F-GraphRAG: Preserving Fine-Grained Facts in Graph Retrieval-Augmented Generation

Shenmeng Chen[1] **Kang Zhang**[2,*]
[1]School of Civil Engineering, Harbin Institute of Technology, China
[2]Graduate School of Information, Production and Systems, Waseda University, Japan
24S133075@stu.hit.edu.cn zhangkang@toki.waseda.jp

Abstract

Retrieval-Augmented Generation (RAG) mitigates factual hallucinations in Large Language Models but struggles with complex reasoning. Graph-based RAG (GraphRAG) addresses this gap yet can exhibit a failure pattern: semantic compression during triple extraction may reduce the preservation of fine-grained facts such as precise quantities, temporal constraints, and entity attributes. We formalize this limitation as the Fine-Grained Fact Retrieval (FGFR) cascade and introduce stage-wise diagnostic metrics, including R_{rep}, R_{sel}, and R_{end}, to quantify information loss. To stress-test FGFR, we construct a diagnostic evaluation suite from classic novels that provide controlled long-context scenarios containing explicit entity attributes and temporal constraints. F-GraphRAG, our framework, features a granularity-aware extractor that preserves atomic details and an information-preserving filter that counters global importance bias. On this evaluation suite, F-GraphRAG outperforms GraphRAG by 12.6% in accuracy while using 33.8% fewer relationships, and achieves gains of 5.3–11.2% across nine public benchmarks. These results suggest that improving evidence completeness can provide a more effective alternative to increasing graph connectivity.

1 Introduction

Large Language Models (LLMs) are prone to factual hallucinations and constrained by static knowledge bases, limiting their application in precision-critical domains (Ji et al., 2023; Zhang et al., 2024; Liang et al., 2024; Li et al., 2023). Retrieval-Augmented Generation (RAG) mitigates these issues by dynamically retrieving external evidence (Lewis et al., 2020; Karpukhin et al., 2020; Guu et al., 2020) but often struggles with complex multi-hop reasoning. Existing structured retrieval-augmented generation frameworks, such as GraphRAG (Edge et al., 2024), construct unified knowledge graphs for superior thematic discovery (Pan et al., 2024). However, this structural shift introduces a critical trade-off: the potential loss of fine-grained facts (e.g. , numerical values, temporal constraints, attributes).

We term this limitation the **Fine-Grained Fact Retrieval (FGFR)** problem, where fine-grained details that are explicitly present in the source corpus may be lost during graph-based retrieval. This loss is particularly pronounced for precise quantities, temporal constraints, and entity attributes, and it motivates a more careful examination of how graph construction and retrieval affect evidence preservation.

Preliminary evaluation further suggests that this failure is underrepresented in conventional QA benchmarks: GraphRAG performs strongly on standard benchmarks but degrades substantially on our fine-grained diagnostic setting. We therefore construct FGQA to explicitly stress-test attribute-level evidence retention (Liu et al., 2024).

To mitigate FGFR, we propose **F-GraphRAG**, which jointly improves fine-grained evidence preservation during graph construction and retrieval. Our contributions are threefold:

1. **Identifying fine-grained fact retrieval degradation in graph-based RAG.** We formulate the Fine-Grained Fact Retrieval (FGFR) cascade and introduce stage-wise diagnostic metrics to quantify information loss in graph retrieval.
2. **Introducing FGQA for fine-grained retrieval diagnosis.** We develop FGQA, a diagnostic evaluation suite designed to expose attribute retention failures that are overlooked by existing QA benchmarks.
3. **Proposing an evidence-preserving graph construction and retrieval framework** that improves fine-grained evidence retention without increasing retrieval complexity.

*Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 48219–48237
October 24-29, 2026 ©2026 Association for Computational Linguistics

2 Related Work

GraphRAG has emerged as a general framework for graph-based retrieval augmentation (Edge et al., 2024; Han et al., 2024). Existing studies optimize efficiency (Guo et al., 2025b; AI, 2025; Zhou et al., 2025), reasoning paths (Chen et al., 2026; Wu et al., 2025; Zhao et al., 2025), or graph organization (Zhu et al., 2025; Hu et al., 2025; Sharma et al., 2025; Gutiérrez et al., 2024), typically relying on compressed relational representations (Huang et al., 2025; Xu et al., 2024; Feng et al., 2026). However, few explicitly optimize for the preservation of query-critical attributes, and existing benchmarks (Liang et al., 2026) and surveys (Pan et al., 2024; Peng et al., 2025) lack per-stage diagnostics. We address three gaps: (1) *construction*, preserving attributes during extraction; (2) *retrieval*, prioritizing evidence completeness; and (3) *evaluation*, providing stage-wise metrics (R_{rep}, R_{sel}, R_{end}) and the FGQA suite for root-cause analysis.

