
|---|---|---|
| LightRAG | Total pipeline | 101.1723 |
| E$^2$RAG | Entity subgraph construction | 108.3549 |
| | Event subgraph construction | 103.6091 |
| | Bipartite-mapping | 0.5076 |
| | **Total pipeline** | 108.8625 |
| GraphRAG | Total pipeline | 208.9904 |

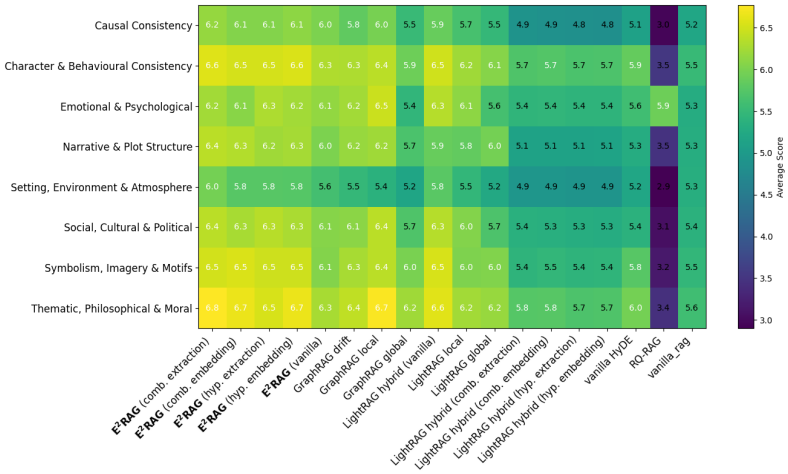


Figure 3: Heatmap of the scoring for each category and mode.

event nodes to provide grounding to the entities to mitigate hallucination. Figure 3 plots the category-average scores (averaged over the three evaluator LLMs). In every category, the top-scoring model is again an E$^2$RAG hypothetical response variant. More details can refer to Appendix K.

**Preprocessing time cost.** Table 3 shows the preprocessing time of the novel *The Phantom of the Opera* for three KG based RAG approaches used in our experiments.

Compared to LightRAG, although E$^2$RAG needs to construct two subgraphs $(\mathcal{G}_{\text{ent}}, \mathcal{G}_{\text{evt}})$ as opposed to just an event KG, and form $\mathcal{B}$ (Bipartite mapping), the two subgraphs' constructions are independent and can be carried out in parallel. Forming $\mathcal{B}$ is sequential after the subgraphs' construction, but for $(\mathcal{G}_{\text{ent}}, \mathcal{G}_{\text{evt}})$ of reasonable size such as in this example, its cost is negligible in practice. As a result, E$^2$RAG has preprocessing time comparable to LightRAG (subject to API query traffic fluctuation), whereas GraphRAG is noticeably slower.

**Token Cost Analysis.** During the KG construction phase, since event extraction is done separately from entity extraction, E$^2$RAG has two times the number of API calls of LightRAG with same max token count $C_{\text{extract}}$, which doubles its worst-case token cost during preprocessing. On the other hand, as the author of LightRAG pointed out, GraphRAG still has much higher token consumption when it comes to KG construction due to massive, repeatedly generated community reports (Guo et al., 2024) (refer to Table 6 for the cost on preprocessing *The Phantom of the Opera*). During retrieval, to ensure a fair comparison, we set the max token $C_{\text{output}}$ for the retrieved chunks to be the same for every KG-based RAG mode; therefore, they all have the same token cost. For every query, the hypothetical response step contributes to exactly one additional API call with the same $C_{\text{output}}$ allowed per API call, which doubles the output token count in the worst case.

**Case study.** In Table 7, we showcase the responses of selected modes used for evaluation for a particular query from *The Hound of the Baskervilles* and the verdicts given by GPT-4.1-mini. The complete verdicts of all three LLM judges can be found in Appendix J.3. Overall, pairing with hypothetical response, E$^2$RAG is able to retrieve contextually relevant chunks—rather than surface-level matches—yielding more thorough and accurate answers. The details on the extraction and retrieval of E$^2$RAG (comb. extraction) mode can be found in Appendix J. Due to the space constraints, complete responses of other modes are provided in the supplementary materials.

## 6 Conclusion

E$^2$RAG keeps every entity mention separate and anchors it to the exact events in which it appears, restoring the temporal and causal context that vanilla RAG and deduplicated KG variants fail to capture. On the new ChronoQA benchmark—designed specifically to test narrative, causal, and character-consistency reasoning—it delivers the best overall and category-specific scores while matching LightRAG's preprocessing cost and latency. Because it builds on off-the-shelf extraction prompts and vector stores, requires no model fine-tuning, and pairs naturally with hypothetical response, E$^2$RAG offers a drop-in upgrade for more faithful retrieval and a foundation for future work on RAG for complex question-answering tasks.

## Limitations

The proposed E$^2$RAG framework focuses on improving the response quality of the retrieval-augmented generation on documents that possess inherent temporal or causal structures—for instance, novels, short stories, scripts, and other narrative-style texts in which characters evolve and events unfold in sequence. Outside of this scope, where information lacks such structures, E$^2$RAG may not offer noticeable benefits compared to methods specifically designed for those contexts.

## References


Alibaba Cloud. Qwen llms. See the "Qwen-Max" section for model details and the snapshot alias.

Chenxin An, Jun Zhang, Ming Zhong, Lei Li, Shansan Gong, Yao Luo, Jingjing Xu, and Lingpeng Kong. 2024. Why does the effective context length of llms fall short? *arXiv preprint arXiv:2410.18745*.

Anthropic. System card addendum: Claude opus 4.1.

Anthropic. 2025. Claude 3.7 sonnet system card. https://www.anthropic.com/claude-3-7-sonnet-system-card. System card, accessed 2025-05-18.

Patrice Béchard and Orlando Marquez Ayala. 2024. Reducing hallucination in structured outputs via retrieval-augmented generation. *arXiv preprint arXiv:2404.08189*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Chi-Min Chan, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, and Jie Fu. 2024. Rq-rag: Learning to refine queries for retrieval augmented generation. *arXiv preprint arXiv:2404.00610*.

Boyu Chen, Zirui Guo, Zidan Yang, Yuluo Chen, Junze Chen, Zhenghao Liu, Chuan Shi, and Cheng Yang. 2025. Pathrag: Pruning graph-based retrieval augmented generation with relational paths. *arXiv preprint arXiv:2502.14902*.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, Ralph M Weischedel, et al. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, pages 837–840. Lisbon.

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

Fiction.live. 2025. Fiction.livebench: The first real-world long context benchmark for writers. Accessed: 2025-05-10.

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

Muhan Gao, TaiMing Lu, Kuai Yu, Adam Byerly, and Daniel Khashabi. 2024. Insights into llm long-context failures: When transformers know but don't tell. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7611–7625.

Google. Gemini 2.5 pro: Model card.

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

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

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Ziyan Jiang, Xueguang Ma, and Wenhu Chen. 2024. Longrag: Enhancing retrieval-augmented generation with long-context llms. *arXiv preprint arXiv:2406.15319*.

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.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. 2024. Long-context llms struggle with long in-context learning. *arXiv preprint arXiv:2404.02060*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Xiang Liu, Peijie Dong, Xuming Hu, and Xiaowen Chu. 2024. Longgenbench: Long-context generation benchmark. *arXiv preprint arXiv:2410.04199*.

Qitan Lv, Jie Wang, Hanzhu Chen, Bin Li, Yongdong Zhang, and Feng Wu. 2024. Coarse-to-fine highlighting: Reducing knowledge hallucination in large language models. *arXiv preprint arXiv:2410.15116*.

Xuanfan Ni, Hengyi Cai, Xiaochi Wei, Shuaiqiang Wang, Dawei Yin, and Piji Li. 2024. Xl $^2$ bench: A benchmark for extremely long context understanding with long-range dependencies. *arXiv preprint arXiv:2404.05446*.

OpenAI. 2024. GPT-4o system card.

OpenAI. 2024. text-embedding-3-small: Openai embedding model. https://platform.openai.com/docs/models/text-embedding-3-small. Model release announced 25 Jan 2024.

OpenAI. 2025. Introducing gpt–4.1 in the api. https://openai.com/index/gpt-4-1. Accessed: 20 May 2025.

OpenAI. 2025. OpenAI o3 and o4-mini System Card.

Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. 2024. Graph retrieval-augmented generation: A survey. *arXiv preprint arXiv:2408.08921*.

James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. 2003. Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34.

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. Iso-timeml: An international standard for semantic annotation. In *LREC*, volume 10, pages 394–397.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*.

Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. 2015. From light to rich ere: Annotation of entities, relations, and events. In *Proceedings of the 3rd workshop on EVENTS: Definition, detection, coreference, and representation*, pages 89–98.

Chandana Sree Mala, Gizem Gezici, and Fosca Giannotti. 2025. Hybrid retrieval for hallucination mitigation in large language models: A comparative analysis. *arXiv e-prints*, pages arXiv–2504.

Minzheng Wang, Longze Chen, Cheng Fu, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, et al. 2024. Leave no document behind: Benchmarking long-context llms with extended multi-doc qa. *arXiv preprint arXiv:2406.17419*.

Shu Wang, Yixiang Fang, Yingli Zhou, Xilin Liu, and Yuchi Ma. 2025. Archrag: Attributed community-based hierarchical retrieval-augmented generation. *arXiv preprint arXiv:2502.09891*.

# A More ChronoQA Details

Table 4: Questions per story (total = 1028).

| Story | # Questions |
|---|---|
| Dracula | 88 |
| The Wonderful Wizard of Oz | 82 |
| The Hound of the Baskervilles | 72 |
| The Phantom of the Opera | 70 |
| The Mysterious Affair at Styles | 69 |
| A Study in Scarlet | 67 |
| The Sign of the Four | 62 |
| The Secret Garden | 61 |
| Lady Susan | 60 |
| The Sorrows of Young Werther | 58 |
| Harry Potter and the Chamber of Secrets | 55 |
| Pride and Prejudice | 54 |
| Dangerous Connections | 51 |
| The Diary of a Nobody | 39 |
| The Adventures of Sherlock Holmes | 34 |
| Harry Potter and the Sorcerer's Stone | 30 |
| The Picture of Dorian Gray | 27 |
| Les Misérables | 25 |
| Anne of Green Gables | 24 |
| **Total** | **1028** |

**Format and release.** Each record is a JSON line with fields `{ "story_id", "story_title", "question_id", "category", "question", "ground_truth", "passages" }`, where `passages` has subfields `{ "start_sentence", "end_sentence", "start_byte", "end_byte", "excerpt"}` which pinpoint the relevant passage(s) from the story that provide the supporting evidence of the ground truth. The generation prompt can be found in Appendix H.1.