Beyond retrieval efficiency and graph topology, another relevant line of research concerns how structured representations preserve information from source documents. Knowledge graph construction inevitably abstracts natural-language evidence into entities and relations, which improves structural accessibility but may remove modifiers, numerical constraints, and contextual qualifiers that are important for downstream questions. Existing GraphRAG evaluations typically measure final answer quality or retrieval relevance, making it difficult to distinguish whether an error originates from graph construction or retrieval selection. In contrast, our work explicitly separates these two sources of information loss and evaluates them with stage-wise retention metrics. This perspective shifts the focus from retrieving more connected evidence to preserving the completeness of query-critical evidence throughout the graph pipeline.

3 Problem Formulation

3.1 GraphRAG Pipeline and FGFR Formalization

A standard GraphRAG pipeline operates on a document corpus D in three stages:

1. **Graph Extraction** Extract : $D \to G$, producing a knowledge graph $G = (V, E)$;
2. **Context Retrieval** Retrieve : $(G, q) \to C_q$, selecting a subset of edges $C_q \subseteq E$ relevant to query q;
3. **Answer Generation** Generate : $(C_q, q) \to a$, producing the final answer a.

We define a fine-grained fact as an attributed quadruple

$$f = (s, p, o, A), \tag{1}$$

where $s, o \in V$ are subject and object entities, p is a predicate, and $A = \{(\tau_1, \nu_1), \ldots, (\tau_k, \nu_k)\}$ is a set of attribute–value pairs encoding task-critical details such as time, quantity, location, material, manner, or purpose. Standard triple extraction retains only the (s, p, o) skeleton and may discard A, introducing a representational bottleneck.

Let F_q denote the set of fine-grained facts that are *essential* to answer query q. We assume that every $f \in F_q$ is explicitly stated in the source corpus D. For diagnosis, for each required fact $f = (s, p, o, A) \in F_q$, we designate a subset $A_q \subseteq A$ of *query-relevant attributes* determined via expert annotation (Cohen's $\kappa = 0.88$; see Appendix C).

Definition 1 (FGFR Failure)**.** We define FGFR as a class of failures where, after extraction and retrieval, at least one fact $f \in F_q$ is either missing entirely from C_q or present in an incomplete form—that is, some attribute–value pair in A_q is lost. A fact is considered completely retrieved only when all query-relevant attributes A_q are preserved in C_q. FGFR is thus a structural representation failure that originates from information loss inside the pipeline.

FGFR manifests as a two-stage cascade:

1. **Representation failure:** Semantic compression during extraction strips away attributes A, so that the fact is not fully preserved in G.
2. **Selection failure:** Graph-based retrieval heuristics (*e.g.*, node degree, PageRank) favor high-frequency relations and prune rare but query-critical edges that carry fine-grained details.

Breaking this cascade requires targeted interventions at both the representation and selection stages.

Under a fixed token budget B (typically 4096 tokens), our objective is to maximize the end-to-end retention of fine-grained facts:

$$\max \ \mathbb{E}_{q \sim Q}\big[R_{\text{end}}(q)\big] \quad \text{s.t.} \quad \text{Len}(C_q) \le B, \tag{2}$$

where $R_{\text{end}}(q)$ denotes the fraction of essential facts fully preserved through both stages.