# B Temporal Typing Details

The extractor labels event–event edges with a small set of relation types using local cues (e.g., "before", "after", "during", "because", "to counter") and paraphrases; the raw cue phrase is stored in edge metadata. We do not construct a global timeline; ordering relies on typed relations plus each node's narrative position.

# C Formal Event Model and Mapping

**Event tuple and occurrence list.** Each event is $e = \langle \text{trigger}, \text{desc}, \tau \rangle$, where *trigger* is the lexical anchor and *desc* is the extractor's short description. $\tau$ stores narrative position where $\tau =$ `source_id` (chunk id); optional byte or sentence offsets are stored as metadata. When a later passage describes the same situation, we deduplicate it to the existing event node and append the new occurrence's $\tau$ to an occurrence list.

**Chunk-gated entity–event edges.** Entity mentions are kept at the mention level and encode their chunk id in the node name (suffix "_chunk-$k$"); let $\text{chunk}(v)$ parse this id. We link an entity mention to an event iff they come from the same chunk and the entity name is present in the event description:

$$\mathcal{B} = \{(v, e) \mid \text{chunk}(v) = \tau(e),\ \text{Name}(v) \subset \text{desc}(e)\}.$$

**Typed event–event edges.** We add a directed link $e \xrightarrow{\rho} e'$ with $\rho \in \mathcal{R}_{\text{link}}$ when a chunk states or implies such a relation; $\mathcal{R}_{\text{link}}$ is an extensible label set produced by the LLM extractor (e.g., PRECEDES, CAUSES, COUNTERACTS, REVEALS, PARALLELS). The raw cue (if any) is stored in edge metadata. edges attached to duplicate event mentions are re-routed to the canonical node during deduplication. For reproducibility, we snapshot the extracted edge labels; retrieval treats all labels uniformly as connectivity (no label-specific weights).

**Effect on retrieval.** Seed nodes are selected by semantic similarity. One-hop expansion over $\mathcal{B}$ preserves the time-specific grounding of entities (only same-chunk events are considered), while typed event edges bring in later clarifying or causally connected events. Event scores then lift their linked entities and one-hop neighboring events, allowing the correct temporal slice to outrank earlier but misleading evidence.

**Concrete illustration.** For chunk $k = 12$, suppose the extractor yields: $e_1 = \langle \textit{drew}, \text{“Hermione draws her wand...”}, \tau = 12 \rangle$ and entity mentions $v_{\text{Hermione},12}$, $v_{\text{wand},12}$. Then $(v_{\text{Hermione},12}, e_1), (v_{\text{wand},12}, e_1) \in \mathcal{B}$. If a later chunk $k = 17$ contains $e_2 = \langle \textit{quivered}, \text{“At Hogwarts, the air quivered...”}, \tau = 17 \rangle$ and states that $e_2$ *reveals* the cause of $e_1$, we add $e_2 \xrightarrow{\text{REVEALS}} e_1$.

**Remark (relation to event formalisms).** Event representation has a rich literature (e.g., ACE, TimeML/-TIMEX3, and Rich ERE). Our design intentionally adopts a minimal schema (trigger+desc+$\tau$) tailored to RAG retrieval and efficient indexing rather than full argument structures or global temporal normalization (Doddington et al., 2004; Pustejovsky et al., 2003, 2010; Song et al., 2015).

## D Query Time Cost

Table 5 shows the average query time for each mode. GraphRAG drift takes the longest time. E$^2$RAG (vanilla) takes marginally longer time compared to LightRAG hybrid mode, likely due to the time taken for searching the additional events on KG. The hypothetical response variants all took slightly longer than their respective base form due to the additional step of generating the hypothetical response. In particular, the combined extraction variant took the longest time due to the extraction process. Nevertheless, the overall query time difference for LightRAG and E$^2$RAG is insignificant with or without hypothetical response, and much shorter compared to all GraphRAG modes.

Table 5: Average query time per method (seconds)

| **Method** | **Avg. Time** |
|---|---|
| GraphRAG drift | 93.154 7 |
| GraphRAG global | 26.232 6 |
| GraphRAG local | 18.014 1 |
| RQ-RAG | 9.926 4 |
| **E$^2$RAG** (comb. extraction) | 8.720 0 |
| **E$^2$RAG** (hyp. embedding) | 8.085 0 |
| **E$^2$RAG** (comb. embedding) | 8.054 4 |
| **E$^2$RAG** (hyp. extraction) | 8.011 6 |
| **E$^2$RAG** (vanilla) | 7.506 5 |
| LightRAG hybrid (comb. extraction) | 7.492 4 |
| LightRAG hybrid (hyp. extraction) | 7.137 9 |
| LightRAG hybrid (hypo. embedding) | 6.714 2 |
| LightRAG hybrid (comb. embedding) | 6.695 1 |
| LightRAG hybrid (vanilla) | 6.156 4 |
| LightRAG local | 5.235 1 |
| vanilla HyDE | 2.910 6 |
| LightRAG global | 1.903 1 |
| vanilla RAG | 1.422 4 |

## E Preprocessing Token Cost

Table 6: Token usage statistics for the three KG-RAG systems

| System | Input tokens | Output tokens | Total tokens |
|---|---|---|---|
| LightRAG | 343 437 | 112 840 | 456 277 |
| **E$^2$RAG** | 598 822 | 208 797 | 807 619 |
| GraphRAG | 880 804 | 440 401 | 1 321 205 |

# F Case Study

In Table 7, we showcase the responses of selected modes used for evaluation for a particular query from *The Hound of the Baskervilles* and the verdicts given by GPT-4.1-mini. The complete verdicts of all three LLM judges can be found in Appendix J.3. Overall, pairing with hypothetical response, E$^2$RAG is able to retrieve contextually relevant chunks—rather than surface-level matches—yielding more thorough and accurate answers.

Table 7: GPT-4.1-mini's verdicts (selected) of the responses to *"Consider the dinner scene in which Holmes details his reasoning about the Lauriston Gardens mystery. How does Watson's narration highlight Holmes's eagerness to explain the logic step by step, and what rhetorical strategies (quoted or paraphrased) does Holmes use to underscore each clue's significance?"*

| Mode | Average Score | Reason |
|---|---|---|
| **E$^2$RAG** (comb. extraction) | 7.0000 | Very thorough and accurate, captures Watson's narration of Holmes's eagerness and detailed rhetorical strategies with direct quotes and logical progression, closely matching the ground truth. |
| LightRAG hybrid (comb. extraction) | 6.3333 | Captures Holmes's eagerness and rhetorical strategies well, including analogies, contrasts, and causal reasoning, with some direct quotes. However, it misses some of the specific step-by-step pacing and the theatrical, emphatic phrasing that the ground truth emphasizes, resulting in a somewhat less vivid depiction. |
| **E$^2$RAG** (vanilla) | 6.0000 | Detailed and faithful to the ground truth, includes Holmes's systematic approach, direct quotes, and rhetorical strategies, capturing both Watson's narration and Holmes's methodical explanation. |
| LightRAG hybrid | 6.0000 | Provides a thorough analysis of Watson's narration and Holmes's rhetorical strategies with examples and direct quotes, closely aligning with the ground truth's emphasis on step-by-step logic and Holmes's delight in explaining. |
| GraphRAG drift | 6.0000 | Mentions Comprehensive and detailed, covers Watson's narration and Holmes's rhetorical strategies including analogies, rhetorical questions, and historical context, closely matching the ground truth. |
| LightRAG local | 5.3333 | Describes Watson's narration and Holmes's rhetorical strategies like analogies and highlighting contradictions, but lacks specific direct quotes and detailed step-by-step logic as in the ground truth. |
| vanilla RAG | 4.3333 | Captures Holmes's eagerness and details several rhetorical strategies with some direct quotes and examples, but includes some inaccuracies and extraneous content not directly related to the Lauriston Gardens dinner scene. |
| vanilla HyDE | 3.6667 | Captures Holmes's eagerness and rhetorical strategies with some direct quotes and analogies, but less focused on the step-by-step logic and specific clues as in the ground truth. |
| RQ-RAG | 3.0000 | Very Very brief and vague; mentions Holmes's eagerness and rhetorical questions but lacks detail, examples, or direct quotes to support the answer. |

# G Hyperparameters

We implement E$^2$RAG within LightRAG's existing codebase. For all experiments, we standardize the chunk size, max token for backbone LLM generation, and chunks appended to be the same as the default setting as LightRAG. We experimented on changing their values but observed no noticeable benefit.

# H Prompts

## H.1 Data Generation Prompt

```
I want to test a few rag systems on their reasoning capabilities and the capability to pick up
nuanced details. Use the document shown below, design queries for it, also provide the ground truth
for each query. In addition, it should focus on testing the RAG system's causal consistency. for
example, 1. it can ask the RAG system how would a certain character behavior in a given context/even
from the document, and see if the answer is consistent with the character's personality/traits at
that specific point in time (character can experience development so their response to the same
thing can very). 2. it can test if the RAG system confuse the event that has not happened but in the
document with the query (the future events that has not happen should not be account for in the
answer). Give the queries and ground truth in json format.the document is here:

DOCUMENT

Focus on Causal Consistency: Apart from character behavior over time and future event confusion,
other causal consistency tests (e.g., testing how well it understands cause-and-effect relationships
in the plot) can also be included. Query Difficulty: query should involve complex, nuanced
reasoning/understanding of the document provided.
Format example:
{
  "queries": [
    {
      "query": "During Harry's first night at Hogwarts, the Gryffindor students climb the moving
      staircases. How does the text describe the corridors and stairways' magical behavior, and what
      is the immediate impact on Harry's sense of direction?",
      "ground_truth": "The staircases sometimes change direction, doors can vanish or move, and some
      require a password or a specific tickle of a doorknob to open. This constant shifting confuses
      new students like Harry, making it easy to get lost early on.",
```

```
      "start_sentence": "The staircases at Hogwarts are famous for moving unexpectedly, often
      depositing unwary students on entirely different floors than intended.",
      "end_sentence": "This enchantment leaves first-years such as Harry feeling hopelessly lost
      during their first nights in the castle.",
      "type": "Causal Consistency"
    }
  ]
}

Rules for the `start_sentence` and `end_sentence` strings:
• They must be *identical substrings* of the document (case-sensitive, byte-for-byte).
• Preserve every original character: spaces, line-breaks, hyphens, quotation marks, etc.
• If the document contains line breaks, represent them in JSON as the two-character sequence.
• Do not add, delete, or normalise any characters—copy-paste only.
• The passage between the two sentences must support the ground-truth answer you give.

It is also important to note that do not explicitly disclose the title/chapter/section number from
which the context of the question is used. Simply described the related event and the characters
invoved to make the context clear.
```

## H.2 Responses Evaluation Prompt

```
You are an expert evaluator of retrieval-augmented generation (RAG) answers.

Scoring rubric (10-point scale):
10 - Matches ground truth exactly or with faithful paraphrase.
7  - Mostly correct; minor omissions or wording differences.
5  - Partially correct; major missing points or inaccuracies.
3  - Mostly incorrect; small overlap.
1  - Off-topic or hallucinated.

Return **only** a valid JSON array, no markdown fences, in this exact shape:
[
  {"mode": "mode\_name", "reason": "short rationale", "score": 9},
  ...
]

If you cannot produce the JSON array, return an object like:
{"error": "description"}.
```

## H.3 Entity, Event Extraction Prompt

```
---Role---

You are a helpful assistant tasked with identifying entities and events in the user's query.

---Goal---

Given the query, list both entities and events. Entities are people, places, organizations, or
objects mentioned in the query, while events are actions, occurrences, or happenings that take place.

---Instructions---

- Output the entities and events in JSON format.
- The JSON should have two keys:
  - "entities" for people, places, organizations, or objects.
  - "events" for actions, occurrences, or happenings.

######################
-Examples-
######################
Example 1:

Query: "How did Napoleon's invasion of Russia affect his empire's strength?"
################
Output:
{{
  "entities": ["Napoleon", "Russia", "Napoleon's empire"],
  "events": ["invasion of Russia", "empire's decline"]
}}
```

```
#############################
Example 2:

Query: "What role did MIT scientists play in the Manhattan Project?"
################
Output:
{{
  "entities": ["MIT", "MIT scientists", "Manhattan Project"],
  "events": ["scientific research", "atomic bomb development"]
}}
#############################
Example 3:

Query: "How did the Industrial Revolution change London's population?"
################
Output:
{{
  "entities": ["London", "London's population", "Industrial Revolution"],
  "events": ["population growth", "urbanization", "industrial development"]
}}
#############################
-Real Data-
######################
Query: {query}
######################
Output:

"""

PROMPTS["naive_rag_response"] = """---Role---

You are a helpful assistant responding to questions about documents provided.


---Goal---

Generate a response of the target length and format that responds to the user's question,
summarizing all information in the input data tables appropriate for the response length and format,
and incorporating any relevant general knowledge.
If you don't know the answer, just say so. Do not make anything up.
Do not include information where the supporting evidence for it is not provided.

---Target response length and format---

{response_type}

---Documents---

{content_data}

Add sections and commentary to the response as appropriate for the length and format. Style the
response in markdown.
```

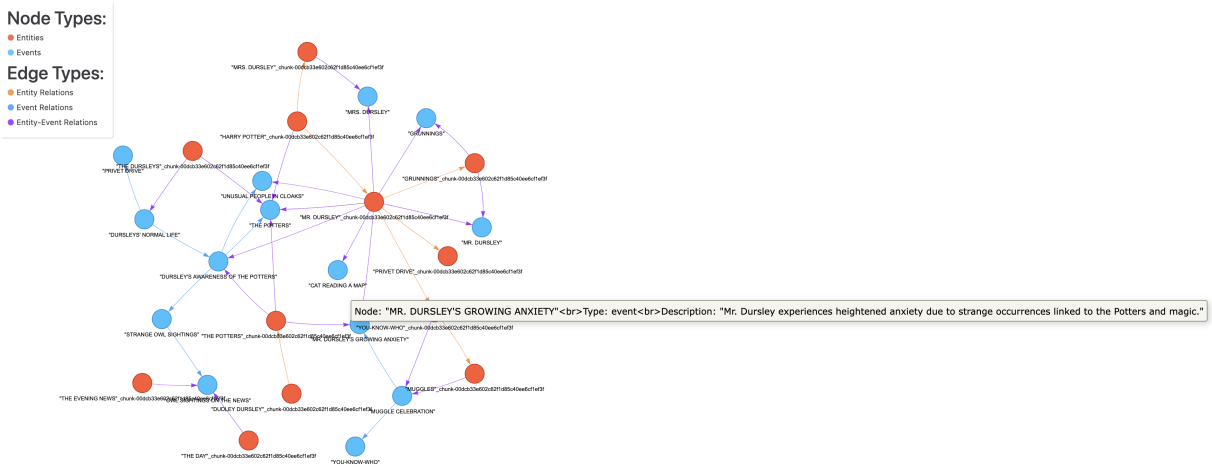


Figure 4: A snapshot of a part of the Entity-Event KG of *Harry Potter and the Sorcerer's Stone*.

# I Entity-Event KG Visualization

# J More Details on Case Study on Mode: $E^2$RAG (comb. extraction)

## J.1 Hypothetical Response

```
In the dinner scene from "The Hound of the Baskervilles," Watson's narration captures Holmes\'s
fervent enthusiasm for elucidating his reasoning behind the Lauriston Gardens mystery. As they sit
at the table, Watson observes Holmes's animated demeanor, noting how "his eyes shone with a keen
light" and how he "leaned forward with an eagerness that betrayed his excitement." This vivid
description not only illustrates Holmes\'s passion for deduction but also emphasizes his desire to
engage Watson in the unraveling of the case, making him feel like a partner in the detective\'s
intellectual adventure.\n\nHolmes employs various rhetorical strategies to accentuate the
significance of each clue as he meticulously lays out his thought process. For instance, he often
uses analogies to simplify complex ideas, comparing the clues to pieces of a puzzle that, when
assembled correctly, reveal the bigger picture. He states, "The world is full of obvious things
which nobody by any chance ever observes," underscoring how critical attention to detail is in
investigative work. This rhetorical device serves to remind Watson-and the reader-of the importance
of keen observation in deciphering the mystery.\n\nMoreover, Holmes frequently employs the technique
of building suspense through his methodical breakdown of the evidence. He carefully dissects each
piece of information, pausing for effect before revealing its implications. By saying, "And yet, it
is the most obvious thing in the world," he not only highlights the clarity of his deductions but
also invites the audience to share in the thrill of discovery. This step-by-step approach not only
showcases Holmes\'s logical prowess but also enhances Watson's narrative by creating a sense of
shared enlightenment, as they both traverse the landscape of clues together.\n\nUltimately,
Watson\'s narration, rich in detail and observation, amplifies Holmes's eagerness to teach and
elucidate, while Holmes's rhetorical strategies serve to elevate the weight of each clue,
transforming the dinner scene into a dynamic exploration of deduction and insight.
```

## J.2 Extracted Data

Extracted entity keywords:

```
['Watson', 'dinner scene', 'The Hound of the Baskervilles', 'Lauriston Gardens', 'Holmes']
```

Extracted event keywords:

```
["Holmes's enthusiasm", 'dinner scene', 'discussion of the Lauriston Gardens mystery', 'breakdown of
evidence', 'building suspense', "underscoring clues' significance", "narration of Holmes's
reasoning", 'detailing reasoning', 'highlighting eagerness', 'explaining logic', 'using rhetorical
strategies', 'shared enlightenment', 'deduction process']
```

Retrieved nodes data:

[\{'type': '"PERSON"', 'description': '"Holmes is a character portrayed as a detective who engages in solving a mystery surr...and expresses confidence in his theories."', 'source_id': 'chunk-e16ab5c062904b8574746080e70db172', 'node_type': 'entity', 'display_color': '#FF5733', 'entity_name': '"HOLMES"_chunk-e16ab5c062904b8574746080e70db172', 'rank': 9\}, ...\{'type': '"EVENT"', 'description': '"Holmes elaborates on his strategy to catch the assassin while navigating the complexities of the situation."', 'source_id': 'chunk-fbff6479d4cfc0843851b4ff06a86927', 'node_type': 'event', 'display_color': '#33C1FF', 'entity_name': '"HOLMES\'S STRATEGY DISCUSSION"', 'rank': 2\}...]

Retrieved edge data:

[\{'src_tgt': (...), 'rank': 37, 'weight': 8.0, 'description': '"I engages in dialogue with Holmes, seeking understanding of the detective\'s methodologies and deductions."', 'keywords': '"narration, learning"', 'source_id': 'chunk-9d367c150b511e67b26c6a5701e7fcb9', 'edge_type': 'entity_relation', 'display_color': '#FF8C33'\}, \{'src_tgt': (...), 'rank': 25, 'edge_type': 'entity_event_relation', 'description': 'Entity \'"Holmes is a skilled detective known for his sharp powers of observation and ...s wheels as a crucial piece of evidence."\'', 'weight': 1.0, 'keywords': 'ENTITY_EVENT_RELATION', 'source_id': 'chunk-9d367c150b511e67b26c6a5701e7fcb9<SEP>chunk-9d367c150b511e67b26c6a5701e7fcb9', 'display_color': '#9933FF'\},...]

Retrieved context data (relevant chunks):

\n-----Sources-----\n```csv\nid,content\r\n0,"and that the depravity of the victim was no condonement in the eyes of the law.\n\nThe more I thought of it the more extraordinary did my companion's hypothesis, that the man had been poisoned, appear. I remembered how he had sniffed his lips, and had no doubt that he had detected something which had given rise to the idea. Then, again, if not poison, what had caused the man's death, since there was neither wound nor marks of strangulation? But, on the other hand, whose blood was that which lay so thickly upon the floor? There were no signs of a struggle, nor had the victim any weapon with which he might have wounded an antagonist. As long as all these questions were unsolved, I felt that sleep would be no easy matter, either for Holmes or myself. His quiet self-confident manner convinced me that he had already formed a theory which explained all the facts, though what it was I could not for an instant conjecture. ...