3.2 Diagnostic Analysis of FGFR

Graph construction can be viewed as an information bottleneck: the source text D is compressed into a set of relational triples, and only this compressed representation is available for downstream

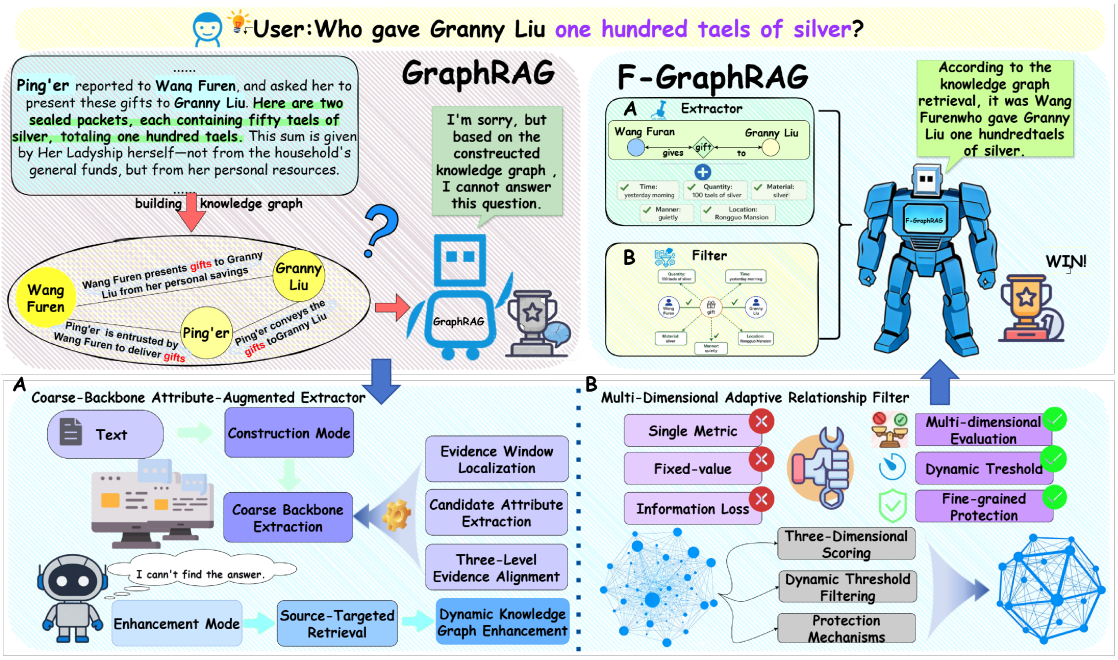


Figure 1: Overcoming FGFR: F-GraphRAG's adaptive components vs. GraphRAG's failure cases.

retrieval (Tishby et al., 2000). If the compression discards attributes that later become relevant to a query, no retrieval policy can recover them. To quantify this loss, we define the *fact compression ratio* for a fact f as

$$CR(f) = \frac{|A_{\text{source}}(f)| - |A_{\text{graph}}(f)|}{|A_{\text{source}}(f)|}, \quad (3)$$

where $A_{\text{source}}(f)$ is the set of attributes explicitly stated in the source text for fact f, and $A_{\text{graph}}(f)$ is the set of attributes retained in the graph representation. Table 1 reports the empirical loss probability for each attribute type, computed over 500 manually annotated diagnostic facts. Quantity and material are the most frequently discarded, while location and descriptor are relatively preserved. This analysis provides direct evidence that triple abstraction systematically biases against certain classes of fine-grained information.

To attribute and quantify information loss across the two stages, we introduce three diagnostic metrics. Let $F_q^G \subseteq F_q$ be the subset of ground-truth facts fully represented in the extracted graph G.

Extraction Integrity $R_{\text{rep}}(q)$ measures the fraction of facts that survive semantic compression:

$$R_{\text{rep}}(q) = \frac{|F_q^G|}{|F_q|}. \quad (4)$$

Filtering Success $R_{\text{sel}}(q)$ measures the fraction of extracted facts retrieved into C_q:

$$R_{\text{sel}}(q) = \begin{cases} \dfrac{|F_q^G \cap C_q^*|}{|F_q^G|}, & F_q^G \neq \emptyset, \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

Here C_q^* denotes the set of facts whose corresponding relation and all query-relevant attributes are present in C_q. Although attributes are atomic units during annotation, we aggregate them into fact-level retention because downstream QA requires the joint preservation of all query-critical attributes. R_{sel} is a conditional metric that isolates selection-stage loss given successful representation; further discussion is provided in Appendix E.