Final response:

In the dinner scene where Sherlock Holmes discusses the Lauriston Gardens mystery with Dr. Watson, Watson \u2019s narration reveals Holmes's enthusiastic and methodical approach to detective work. This moment underscores not only Holmes\u2019s intellectual prowess but also his eagerness to engage Watson in the reasoning process that leads to his deductions. Watson\u2019s observations during this discussion reflect his genuine admiration for Holmes's analytical mind, which illuminates the depth of the detective\u2019s insight and reasoning.\n\n### Watson\u2019s Narrative Perspective\n\nWatson, as the narrator, captures the intensity of Holmes\u2019s engagement with the case. He notes how Holmes displays a \"quiet self-confident manner,\" suggesting that Holmes is not only certain of his deductions but also eager to articulate the thought process that led him there. For instance, Watson recounts moments where Holmes transitions on topics with excitement, indicating how much he relishes sharing the intricacies of his method \u2014 \u201cThe plot thickens,\u201d he exclaims at key junctures, revealing his thrill at unraveling the mystery. This narrative style enhances Holmes's character as not just a brilliant detective but also a passionate educator, inviting Watson\u2014and, by extension, the reader\u2014into his deductive world.

\n\n### Rhetorical Strategies Used by Holmes\n\nHolmes employs several rhetorical strategies to underscore the significance of each clue in his explanation. These strategies include:\n\n1. **Logical Sequencing**: Holmes emphasizes the importance of reasoning backwards and forwards, which he articulately illustrates when he examines the physical evidence. For example, he reasons from the cab\u2019s ruts and horse\u2019s hoof prints left on the wet soil to deduce that two men must have visited the house. Here, he insists that \"it is a capital mistake to theorize before you have all the evidence,\" highlighting his disciplined approach to deduction that relies on empirical observations rather than mere speculation.\n\n2. **Detailed Observation**: Holmes's acute observations serve as pivotal points in his reasoning. He recognizes that the \"very first thing\" he observed was the impressions left by the cab wheels. By meticulously detailing this observation, he conveys to Watson\u2014and the audience\u2014the fundamental role of minute details in solving a case: \"There is no branch of detective science which is so important and so much neglected as the art of tracing footsteps.\"\n\n3. **Analytical Clarity**: As he breaks down how he calculated the height and nature of the suspects based on their footprints and style of dress, Holmes uses clear, methodical language. He moves from discussing the physical evidence to the psychological implications, illustrating how the spatial and temporal context plays a critical role in understanding the crime. For example, he reflects on the footprints and concludes, \u201cIt is simple enough\u2026 I could read all that in the dust.\u201d \n\n4. **Connecting to Broader Concepts**: Holmes often draws on broader concepts of criminality and human behavior, thereby situating his deductions within a larger framework of detective work. He references historical cases and theories that bolster his arguments, explaining his reasoning not just as a technical endeavor, but as a complex interplay of human motivations and actions.\n\n### Conclusion\n\nHolmes's eagerness to share his deductive reasoning exemplifies the detective's character as someone deeply engaged in the pursuit of knowledge and understanding. Through Watson\u2019s narration, readers see an interplay of admiration, curiosity, and insight as Holmes articulates the significance of each clue. The combination of logical sequencing, detailed observation, analytical clarity, and broader contextual reasoning allows Holmes to construct a compelling narrative that not only uncovers the truth but also elevates the art of detection itself. This dinner scene, rich with multi-layered reasoning, highlights the intricacies of Holmes's mind, drawing readers closer to the fascinating world of deduction.

## J.3 Complete Verdicts from some LLM Judges

Table 8: GPT-4.1-mini's verdicts (part 1) of the responses to *"Consider the dinner scene in which Holmes details his reasoning about the Lauriston Gardens mystery. How does Watson's narration highlight Holmes's eagerness to explain the logic step by step, and what rhetorical strategies (quoted or paraphrased) does Holmes use to underscore each clue's significance?"*

| Mode | Average Score | Reason |
|---|---|---|
| E²RAG (comb. extraction) | 7.0000 | Very thorough and accurate, captures Watson's narration of Holmes's eagerness and detailed rhetorical strategies with direct quotes and logical progression, closely matching the ground truth. |
| E²RAG (comb. embedding) | 7.0000 | Comprehensive and well-aligned with the ground truth, includes detailed analysis of Watson's narration, Holmes's eagerness, and rhetorical strategies with direct quotes. |
| E²RAG (hyp. embedding) | 7.0000 | Detailed and faithful, includes Holmes's eagerness, rhetorical strategies, direct quotes, and Watson's narration, effectively capturing the essence of the ground truth. |
| LightRAG hybrid (hyp. extraction) | 6.6667 | Highly detailed and faithful to the ground truth, includes Holmes's systematic reasoning, direct quotes, rhetorical questions, and Watson's narration highlighting Holmes's eagerness. |
| LightRAG hybrid (comb. extraction) | 6.3333 | Captures Holmes's eagerness and rhetorical strategies well, including analogies, contrasts, and causal reasoning, with some direct quotes. However, it misses some of the specific step-by-step pacing and the theatrical, emphatic phrasing that the ground truth emphasizes, resulting in a somewhat less vivid depiction. |
| E²RAG (hyp. extraction) | 6.3333 | Provides a detailed and faithful account of Watson's narration and Holmes's eagerness, including direct quotes and rhetorical strategies such as deductive reasoning, metaphor, and analogy. However, it lacks some of the exact emphatic openings, instructive questions, and dismissive flourishes explicitly cited in the ground truth, making it slightly less precise. |
| LightRAG hybrid (comb. embedding) | 6.3333 | Providing a solid overview of Holmes's eagerness and rhetorical strategies with direct quotes and logical reasoning. However, it does not fully capture the breathless pacing and specific emphatic or dismissive phrases that characterize the ground truth. |
| LightRAG hybrid (hyp. embedding) | 6.3333 | Nearly identical to hybrid_combined_extraction in content and style, with good coverage of Watson's narration and Holmes's rhetorical strategies but lacking the full range of direct quotes and the theatrical emphasis present in the ground truth. |
| E²RAG (vanilla) | 6.0000 | Detailed and faithful to the ground truth, includes Holmes's systematic approach, direct quotes, and rhetorical strategies, capturing both Watson's narration and Holmes's methodical explanation. |
| GraphRAG global | 6.0000 | Addresses Strong answer highlighting Holmes's eagerness and rhetorical strategies like analogies and emphatic language, with direct quotes and good alignment to the ground truth. |
| LightRAG hybrid | 6.0000 | Provides a thorough analysis of Watson's narration and Holmes's rhetorical strategies with examples and direct quotes, closely aligning with the ground truth's emphasis on step-by-step logic and Holmes's delight in explaining. |
| GraphRAG drift | 6.0000 | Mentions Comprehensive and detailed, covers Watson's narration and Holmes's rhetorical strategies including analogies, rhetorical questions, and historical context, closely matching the ground truth. |
| GraphRAG local | 5.6667 | Good summary of Watson's narration and Holmes's rhetorical strategies including logical reasoning and analogies, with some direct quotes, but less detailed than the ground truth. |
| LightRAG global | 5.6667 | Effectively highlights Holmes's eagerness and rhetorical strategies such as analogies and rhetorical questions with some direct quotes, closely matching the ground truth's focus on Holmes's methodical explanation. |

Table 8: GPT-4.1-mini's verdicts (continued)

| Mode | Average Score | Reason |
|---|---|---|
| LightRAG local | 5.3333 | Describes Watson's narration and Holmes's rhetorical strategies like analogies and highlighting contradictions, but lacks specific direct quotes and detailed step-by-step logic as in the ground truth. |
| vanilla RAG | 4.3333 | Captures Holmes's eagerness and details several rhetorical strategies with some direct quotes and examples, but includes some inaccuracies and extraneous content not directly related to the Lauriston Gardens dinner scene. |
| vanilla HyDE | 3.6667 | Captures Holmes's eagerness and rhetorical strategies with some direct quotes and analogies, but less focused on the step-by-step logic and specific clues as in the ground truth. |
| RQ-RAG | 3.0000 | Very Very brief and vague; mentions Holmes's eagerness and rhetorical questions but lacks detail, examples, or direct quotes to support the answer. |

Table 9: Claude-3.7-Sonnet's verdicts (part 1) of the responses to *"Consider the dinner scene in which Holmes details his reasoning about the Lauriston Gardens mystery. How does Watson's narration highlight Holmes's eagerness to explain the logic step by step, and what rhetorical strategies (quoted or paraphrased) does Holmes use to underscore each clue's significance?"*

| Mode | Average Score | Reason |
|---|---|---|
| E²RAG (comb. extraction) | 7.0000 | Mentions footprints and cab marks, and discusses Holmes's eagerness to explain his reasoning. However, it doesn't capture the specific rhetorical pattern of short declarations followed by challenges to Watson. |
| E²RAG (comb. embedding) | 7.0000 | Discusses Holmes's eagerness and mentions footprints and stride length, but doesn't fully capture the rhetorical pattern of short declarations followed by challenges to Watson described in the ground truth. |
| E²RAG (hyp. embedding) | 7.0000 | Mentions cab marks and footprints, and discusses Holmes's eagerness to explain his reasoning. However, it doesn't capture the specific rhetorical pattern of short declarations followed by challenges to Watson. |
| LightRAG hybrid (hyp. extraction) | 6.6667 | Mentions key elements like cab wheel ruts and stride length, and discusses Holmes's eagerness to explain his reasoning. Includes some specific rhetorical strategies but doesn't fully capture the rapid, orderly exposition and the pattern of declarations followed by challenges to Watson described in the ground truth. |
| LightRAG hybrid (comb. extraction) | 6.3333 | Accurately mentions cab tracks and Holmes's eagerness to explain his reasoning step by step. References some rhetorical strategies like analogies and contrasts, but doesn't fully capture all the specific elements (fingernails, puddle width) or the rhetorical pattern of emphatic openings and dismissive flourishes in the ground truth. |
| E²RAG (hyp. extraction) | 6.3333 | Mentions cab marks and footprints but doesn't specifically address Holmes's eagerness to explain step by step or his rhetorical pattern of making declarations followed by challenging Watson. |
| LightRAG hybrid (comb. embedding) | 6.3333 | Mentions cab tracks and Holmes's eagerness to explain his reasoning. Includes some rhetorical strategies but doesn't fully capture all the specific elements or the rhetorical pattern described in the ground truth. |
| LightRAG hybrid (hyp. embedding) | 6.3333 | Mentions cab tracks and Holmes's eagerness to explain his reasoning. Includes some rhetorical strategies but doesn't fully capture all the specific elements or the rhetorical pattern described in the ground truth. |
| E²RAG (vanilla) | 6.0000 | Mentions some key elements like footprints and cab marks but doesn't specifically address Holmes's eagerness to explain step by step or his rhetorical pattern of making declarations followed by challenging Watson. |
| GraphRAG global | 6.0000 | Addresses Holmes's eagerness to explain his reasoning but lacks specific references to footprints, stride lengths, and cab marks. The rhetorical strategies mentioned are not specific to the text. |
| LightRAG hybrid | 6.0000 | Provides a detailed analysis of Holmes's eagerness and rhetorical strategies but doesn't specifically mention the footprints, stride lengths, and cab marks that are central to the ground truth. The rhetorical strategies are more generalized than specific. |
| GraphRAG drift | 6.0000 | Mentions Holmes's eagerness and some rhetorical strategies but lacks specific references to footprints, stride lengths, and cab marks. The analysis is somewhat generic rather than text-specific. |