End-to-End Retention $R_{\text{end}}(q)$ is the product of the two stage-wise metrics:

$$R_{\text{end}}(q) = R_{\text{rep}}(q) \times R_{\text{sel}}(q). \quad (6)$$

Table 1: Attribute loss probability under standard triple extraction (500 annotated facts).

Attribute Type	Loss Probability
Quantity	67%
Material	58%
Temporal	49%
Manner	43%
Purpose	35%
Appearance	32%
Location	31%
Descriptor	27%

Note: The fraction of facts where an attribute present in the source is missing from the graph.

This multiplicative composition captures the joint probability that a required fact is represented and then retrieved. R_{end} is the primary metric for end-to-end evaluation.

We report macro-averaged values $\overline{R_{\text{rep}}}$, $\overline{R_{\text{sel}}}$, and $\overline{R_{\text{end}}}$ across all queries in a benchmark. A preliminary analysis using the standard GraphRAG pipeline reveals severe degradation at both stages.

4 Methodology

4.1 Notation and Framework Overview

Our framework reuses the same property graph storage and indexing backend as standard GraphRAG. The key difference is that we materialize the attribute set A as edge properties. The document corpus is denoted as D, and the knowledge graph as $G = (V, E)$. Each relation $r \in E$ may carry an attribute set $A_r = \{(\tau_1, \nu_1), \dots\}$.

F-GraphRAG contains two coordinated components (Figure 1): a **granularity-aware extractor** for improving R_{rep}, and an **information-preserving filter** for improving R_{sel}. In rare cases where no edges are retrieved, a fallback dense retrieval is used (see Appendix D.5); this safeguard is not a core architectural component and its contribution is examined in the ablation study.

4.2 Granularity-Aware Relationship Extraction

To mitigate semantic compression, we formulate graph construction as an evidence-preserving objective. The extractor aims to preserve the completeness of evidential facts by assigning each relation an *attribute completeness score*:

$$C(r) = \sum_{\tau \in \mathcal{T}} w_\tau \cdot \mathbb{1}[(\tau, \cdot) \in A_r], \qquad (7)$$

where $\mathcal{T}$ is the set of eight attribute types (time, quantity, location, manner, material, purpose, descriptor, appearance; details in Appendix D.1) and w_τ are type-specific weights (default 1.0). This score measures the informational richness of a relation independently of any specific query.

The coarse triple extractor additionally produces a **base confidence score** $S_{\text{base}}(r)$ for each triple, reflecting its syntactic and semantic plausibility. We then combine this base score with the attribute completeness score to obtain a priority:

$$\text{priority}(r) = S_{\text{base}}(r) + \alpha \cdot C(r), \qquad (8)$$

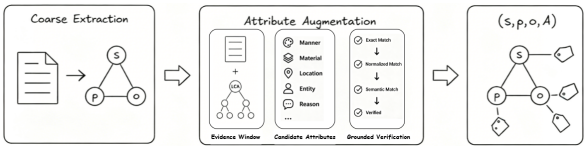


Figure 2: Coarse backbone extraction followed by evidence-grounded attribute augmentation.

Algorithm 1 Coarse Backbone + Evidence-Grounded Attribute Augmentation

Require: Sentence t, coarse triple extractor $\mathcal{E}_c$, attribute extractor $\mathcal{E}_a$
Ensure: Set of relations $\mathcal{R}$

function EXTRACT(t)
  $\mathcal{R}_c \leftarrow \mathcal{E}_c(t)$ ▷ Each triple (s, p, o) includes base score S_{base}
  $\mathcal{R} \leftarrow \emptyset$
  for $(s, p, o, S_{\text{base}}) \in \mathcal{R}_c$ **do**
    window $\leftarrow$ getEvidenceWindow(t, s, o)
    candidates $\leftarrow \mathcal{E}_a$(window)
    $A \leftarrow \emptyset$
    for $(\tau, \nu) \in$ candidates **do**
      if aligned$(\nu,$ window$)$ **then**
        $A \leftarrow A \cup \{(\tau, \nu)\}$
      end if
    end for
    $C(r) \leftarrow \sum_\tau w_\tau \cdot \mathbb{1}[(\tau, \cdot) \in A]$
    priority$(r) \leftarrow S_{\text{base}} + \alpha \cdot C(r)$
    $\mathcal{R} \leftarrow \mathcal{R} \cup \{(s, p, o, A, \text{priority})\}$
  end for
  Sort $\mathcal{R}$ by priority descending
  Retain top $K = 50$ relations per 1k tokens
  return $\mathcal{R}$
end function

where α controls the trade-off between base relation quality and attribute richness. This formulation ensures that the parameter α directly affects the ranking of candidate relations, making the subsequent sensitivity analysis meaningful. The default value is $\alpha = 0.5$ (selected via Bayesian optimization, see Appendix D.8). Candidate relations are ranked by priority within each text chunk, and only the top-K relations ($K = 50$) are retained. Algorithm 1 formalizes the process.

4.3 Information-Preserving Relationship Selection

To address retrieval-stage evidence loss, we formulate relationship selection as an information preservation problem rather than a popularity ranking problem. The filtering process consists of two steps: **coarse pruning** and **adaptive fine selection**. Coarse pruning obtains a k-hop candidate subgraph around query entities, while adaptive fine selection applies scoring and dynamic thresholding. We obtain seed entities by prompting the LLM (DeepSeek-V3, temperature 0.1) to extract entity

Table 2: Recall of query-critical facts by degree group.

Edge Degree Group	GraphRAG	F-GraphRAG
High	72%	79%
Medium	45%	68%
Low	20%	65%

Note: Recall indicates the percentage of query-critical facts successfully retrieved into the context.

mentions from the query and link them to graph nodes via exact match or Levenshtein fuzzy matching (threshold 0.9). The same LLM decomposes the query intent into high-level conceptual keywords (e.g. , "funeral") and low-level attribute keywords (e.g. , "49 days"), which respectively guide k-hop candidate expansion and scoring in the subsequent filtering stage. Importantly, entity extraction and query decomposition are query parsing steps, not generation; for fair comparison, the identical LLM backend is used for all graph-based methods that require seed entity extraction.

For each candidate relation, we compute three complementary scores: S_{attr} for evidence completeness, S_{ling} for query compatibility, and S_{net} for structural continuity (formal definitions in Appendix D.7). Their weighted combination is

$$S_{\text{total}}(r) = w_{\text{attr}} \cdot S_{\text{attr}}(r) + w_{\text{ling}} \cdot S_{\text{ling}}(r) + w_{\text{net}} \cdot S_{\text{net}}(r), \tag{9}$$

where $w_{\text{attr}}, w_{\text{ling}}, w_{\text{net}}$ are query-level weights derived from score separability, with a fixed fallback weighting used when separability is degenerate.

A leave-one-out analysis on a 500-query diagnostic subset (Table 3) shows that the evidence completeness dimension contributes most to R_{sel} improvement (+18 pp), followed by the query compatibility dimension (+9 pp) and the structural continuity dimension (+5 pp). The structural continuity dimension thus serves a protective role and does not reintroduce global-importance bias.

Low-frequency fact analysis. To directly test the claim that global importance bias harms rare facts, we partition the diagnostic facts into three groups based on the degree of their corresponding graph edge: high, medium, and low. Table 2 reports the recall of these groups by GraphRAG and F-GraphRAG. Rare facts (low degree) are disproportionately lost by GraphRAG (only 20% recall), whereas F-GraphRAG recovers 65%. This supports the design of our information-preserving filter, which explicitly rewards attribute completeness rather than structural prominence.

Table 3: Leave-one-out ablation of information-preserving filtering dimensions (500-query subset).

Removed Dimension	R_{sel}	ΔR_{sel}
Full filter (no removal)	73.0	–
– Evidence completeness	55.0	–18.0
– Query compatibility	64.0	–9.0
– Structural continuity	68.0	–5.0

Note: Evidence completeness is the primary driver; structural continuity serves a protective role.

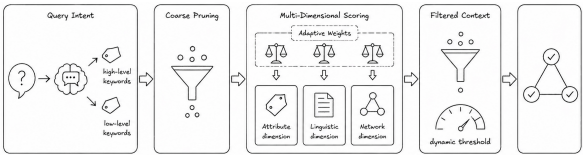


Figure 3: Information-preserving filtering pipeline with three-dimensional scoring and adaptive thresholding.