Table 9: Claude-3.7-Sonnet's verdicts (continued)

| Mode | Average Score | Reason |
|---|---|---|
| GraphRAG local | 5.6667 | Discusses Holmes's eagerness and some rhetorical strategies but lacks specific references to the footprints, stride lengths, and cab marks mentioned in the ground truth. The analysis is somewhat generic. |
| LightRAG local | 5.3333 | Discusses Holmes's eagerness and methodical reasoning but lacks specific references to footprints, stride lengths, and cab marks mentioned in the ground truth. The rhetorical strategies are generalized rather than specific to the text. |
| LightRAG global | 5.6667 | Addresses Holmes's eagerness to explain his logic but lacks specific references to the footprints, stride lengths, and cab marks. Mentions rhetorical questions but doesn't capture Holmes's pattern of short declarations followed by challenges to Watson. |
| vanilla RAG | 4.3333 | Contains some relevant elements about Holmes's analytical reasoning but misquotes Holmes and includes fabricated dialogue. The analysis lacks specific references to Watson's narration of Holmes's eagerness and the step-by-step logic mentioned in the ground truth. |
| vanilla HyDE | 3.6667 | Discusses a different story ('The Hound of the Baskervilles') and doesn't mention the specific elements from the ground truth like footprints, stride lengths, and cab marks. |
| RQ-RAG | 3.0000 | Very brief and generic. While it mentions Holmes's eagerness and rhetorical strategies, it lacks specific details about footprints, stride lengths, and cab marks. The answer is too vague and underdeveloped. |

Table 10: GPT-4o's verdicts (part 1) of the responses to *"Consider the dinner scene in which Holmes details his reasoning about the Lauriston Gardens mystery. How does Watson's narration highlight Holmes's eagerness to explain the logic step by step, and what rhetorical strategies (quoted or paraphrased) does Holmes use to underscore each clue's significance?"*

| Mode | Average Score | Reason |
|---|---|---|
| E$^2$RAG (comb. extraction) | 7.0000 | Partially correct; discusses Holmes's eagerness and some rhetorical strategies but lacks specific examples. |
| E$^2$RAG (comb. embedding) | 7.0000 | Partially correct; discusses Holmes's eagerness and some rhetorical strategies but lacks specific examples. |
| E$^2$RAG (hyp. embedding) | 7.0000 | Partially correct; discusses Holmes's eagerness and some rhetorical strategies but lacks specific examples. |
| LightRAG hybrid (hyp. extraction) | 6.6667 | Partially correct; discusses Holmes's eagerness and some rhetorical strategies but lacks specific examples. |
| LightRAG hybrid (comb. extraction) | 6.3333 | Partially correct; discusses Holmes's eagerness and some rhetorical strategies but lacks specific examples. |
| E$^2$RAG (hyp. extraction) | 6.3333 | Partially correct; discusses Holmes's eagerness and some rhetorical strategies but lacks specific examples. |
| LightRAG hybrid (comb. embedding) | 6.3333 | Partially correct; discusses Holmes's eagerness and some rhetorical strategies but lacks specific examples. |
| LightRAG hybrid (hyp. embedding) | 6.3333 | Partially correct; discusses Holmes's eagerness and some rhetorical strategies but lacks specific examples. |
| E$^2$RAG (vanilla) | 6.0000 | Partially correct; discusses Holmes's enthusiasm and some rhetorical strategies but lacks specific examples. |
| GraphRAG global | 6.0000 | Addresses Partially correct; discusses Holmes's eagerness and some rhetorical strategies but lacks specific examples. |
| LightRAG hybrid | 6.0000 | Partially correct; mentions Holmes's eagerness and some rhetorical strategies but lacks specific examples. |
| GraphRAG drift | 6.0000 | Mentions Partially correct; discusses Holmes's eagerness and some rhetorical strategies but lacks specific examples. |
| GraphRAG local | 5.6667 | Partially correct; discusses Holmes's eagerness and some rhetorical strategies but lacks specific examples. |
| LightRAG local | 5.3333 | Partially correct; mentions Holmes's enthusiasm and some rhetorical strategies but lacks specific examples. |
| LightRAG global | 5.6667 | Partially correct; discusses Holmes's eagerness and rhetorical strategies but lacks specific examples from the scene. |
| vanilla RAG | 4.3333 | Mostly incorrect; lacks specific details about Holmes's rhetorical strategies and Watson's narration. |
| vanilla HyDE | 3.6667 | Off-topic; discusses 'The Hound of the Baskervilles' instead of Lauriston Gardens. |
| RQ-RAG | 3.0000 | Very Mostly incorrect; brief mention of rhetorical strategies without specific examples or context. |

## K Additional Results

Table 11 to Table 16 show the ranking results of each individual LLM evaluator. It can be observed each mode's ranking remains relatively stable. On the other hand, different evaluators have different standards, as the scores given by Claude-3-7-sonnet are generally low while the scores given by GPT-4.1-mini are generally high across the board. Table 17 to Table 24 show the ranking results for the questions based on category, averaged across all three evaluator LLMs.

Table 11: Average scores for `Claude-3-7-sonnet`

| Rank | Mode | Avg Score | Total |
|---|---|---|---|
| 1 | **E$^2$RAG** (comb. extraction) | 4.6900 | 4 790 |
| 2 | **E$^2$RAG** (hyp. embedding) | 4.6600 | 4 675 |
| 3 | GraphRAG local | 4.6300 | 4 736 |
| 4 | **E$^2$RAG** (vanilla) | 4.6200 | 4 718 |
| 5 | LightRAG hybrid | 4.6000 | 4 626 |
| 6 | **E$^2$RAG** (hyp. extraction) | 4.5900 | 4 701 |
| 7 | **E$^2$RAG** (comb. embedding) | 4.4500 | 4 455 |
| 8 | GraphRAG drift | 4.4100 | 4 511 |
| 9 | LightRAG local | 4.3800 | 4 492 |
| 10 | LightRAG global | 4.2700 | 4 361 |
| 11 | GraphRAG global | 4.0600 | 4 139 |
| 12 | vanilla RAG | 3.9700 | 4 067 |
| 13 | vanilla HyDE | 3.8100 | 3 902 |
| 14 | LightRAG hybrid (hyp. embedding) | 3.7600 | 3 843 |
| 15 | LightRAG hybrid (comb. extraction) | 3.7600 | 3 836 |
| 16 | LightRAG hybrid (comb. embedding) | 3.7500 | 3 834 |
| 17 | LightRAG hybrid (hyp. extraction) | 3.7200 | 3 791 |
| 18 | RQ-RAG | 2.6500 | 2 604 |

Table 12: Average scores for `GPT-4o`

| Rank | Mode | Avg Score | Total |
|---|---|---|---|
| 1 | GraphRAG local | 6.6700 | 6 846 |
| 2 | **E$^2$RAG** (comb. extraction) | 6.6500 | 6 829 |
| 3 | GraphRAG drift | 6.6400 | 6 811 |
| 4 | LightRAG hybrid | 6.6200 | 6 795 |
| 5 | **E$^2$RAG** (comb. embedding) | 6.6000 | 6 731 |
| 6 | **E$^2$RAG** (hyp. extraction) | 6.5700 | 6 752 |
| 7 | **E$^2$RAG** (hyp. embedding) | 6.5300 | 6 710 |
| 8 | **E$^2$RAG** (vanilla) | 6.5200 | 6 700 |
| 9 | LightRAG local | 6.4400 | 6 623 |
| 10 | GraphRAG global | 6.3600 | 6 525 |
| 11 | LightRAG global | 6.3300 | 6 509 |
| 12 | LightRAG hybrid (comb. embedding) | 6.2000 | 6 297 |
| 13 | LightRAG hybrid (comb. extraction) | 6.1900 | 6 349 |
| 14 | LightRAG hybrid (hyp. extraction) | 6.1900 | 6 343 |
| 15 | LightRAG hybrid (hyp. embedding) | 6.1700 | 6 264 |
| 16 | vanilla HyDE | 6.0800 | 6 243 |
| 17 | vanilla RAG | 5.9200 | 6 083 |
| 18 | RQ-RAG | 3.9800 | 3 961 |

Table 13: Average scores for `GPT-4.1-mini`

| Rank | Mode | Avg Score | Total |
|---|---|---|---|
| 1 | **E$^2$RAG** (comb. extraction) | 8.8700 | 9 102 |
| 2 | **E$^2$RAG** (comb. embedding) | 8.8000 | 9 028 |
| 3 | **E$^2$RAG** (hyp. extraction) | 8.6400 | 8 877 |
| 4 | **E$^2$RAG** (hyp. embedding) | 8.6400 | 8 856 |
| 5 | GraphRAG drift | 8.4300 | 8 653 |
| 6 | GraphRAG local | 8.4200 | 8 649 |
| 7 | **E$^2$RAG** (vanilla) | 8.3600 | 8 586 |
| 8 | LightRAG hybrid | 8.2600 | 8 479 |
| 9 | vanilla HyDE | 8.0200 | 8 221 |
| 10 | GraphRAG global | 7.9500 | 8 161 |
| 11 | LightRAG local | 7.6900 | 7 897 |
| 12 | LightRAG global | 7.5900 | 7 800 |
| 13 | LightRAG hybrid (comb. embedding) | 7.1000 | 7 126 |
| 14 | LightRAG hybrid (comb. extraction) | 7.0900 | 7 241 |
| 15 | LightRAG hybrid (hyp. embedding) | 7.0100 | 7 054 |
| 16 | LightRAG hybrid (hyp. extraction) | 6.9900 | 7 157 |
| 17 | vanilla RAG | 6.2300 | 6 399 |
| 18 | RQ-RAG | 3.9100 | 3 871 |