The comprehensive score is used with an adaptive threshold $\theta = \text{mean}(\mathcal{S}) + 1.5 \cdot \text{std}(\mathcal{S})$ to select the final edge set, with explicit protection for semantic outliers and structural bridges. The full filtering procedure is given in Algorithm 2 (Appendix D.6.1).

5 Experiments

5.1 Experimental Setup

FGQA Evaluation Suite. FGQA is a diagnostic evaluation suite for analyzing fine-grained fact retrieval failures in long-context graph-based retrieval systems. It contains 10,000 fine-grained questions from ten classic novels (five Chinese, five Western) and 1,000 global reasoning questions from *Dream of the Red Chamber*. The construction protocol is detailed in Appendix B and Appendix C.

Baselines and Metrics. We compare against GraphRAG (Edge et al., 2024), LightRAG (Guo et al., 2025b), FastGraphRAG (AI, 2025), E^{2}GraphRAG (Zhao et al., 2025), GraphRAG-FI (Guo et al., 2025a), QCG-RAG (Wu et al., 2025), and CheapRAG (Zhou et al., 2025), as well as 16 dense retrieval models (complete list in Appendix J). We evaluate accuracy as factual correctness using an LLM-based evaluator with human verification (agreement $\kappa = 0.82$). All graph-based methods use the same embedding model (`text-embedding-3-large`), generation model (DeepSeek-V3, temperature=0.1), and context budget of $B = 4096$ tokens. Dense retrieval baselines use their respective embedding models while sharing the same generation backbone, context budget, and evaluation protocol. Fur-

Table 4: Performance drops sharply on FGQA compared to standard QA benchmarks.

Dataset	Attr. Den.	GraphRAG	F-GraphRAG	Gap
HotpotQA	Low	91.3%	92.1%	0.8%
NQ	Low	89.8%	90.5%	0.7%
Podcast	Low	88.2%	88.9%	0.7%
News	Low	86.5%	87.2%	0.7%
FGQA	High	56.1%	68.7%	12.6%

Note: Attr. Den. = Attribute Density. Gap = F-GraphRAG − GraphRAG. On a high-density subset of News, the gap increases to 7.1% (see Appendix B.3).

ther implementation details (data split, hyperparameter tuning, seeds) are in Appendix F.

5.2 Main Results and Diagnostic Evidence

5.2.1 FGQA Diagnostic Evaluation

Table 4 shows that GraphRAG achieves high scores on standard benchmarks but drops sharply on FGQA, exposing the FGFR cascade. F-GraphRAG narrows this gap. Table 5 reports fine-grained QA accuracy; F-GraphRAG outperforms all baselines, achieving 68.7% accuracy.

A key finding is the strong correlation between attribute density and performance degradation. Across FGQA and the nine public benchmark subsets ($n = 10$), attribute density was positively correlated with the performance drop of GraphRAG relative to low-density benchmarks (Pearson $r = 0.78$, $p = 0.008$). This indicates that existing benchmarks with low attribute density mask the FGFR problem, while attribute-rich tasks (e.g., FGQA) reveal it.

Importantly, FGQA is intended as a diagnostic stress test rather than a replacement for conventional QA benchmarks. Standard benchmarks primarily evaluate whether a retriever can locate semantically relevant evidence, whereas FGQA additionally tests whether the retrieved representation preserves the attributes required to reconstruct the answer precisely. This distinction is critical for graph-based retrieval, because a relation may remain topically relevant even after quantities, materials, temporal qualifiers, or descriptive modifiers have been removed during graph construction. Consequently, strong retrieval performance on conventional benchmarks does not necessarily imply faithful preservation of fine-grained evidence. FGQA therefore complements existing evaluations by isolating a failure mode that is difficult to observe from end-to-end accuracy alone.

Table 5: Fine-Grained QA Accuracy on FGQA.