Table 14: Average scores for `claude-opus-4-1-20250805`

| Rank | Mode | Avg Score | Total |
|---|---|---|---|
| 1 | **E$^2$RAG** (comb. extraction) | 5.9600 | 5 397 |
| 2 | **E$^2$RAG** (hyp. embedding) | 5.9400 | 5 354 |
| 3 | **E$^2$RAG** (hyp. extraction) | 5.8500 | 5 297 |
| 4 | **E$^2$RAG** (comb. embedding) | 5.8500 | 5 255 |
| 5 | vanilla RAG | 5.6500 | 4 918 |
| 6 | **E$^2$RAG** (vanilla) | 5.6500 | 4 913 |
| 7 | GraphRAG local | 5.5900 | 5 063 |
| 8 | GraphRAG drift | 5.5200 | 5 003 |
| 9 | LightRAG hybrid | 5.5100 | 4 781 |
| 10 | LightRAG local | 5.3400 | 4 642 |
| 11 | LightRAG global | 5.0200 | 4 367 |
| 12 | GraphRAG global | 4.8400 | 4 381 |
| 13 | vanilla HyDE | 4.6300 | 4 194 |
| 14 | LightRAG hybrid (comb. embedding) | 4.4500 | 4 030 |
| 15 | LightRAG hybrid (comb. extraction) | 4.4300 | 4 015 |
| 16 | LightRAG hybrid (hyp. embedding) | 4.4300 | 4 009 |
| 17 | LightRAG hybrid (hyp. extraction) | 4.4100 | 3 998 |
| 18 | RQ-RAG | 3.7100 | 2 131 |

Table 15: Average scores for `gemini-2.5-pro`

| Rank | Mode | Avg Score | Total |
|---|---|---|---|
| 1 | **E$^2$RAG** (hyp. embedding) | 5.6000 | 5 703 |
| 2 | **E$^2$RAG** (hyp. extraction) | 5.5900 | 5 720 |
| 3 | **E$^2$RAG** (comb. extraction) | 5.5200 | 5 651 |
| 4 | **E$^2$RAG** (comb. embedding) | 5.4600 | 5 461 |
| 5 | LightRAG local | 5.4500 | 5 381 |
| 6 | GraphRAG local | 5.4400 | 5 570 |
| 7 | LightRAG hybrid | 5.3800 | 5 201 |
| 8 | **E$^2$RAG** (vanilla) | 5.3700 | 5 290 |
| 9 | LightRAG global | 5.2000 | 5 136 |
| 10 | GraphRAG drift | 4.9700 | 5 090 |
| 11 | GraphRAG global | 4.8100 | 4 920 |
| 12 | vanilla RAG | 4.7200 | 4 657 |
| 13 | vanilla HyDE | 4.3600 | 4 397 |
| 14 | LightRAG hybrid (hyp. extraction) | 4.2300 | 4 286 |
| 15 | LightRAG hybrid (hyp. embedding) | 4.2300 | 4 228 |
| 16 | LightRAG hybrid (comb. extraction) | 4.2200 | 4 284 |
| 17 | LightRAG hybrid (comb. embedding) | 4.2200 | 4 054 |
| 18 | RQ-RAG | 2.1200 | 1 942 |

Table 16: Average scores for `qwen-max-2025-01-25`

| Rank | Mode | Avg Score | Total |
|---|---|---|---|
| 1 | LightRAG hybrid | 6.9300 | 6 858 |
| 2 | **E$^2$RAG** (comb. embedding) | 6.6800 | 6 111 |
| 3 | **E$^2$RAG** (hyp. embedding) | 6.6700 | 6 127 |
| 4 | **E$^2$RAG** (comb. extraction) | 6.6500 | 6 812 |
| 5 | GraphRAG drift | 6.5800 | 6 729 |
| 6 | **E$^2$RAG** (hyp. extraction) | 6.5800 | 6 729 |
| 7 | GraphRAG local | 6.5500 | 6 695 |
| 8 | LightRAG local | 6.4500 | 6 380 |
| 9 | LightRAG hybrid (comb. embedding) | 6.4000 | 6 342 |
| 10 | LightRAG hybrid (comb. extraction) | 6.3800 | 6 510 |
| 11 | LightRAG hybrid (hyp. embedding) | 6.3600 | 6 301 |
| 12 | GraphRAG global | 6.3500 | 6 493 |
| 13 | LightRAG hybrid (hyp. extraction) | 6.3400 | 6 468 |
| 14 | LightRAG global | 6.3400 | 6 274 |
| 15 | vanilla HyDE | 6.2600 | 6 419 |
| 16 | **E$^2$RAG** (vanilla) | 6.1700 | 6 032 |
| 17 | vanilla RAG | 5.8600 | 5 795 |
| 18 | RQ-RAG | 3.7200 | 3 061 |

Table 17: Causal Consistency category: average scores

| Rank | Mode | Avg Score | Total |
|---|---|---|---|
| 1 | **E$^2$RAG** (comb. extraction) | 6.1600 | 10 020 |
| 2 | **E$^2$RAG** (hyp. embedding) | 6.1200 | 9 665 |
| 3 | **E$^2$RAG** (comb. embedding) | 6.0700 | 9 488 |
| 4 | **E$^2$RAG** (hyp. extraction) | 6.0500 | 9 855 |
| 5 | **E$^2$RAG** (vanilla) | 6.0100 | 9 705 |
| 6 | GraphRAG local | 5.9800 | 9 710 |
| 7 | LightRAG hybrid | 5.8700 | 9 405 |
| 8 | GraphRAG drift | 5.8300 | 9 497 |
| 9 | LightRAG local | 5.6500 | 9 144 |
| 10 | GraphRAG global | 5.5200 | 8 973 |
| 11 | LightRAG global | 5.4900 | 8 877 |
| 12 | vanilla RAG | 5.1900 | 8 399 |
| 13 | vanilla HyDE | 5.1100 | 8 290 |
| 14 | LightRAG hybrid (comb. extraction) | 4.8900 | 7 912 |
| 15 | LightRAG hybrid (comb. embedding) | 4.8600 | 7 685 |
| 16 | LightRAG hybrid (hyp. extraction) | 4.8500 | 7 857 |
| 17 | LightRAG hybrid (hyp. embedding) | 4.8300 | 7 707 |
| 18 | RQ-RAG | 2.9500 | 3 387 |

Table 18: Character Consistency category: average scores

| Rank | Mode | Avg Score | Total |
|---|---|---|---|
| 1 | **E$^2$RAG** (comb. extraction) | 6.5700 | 15 193 |
| 2 | **E$^2$RAG** (hyp. embedding) | 6.5500 | 14 894 |
| 3 | **E$^2$RAG** (hyp. extraction) | 6.5100 | 15 068 |
| 4 | **E$^2$RAG** (comb. embedding) | 6.5100 | 14 730 |
| 5 | LightRAG hybrid | 6.4800 | 14 851 |
| 6 | GraphRAG local | 6.3700 | 14 747 |
| 7 | **E$^2$RAG** (vanilla) | 6.3100 | 14 446 |
| 8 | GraphRAG drift | 6.2900 | 14 539 |
| 9 | LightRAG local | 6.2100 | 14 281 |
| 10 | LightRAG global | 6.0600 | 13 921 |
| 11 | GraphRAG global | 5.9100 | 13 679 |
| 12 | vanilla HyDE | 5.8500 | 13 513 |
| 13 | LightRAG hybrid (comb. embedding) | 5.7200 | 13 029 |
| 14 | LightRAG hybrid (comb. extraction) | 5.7100 | 13 170 |
| 15 | LightRAG hybrid (hyp. embedding) | 5.6900 | 13 017 |
| 16 | LightRAG hybrid (hyp. extraction) | 5.6700 | 13 101 |
| 17 | vanilla RAG | 5.5100 | 12 653 |
| 18 | RQ-RAG | 3.5300 | 9 321 |

Table 19: Emotional and Psychological category: average scores

| Rank | Mode | Avg Score | Total |
|---|---|---|---|
| 1 | GraphRAG local | 6.460 0 | 1 576 |
| 2 | LightRAG hybrid | 6.340 0 | 1 515 |
| 3 | **E$^2$RAG** (hyp. extraction) | 6.300 0 | 1 536 |
| 4 | **E$^2$RAG** (comb. extraction) | 6.240 0 | 1 522 |
| 5 | GraphRAG drift | 6.230 0 | 1 521 |
| 6 | **E$^2$RAG** (hyp. embedding) | 6.230 0 | 1 495 |
| 7 | **E$^2$RAG** (vanilla) | 6.140 0 | 1 498 |
| 8 | LightRAG local | 6.110 0 | 1 496 |
| 9 | **E$^2$RAG** (comb. embedding) | 6.080 0 | 1 454 |
| 10 | RQ-RAG | 5.920 0 | 1 383 |
| 11 | LightRAG global | 5.620 0 | 1 372 |
| 12 | vanilla HyDE | 5.580 0 | 1 362 |
| 13 | GraphRAG global | 5.450 0 | 1 324 |
| 14 | LightRAG hybrid (comb. embedding) | 5.450 0 | 1 307 |
| 15 | LightRAG hybrid (comb. extraction) | 5.430 0 | 1 325 |
| 16 | LightRAG hybrid (hyp. extraction) | 5.420 0 | 1 316 |
| 17 | LightRAG hybrid (hyp. embedding) | 5.420 0 | 1 311 |
| 18 | vanilla RAG | 5.290 0 | 1 291 |

Table 20: Narrative and Plot Structure category: average scores

| Rank | Mode | Avg Score | Total |
|---|---|---|---|
| 1 | **E$^2$RAG** (comb. extraction) | 6.380 0 | 1 664 |
| 2 | **E$^2$RAG** (comb. embedding) | 6.330 0 | 1 602 |
| 3 | **E$^2$RAG** (hyp. embedding) | 6.250 0 | 1 595 |
| 4 | **E$^2$RAG** (hyp. extraction) | 6.230 0 | 1 625 |
| 5 | GraphRAG local | 6.210 0 | 1 622 |
| 6 | GraphRAG drift | 6.170 0 | 1 611 |
| 7 | **E$^2$RAG** (vanilla) | 6.000 0 | 1 525 |
| 8 | LightRAG global | 5.960 0 | 1 521 |
| 9 | LightRAG hybrid | 5.880 0 | 1 499 |
| 10 | LightRAG local | 5.840 0 | 1 490 |
| 11 | GraphRAG global | 5.680 0 | 1 483 |
| 12 | vanilla RAG | 5.310 0 | 1 355 |
| 13 | vanilla HyDE | 5.300 0 | 1 377 |
| 14 | LightRAG hybrid (hyp. embedding) | 5.150 0 | 1 308 |
| 15 | LightRAG hybrid (comb. extraction) | 5.130 0 | 1 335 |
| 16 | LightRAG hybrid (comb. embedding) | 5.130 0 | 1 304 |
| 17 | LightRAG hybrid (hyp. extraction) | 5.120 0 | 1 332 |
| 18 | RQ-RAG | 3.470 0 | 1 249 |

Table 21: Setting, Environment and Atmosphere category: average scores

| Rank | Mode | Avg Score | Total |
|---|---|---|---|
| 1 | **E$^2$RAG** (comb. extraction) | 5.960 0 | 1 956 |
| 2 | **E$^2$RAG** (hyp. extraction) | 5.770 0 | 1 897 |
| 3 | LightRAG hybrid | 5.770 0 | 1 882 |
| 4 | **E$^2$RAG** (comb. embedding) | 5.770 0 | 1 874 |
| 5 | **E$^2$RAG** (hyp. embedding) | 5.750 0 | 1 863 |
| 6 | **E$^2$RAG** (vanilla) | 5.620 0 | 1 827 |
| 7 | GraphRAG drift | 5.480 0 | 1 802 |
| 8 | LightRAG local | 5.460 0 | 1 781 |
| 9 | GraphRAG local | 5.370 0 | 1 767 |
| 10 | vanilla RAG | 5.330 0 | 1 738 |
| 11 | LightRAG global | 5.230 0 | 1 705 |
| 12 | vanilla HyDE | 5.210 0 | 1 713 |
| 13 | GraphRAG global | 5.170 0 | 1 697 |
| 14 | LightRAG hybrid (comb. extraction) | 4.940 0 | 1 614 |
| 15 | LightRAG hybrid (comb. embedding) | 4.940 0 | 1 595 |
| 16 | LightRAG hybrid (hyp. extraction) | 4.910 0 | 1 605 |
| 17 | LightRAG hybrid (hyp. embedding) | 4.900 0 | 1 587 |
| 18 | RQ-RAG | 2.900 0 | 909 |

Table 22: Social, Cultural and Political category: average scores

| Rank | Mode | Avg Score | Total |
|---|---|---|---|
| 1 | **E$^2$RAG** (comb. extraction) | 6.3700 | 2 750 |
| 2 | GraphRAG local | 6.3600 | 2 741 |
| 3 | **E$^2$RAG** (hyp. extraction) | 6.3400 | 2 727 |
| 4 | LightRAG hybrid | 6.3400 | 2 529 |
| 5 | **E$^2$RAG** (hyp. embedding) | 6.2700 | 2 629 |
| 6 | **E$^2$RAG** (comb. embedding) | 6.2500 | 2 594 |
| 7 | GraphRAG drift | 6.0800 | 2 620 |
| 8 | **E$^2$RAG** (vanilla) | 6.0600 | 2 448 |
| 9 | LightRAG local | 6.0400 | 2 442 |
| 10 | GraphRAG global | 5.6900 | 2 451 |
| 11 | LightRAG global | 5.6800 | 2 293 |
| 12 | vanilla RAG | 5.3900 | 2 177 |
| 13 | vanilla HyDE | 5.3800 | 2 302 |
| 14 | LightRAG hybrid (comb. extraction) | 5.3500 | 2 289 |
| 15 | LightRAG hybrid (comb. embedding) | 5.3300 | 2 224 |
| 16 | LightRAG hybrid (hyp. extraction) | 5.3100 | 2 274 |
| 17 | LightRAG hybrid (hyp. embedding) | 5.2900 | 2 234 |
| 18 | RQ-RAG | 3.0600 | 1 398 |

Table 23: Symbolism, Imagery and Motifs category: average scores

| Rank | Mode | Avg Score | Total |
|---|---|---|---|
| 1 | **E$^2$RAG** (comb. embedding) | 6.5600 | 2 590 |
| 2 | LightRAG hybrid | 6.5400 | 2 446 |
| 3 | **E$^2$RAG** (comb. extraction) | 6.5100 | 2 610 |
| 4 | **E$^2$RAG** (hyp. extraction) | 6.4800 | 2 599 |
| 5 | **E$^2$RAG** (hyp. embedding) | 6.4800 | 2 552 |
| 6 | GraphRAG local | 6.3700 | 2 548 |
| 7 | GraphRAG drift | 6.2700 | 2 508 |
| 8 | **E$^2$RAG** (vanilla) | 6.0600 | 2 278 |
| 9 | LightRAG global | 6.0400 | 2 278 |
| 10 | LightRAG local | 6.0200 | 2 270 |
| 11 | GraphRAG global | 5.9900 | 2 394 |
| 12 | vanilla HyDE | 5.7700 | 2 312 |
| 13 | LightRAG hybrid (comb. embedding) | 5.4600 | 2 164 |
| 14 | vanilla RAG | 5.4600 | 2 057 |
| 15 | LightRAG hybrid (hyp. embedding) | 5.4300 | 2 163 |
| 16 | LightRAG hybrid (comb. extraction) | 5.4200 | 2 172 |
| 17 | LightRAG hybrid (hyp. extraction) | 5.4100 | 2 170 |
| 18 | RQ-RAG | 3.1700 | 1 346 |

Table 24: Thematic, Philosophical and Moral category: average scores

| Rank | Mode | Avg Score | Total |
|---|---|---|---|
| 1 | **E$^2$RAG** (comb. extraction) | 6.7600 | 2 866 |
| 2 | GraphRAG local | 6.7300 | 2 848 |
| 3 | **E$^2$RAG** (hyp. embedding) | 6.6500 | 2 732 |
| 4 | **E$^2$RAG** (comb. embedding) | 6.6400 | 2 709 |
| 5 | LightRAG hybrid | 6.5800 | 2 613 |
| 6 | **E$^2$RAG** (hyp. extraction) | 6.5300 | 2 769 |
| 7 | GraphRAG drift | 6.4000 | 2 699 |
| 8 | **E$^2$RAG** (vanilla) | 6.2500 | 2 512 |
| 9 | LightRAG local | 6.2300 | 2 511 |
| 10 | GraphRAG global | 6.2000 | 2 618 |
| 11 | LightRAG global | 6.1700 | 2 480 |
| 12 | vanilla HyDE | 5.9500 | 2 507 |
| 13 | LightRAG hybrid (comb. embedding) | 5.7500 | 2 375 |
| 14 | LightRAG hybrid (comb. extraction) | 5.7400 | 2 418 |
| 15 | LightRAG hybrid (hyp. embedding) | 5.7000 | 2 372 |
| 16 | LightRAG hybrid (hyp. extraction) | 5.6900 | 2 388 |
| 17 | vanilla RAG | 5.5800 | 2 249 |
| 18 | RQ-RAG | 3.4300 | 1 528 |

## L Pseudocode

**Algorithm 1** Entity–Event KG Insertion (Pre-processing)

**Require:** document $D$; entity, event extractor $g(\cdot)$;
text encoder $f(\cdot)$; top-$L$ chunk size $m$
**Ensure:** graphs $\mathcal{G}_{\text{ent}}, \mathcal{G}_{\text{evt}}$; bipartite edge set $\mathcal{B}$; vector stores $\mathcal{I}_{\text{ent}}, \mathcal{I}_{\text{evt}}$

**procedure** INSERTDOCUMENT($D$)
  **Step 1: chunking**
  $\{c_\ell\}_{\ell=1}^{L} \leftarrow \text{CHUNK}(D, m)$
  **Step 2: node extraction**
  **for** $\ell \leftarrow 1$ **to** $L$ **do**
    $V_{\text{ent}}^{(\ell)}, V_{\text{evt}}^{(\ell)} \leftarrow g(c_\ell)$
  $V_{\text{ent}} \leftarrow \bigcup_\ell V_{\text{ent}}^{(\ell)};\ V_{\text{evt}} \leftarrow \bigcup_\ell V_{\text{evt}}^{(\ell)}$
  **Step 3: build edge sets**
  $E_{\text{ent}}, E_{\text{evt}} \leftarrow \text{INTRACHUNKEDGES}(\{c_\ell\}, V_{\text{ent}}, V_{\text{evt}})$
  $\mathcal{B} \leftarrow \text{BUILDBIPARTITE}(V_{\text{ent}}, V_{\text{evt}})$
  **Step 4: embed and index**
  **for all** $v \in V_{\text{ent}}$ **do** ▷ entities
    $\mathcal{I}_{\text{ent}}.\text{ADD}\big(v, f(\text{CANON}(v))\big)$
  **for all** $e \in V_{\text{evt}}$ **do** ▷ events
    $\mathcal{I}_{\text{evt}}.\text{ADD}\big(e, f(\text{CANON}(e))\big)$
  **return** $\mathcal{G}_{\text{ent}} = (V_{\text{ent}}, E_{\text{ent}}),\ \mathcal{G}_{\text{evt}} = (V_{\text{evt}}, E_{\text{evt}}),\ \mathcal{B}$

**Algorithm 2** BUILDBIPARTITE: connect entity mentions to events

**Require:** entity nodes $V_{\text{ent}}$, event nodes $V_{\text{evt}}$
**Ensure:** bipartite edge set $\mathcal{B}$

**function** BUILDBIPARTITE($V_{\text{ent}}, V_{\text{evt}}$)
  $\mathcal{B} \leftarrow \varnothing$
  */* surface-form string match inside the same chunk */*
  **for all** $v \in V_{\text{ent}}$ **do**
    $c \leftarrow \text{CHUNKID}(v);\ S \leftarrow \text{NAME}(v)$
    **for all** $e \in V_{\text{evt}}$ **with** $\text{CHUNKID}(e) = c$ **do**
      **if** $S \subset \text{DESC}(e)$ **then**
        $\mathcal{B} \leftarrow \mathcal{B} \cup \{(v, e)\}$
  **return** $\mathcal{B}$