Category	Model	Accuracy
Graph Retrieval	**F-GraphRAG (Ours)**	**68.7% ± 0.8%**†
	F-GraphRAG w/o fallback	66.8% ± 0.9%
	GraphRAG-FI	64.4% ± 1.0%
	QCG-RAG	64.3% ± 1.0%
	E^{2}GraphRAG	62.1% ± 1.1%
	GraphRAG	56.1% ± 1.2%
	FastGraphRAG	59.8% ± 1.3%
	LightRAG	56.7% ± 1.4%
	CheapRAG	60.5% ± 1.1%
Dense Retrieval	text-embedding-3-large	65.9% ± 0.8%
	DashScope-text-embedding-v4	63.6% ± 0.7%
	stella-large–v2	61.7% ± 0.9%
	bge-large	60.3% ± 0.8%

Note: $^{\dagger}p < 0.001$ vs. GraphRAG (paired t-test with Bonferroni correction at query level).

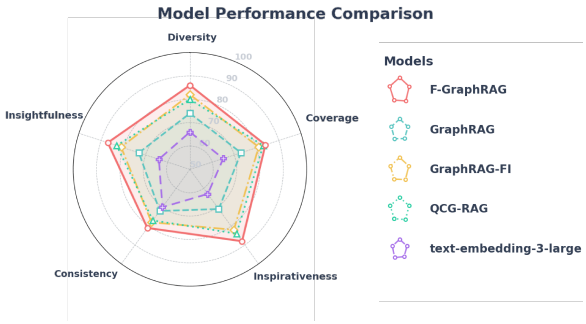


Figure 4: Radar chart comparing five representative methods on five global reasoning quality dimensions.

5.2.2 Cross-domain Validation

To evaluate whether FGFR mitigation transfers across domains, we construct fine-grained evaluation subsets from nine public benchmarks spanning biomedical (BioASQ (Tsatsaronis et al., 2015)), financial (ConvFinQA (Chen et al., 2022)), legal (CUAD (Hendrycks et al., 2021)), news (DROP (Dua et al., 2019)), governmental (GovReport (Huang et al., 2021)), historical (LongBench (Bai et al., 2024)), scientific (Qasper (Dasigi et al., 2021)), educational (RACE (Lai et al., 2017)), and technical (TechQA (Castelli et al., 2020)) domains. For each dataset, we construct two subsets: a random subset of 1,000 questions and a high-density subset of 1,000 questions sampled from the top quartile of attribute density. Both subsets are re-annotated with query-relevant attribute sets A_q following the same protocol (Appendix B.3). The random subsets preserve the original attribute density distribution, while the high-density subsets provide stress-test conditions.

As shown in Table 6, F-GraphRAG consis-

Table 6: Cross-domain validation of FGFR mitigation on curated fine-grained subsets.

Domain	Dataset	GraphRAG	F-GraphRAG
Novel	FGQA	56.1 (1.2)	**68.7 (0.8)**†
Biomedical	BioASQ	52.3 (1.8)	**63.5 (1.5)**†
Financial	ConvFinQA	48.7 (2.1)	**58.1 (1.7)**†
Legal	CUAD	55.2 (1.9)	**66.0 (1.6)**†
News	DROP	61.8 (1.5)	**68.9 (1.3)**†
Government	GovReport	57.4 (2.0)	**64.2 (1.8)**
Historical	LongBench	53.9 (2.2)	**62.5 (1.9)**†
Scientific	Qasper	59.6 (1.7)	**66.8 (1.4)**
Education	RACE	63.1 (1.4)	**70.2 (1.2)**
Technical	TechQA	56.8 (1.6)	**62.1 (1.5)**

Note: † indicates significant improvement at $p < 0.05$ by paired t-test with Bonferroni correction at query level.

tently improves performance on attribute-rich subsets, with gains observed on most domains. Improvements range from +5.3% to +11.2% on high-density subsets, and 5 out of 9 datasets achieve statistically significant improvements (paired t-test, $p < 0.05$ after Bonferroni correction). A similar trend holds for F-GraphRAG gains: across the same ten datasets, gains were larger in domains with higher attribute density (Pearson $r = 0.72$, $p = 0.02$). On random subsets, F-GraphRAG shows no degradation (all gains positive), confirming that it does not harm general retrieval scenarios. Detailed random-subset results are in Appendix L.