**Algorithm 3** Entity–Event KG RETRIEVE (Inference-time)

**Require:** query $q$ (or hypothetical response transformed text),
entity store $\mathcal{I}_{\text{ent}}$, event store $\mathcal{I}_{\text{evt}}$,
bipartite edge set $\mathcal{B}$, similarity encoder $f(\cdot)$,
entity, event extractor $g(\cdot)$, top-$k$ parameter $k$
**Ensure:** ranked context bundle $C$ (passages + linearized subgraph)
**procedure** RETRIEVE($q$)
  **Cue extraction**
  $S_{\text{ent}}, S_{\text{evt}} \leftarrow g(q)$
  $V_q \leftarrow \text{LOOKUP}(S_{\text{ent}}, \mathcal{I}_{\text{ent}}) \cup \text{LOOKUP}(S_{\text{evt}}, \mathcal{I}_{\text{evt}})$
  **One-hop expansion across $\mathcal{B}$**
  $V_q^+ \leftarrow V_q \cup \{e \mid (v,e) \in \mathcal{B},\ v \in V_q\} \cup \{v \mid (v,e) \in \mathcal{B},\ e \in V_q\}$
  **Similarity ranking**
  $\mathcal{P} \leftarrow \{\text{CHUNK}(x) \mid x \in V_q^+\}$ ▷ candidate passages
  $\mathbf{z} \leftarrow f(q)$
  $\text{scores}[p] \leftarrow \langle f(p), \mathbf{z} \rangle \quad \forall p \in \mathcal{P}$
  $\mathcal{P}_k \leftarrow \text{TOPK}(\text{scores}, k)$
  **Context assembly**
  $G_{\text{sub}} \leftarrow (\, V_q^+,\ (E_{\text{ent}} \cup E_{\text{evt}} \cup \mathcal{B}) \cap (V_q^+ \times V_q^+)\,)$
  $C \leftarrow \text{FORMATCONTEXT}(\mathcal{P}_k, G_{\text{sub}})$
  **return** $C$

## M Evaluation Prompts and Settings

**Verifier prompt (used for passage-grounded answer checking).**

```
You are an expert fact-checker. Your task is to evaluate if a given answer is
fully supported by the provided context passage.

Context Passage:
{passage}

Question:
{question}

Answer to Verify:
{answer}

Instructions:
Read the passage, question, and answer carefully. Determine if every single
claim in the "Answer to Verify" is factually supported by the "Context Passage".
The answer must be 100% accurate based only on the provided text. Any
information, even if plausible, that is not in the passage makes the answer
incorrect.

Respond with a single JSON object containing two keys: "verdict" (either
"Correct" or "Incorrect") and "reason" (a brief explanation for your verdict).

Example:
{
  "verdict": "Incorrect",
  "reason": "The passage states the axe was made of tin, but the answer claims
```

```
            it was steel."
}
```

**Corruptor prompt (for synthetic negative answers in verifier stress tests).**

```
You are a data corruptor. Your task is to subtly alter a given answer based on
a specific error type, using only the information in the provided context
passage. The corrupted answer must be plausible but factually incorrect
according to the passage.

Context Passage:
{passage}

Question:
{question}

Original Correct Answer:
{answer}

Corruption Task:
Rewrite the 'Original Correct Answer' to introduce the following error:
{error_type}.

- Reversing Causality: Swap cause and effect.
- Fabricating Motivation: Invent a plausible but unsupported reason.
- Swapping Entities: Swap roles of two characters or objects.
- Adding Plausible but Incorrect Details: Add a small, believable but unsupported
  detail.

Do not state that you are corrupting the data. Just provide the new, subtly
incorrect answer.

Corrupted Answer:
```

**Generation guardrails for the backbone LLM.** The response generator is instructed to rely only on retrieved passages and to abstain when evidence is missing:

> *Generate a response of the target length and format that responds to the user's question, summarizing information in the input data tables appropriate for the response length and format, and incorporating any relevant general knowledge. If you do not know the answer, say so. Do not make anything up. Do not include information where the supporting evidence is not provided.*

## N LLM Versions and Reproducibility

**Backbone model for all compared RAG pipelines.** `gpt-4o-mini-2024-07-18` (weak backbone model is chosen to test RAG's capability).

**Embedding model for all compared RAG pipelines.** `text-embedding-3-small` (default choice of many baselines.

**LLM judges (six-model jury).** `claude-3-7-sonnet-20250219`, `gpt-4o-2024-08-06`, `claude-opus-4-1-20250805`, `gpt-4.1-mini-2025-04-14`, `gemini-2.5-pro` (stable), `qwen-max-2025-01-25`. LLM judeges from different model families are selected to prevent model biases.

**Decoding settings for evaluation.** Temperature 0 (except for `gpt-5-2025-08-07` which uses the default temperature), top-$p$ 1.0. We average scores across the fix judges. To mitigate vendor updates, we release raw judge outputs and per-item scores, and we record exact API model names and timestamps.
**Deterministic preprocessing and retrieval.** All indices are frozen before evaluation. We log random seeds for any sampling in chunking or negative sampling.

## O Verifier Robustness: Large-Scale Sanity Check

We tested the verifier on 100 items: 50 correct answers and 50 corrupted answers created by the corruptor prompt in Appendix M. Overall accuracy was 93%. The confusion matrix is below.

Table 25: Verifier confusion matrix on 100 items.

| | Predicted Correct | Predicted Incorrect |
|---|---|---|
| Actual Correct | 44 | 6 |
| Actual Incorrect | 1 | 49 |

The verifier detected 49 out of 50 corrupted cases (98%), with six false positives on correct answers. This conservative bias is suitable for dataset curation where soundness is preferred over recall.

## P Deterministic Supplementary Metrics

We computed BLEU, BERTScore (F1), and NLI entailment as supplementary, fully deterministic metrics. The rankings diverge from the LLM judges for ChronoQA's temporally grounded questions.

Table 26: Deterministic metrics on ChronoQA. Higher is better.

| Mode | BERTScore (F1) | NLI (Entailment) | BLEU |
|---|---|---|---|
| RQ-RAG | 0.5838 | 0.6440 | 0.0178 |
| vanilla RAG | 0.5569 | 0.4606 | 0.0109 |
| E$^2$RAG (comb. embedding) | 0.5520 | 0.4914 | 0.0102 |
| E$^2$RAG (hyp. embedding) | 0.5515 | 0.5256 | 0.0104 |
| E$^2$RAG (comb. extraction) | 0.5515 | 0.4622 | 0.0104 |
| E$^2$RAG (vanilla) | 0.5504 | 0.4814 | 0.0097 |
| E$^2$RAG (hyp. extraction) | 0.5492 | 0.4604 | 0.0103 |
| LightRAG local | 0.5491 | 0.4634 | 0.0094 |
| LightRAG global | 0.5489 | 0.4870 | 0.0098 |
| LightRAG hybrid | 0.5484 | 0.4821 | 0.0095 |
| vanilla HyDE | 0.5446 | 0.5311 | 0.0084 |
| GraphRAG global | 0.5399 | 0.2628 | 0.0133 |
| LightRAG hybrid (hyp. extraction) | 0.5311 | 0.5167 | 0.0070 |
| LightRAG hybrid (hyp. embedding) | 0.5311 | 0.5304 | 0.0071 |
| LightRAG hybrid (comb. extraction) | 0.5310 | 0.5204 | 0.0070 |
| LightRAG hybrid (comb. embedding) | 0.5306 | 0.5246 | 0.0070 |
| GraphRAG local | 0.5147 | 0.6698 | 0.0124 |
| GraphRAG drift | 0.5080 | 0.6898 | 0.0113 |

*Model checkpoints (Hugging Face):* BERTScore uses `microsoft/deberta-xlarge-mnli`. NLI uses `cross-encoder/nli-deberta-v3-large`.

**Interpretation.** The discrepancy is stark:

- RQ-RAG, which was ranked last by LLM judges for producing simplistic or incorrect answers, is paradoxically ranked first on both BERTScore and BLEU score.
- Conversely, our top-performing method, E$^2$RAG, ranks only modestly or even poorly across all three deterministic metrics.
- GraphRAG drift is ranked first by NLI but last by BERTScore, which shows that these metrics can reward opposing signals. We should therefore exercise caution when interpreting their numbers.

In contrast, the individual rankings from our three LLM judges show strong consistency (see Appendix K), which makes them more reliable for this task.

This mismatch shows that metrics focused on lexical overlap or static entailment are poor proxies for the temporal and causal reasoning that ChronoQAmeasures. This analysis strengthens the justification for our LLM-based evaluation.

## Q Generalization to an External Temporal Benchmark

We evaluated on 236 *Complex-TR* items with context length greater than 10,000 tokens. We report average judge scores using the same three LLM judges as in the main evaluation in Table 27. As the results show, our methods rank at the top. While GraphRAG drift is marginally better than one E$^2$RAG (hyp. embedding), all other E$^2$RAG variants outperform it. Notably, the strong performance of E$^2$RAG (vanilla) indicates that our core method can perform well even without the hypothetical document generation component on this benchmark.

Table 27: Results on long-context subset of Complex-TR (236 items).

| Mode | Avg. Score |
|---|---|
| E$^2$RAG (comb. extraction) | 6.2048 |
| E$^2$RAG (vanilla) | 6.1872 |
| E$^2$RAG (comb. embedding) | 6.1412 |
| E$^2$RAG (hyp. extraction) | 6.0904 |
| GraphRAG drift | 6.0692 |
| E$^2$RAG (hyp. embedding) | 6.0664 |
| LightRAG hybrid | 6.0383 |
| vanilla RAG | 6.0255 |
| vanilla HyDE | 6.0141 |
| GraphRAG local | 5.9901 |
| LightRAG global | 5.9745 |
| LightRAG local | 5.9191 |
| LightRAG hybrid (hyp. extraction) | 5.6662 |
| LightRAG hybrid (hyp. embedding) | 5.6620 |
| LightRAG hybrid (comb. extraction) | 5.6577 |
| LightRAG hybrid (comb. embedding) | 5.6549 |
| GraphRAG global | 5.6257 |
| RQ-RAG | 1.0170 |

## R Free-Form vs. Multiple-Choice: A Demonstration

We convert a ChronoQA item into multiple choice and compare outcomes.
**Question (Les Misérables).** When Valjean rescues Marius from the barricade, does it occur before Javert's death, and how does the script clarify the order to avoid mixing future details with the immediate rescue?
**Options.** A) No, Javert's death occurs first, ...
B) Valjean saves Marius first, ... later Javert commits suicide, ...
C) The events are shown simultaneously, ...
D) Valjean begins the rescue, ... Javert's death occurs in the middle, ...
**Observations.** In free-form, the vanilla RAG baseline received judge scores of 1, 3, and 3. In multiple choice, the same system selects the ground truth option (B) easily. This illustrates that MCQ format can mask reasoning weaknesses that ChronoQA exposes.

## S Native Long Context vs. RAG Augmentation

Even strong long context models benefit from retrieval guidance, likely because effective context use can lag behind maximum window length.

Table 28: Native long-context performance and augmentation with E$^2$RAG.

| Setting | Avg. Score |
|---|---|
| Claude-3.5-Sonnet (native long context, no RAG) | 7.0604 |
| Gemini-2.5-Pro (native long context, no RAG) | 8.5323 |
| Gemini-2.5-Pro + E$^2$RAG (comb. embedding) | 8.6262 |