Beyond domain generalization, we also examine robustness across different generation backbones. Keeping the graph construction, attribute extraction, and retrieval modules unchanged, we replace only the answer generation model and evaluate on an FGQA subset (500 queries). Table 7 reports consistent gains for F-GraphRAG across three generation models (+9.5% to +11.8%), confirming that improvements stem from evidence quality in the retrieval stage rather than model-specific preferences. This robustness follows from three design factors: (1) the core innovation operates at the retrieval stage, decoupled from the generation model; (2) the LLM calls in our framework (query decomposition, entity linking) are auxiliary structured extraction tasks whose outputs are constrained by the input query and graph evidence; (3) the consistency of gains across models reduces the likelihood of model-specific artifacts.

5.2.3 Ablation and Component Contributions

We isolate the contributions of the two core components by removing each. Table 8 shows that removing the information-preserving filter costs

Table 7: Accuracy across generation models on an FGQA subset (500 queries).

Generation Model	GraphRAG	F-GraphRAG	Δ
DeepSeek-V3	58.1%	69.0%	+10.9%
GPT-4o	58.3%	70.1%	+11.8%
Qwen2-72B	54.7%	64.2%	+9.5%

Note: Δ = F-GraphRAG − GraphRAG. Standard deviations are provided in Appendix M.

Table 8: Accuracy under component ablation.

Configuration	Extractor	Filter	FFR	Accuracy
F-GraphRAG	✓	✓	✓	68.7 (0.8)
w/o FFR	✓	✓	×	66.8 (0.9)
w/o filter	✓	×	✓	65.8 (0.9)
w/o extractor	×	✓	✓	61.4 (1.0)
w/o filter + FFR	✓	×	×	65.3 (0.9)
w/o extractor + FFR	×	✓	×	61.1 (1.1)
w/o extractor + filter	×	×	✓	58.2 (1.1)
GraphRAG	×	×	×	56.1 (1.2)

Note: ✓= present; × = removed. FFR = Fallback Fact Retrieval.

2.9 points, removing the granularity-aware extractor costs 7.3 points, and removing both components (while retaining the fallback) reduces accuracy to 58.2%, a 10.5-point drop. Fully reverting to the GraphRAG baseline (removing both components and the fallback) yields a total gap of 12.6 points. The fallback contributes +1.9 points.

Stage-wise diagnostic metrics on a 500-query annotated subset (Table 9) confirm the root-cause attribution: the extractor lifts R_{rep} from 58% to 76%, the filter boosts R_{sel} from 52% to 73%, and their combination yields $R_{\text{end}} = 55\%$ (vs. 30% for GraphRAG). Root-cause analysis attributes 62% of failures to representation loss and 24% to selection failure, aligning with our two-pronged design.

5.2.4 Stage-wise Retention Analysis

To further link diagnostic metrics to task performance, we analyze the relationship between $R_{\text{end}}(q)$ and answer accuracy. We partition the 500 annotated diagnostic queries into quintiles based on their R_{end} values. Table 10 shows a clear monotonic increase in accuracy from the lowest to the highest quintile, confirming that end-to-end fact retention is a strong predictor of generation quality. This trend also supports the diagnostic validity of R_{end}: the metric is not merely sensitive to intermediate graph properties, but tracks whether preserved evidence ultimately contributes to correct answers.

Table 9: Stage-wise diagnostic metrics under progressive component removal.

Configuration	E	F	R_{rep}	R_{sel}	R_{end}
F-GraphRAG	✓	✓	76(3.8)	73(4.3)	55(4.5)
− Filter	✓	×	76(3.8)	62(5.0)	51(4.2)
− Extractor	×	✓	58(4.9)	67(4.3)	42(4.2)
− Extractor & Filter	×	×	58(4.9)	52(5.0)	30(4.6)

Note: ✓= present; × = removed.

Table 10: Correlation between $R_{\text{end}}(q)$ and accuracy.

Quintile	Mean R_{end}	Accuracy
1 (Lowest)	0.15	35.0% ± 3.1%
2	0.38	55.0% ± 3.8%
3	0.55	70.0% ± 4.2%
4	0.75	85.0% ± 3.5%
5 (Highest)	0.92	98.5% ± 2.8%

5.2.5 Hyperparameter Sensitivity

Hyperparameter sensitivity analyses (Tables 11 and 12) show filter coefficients vary at most 0.7 percentage points in accuracy under $\pm 20\%$ perturbations. For priority coefficient α, a $\pm 20\%$ perturbation around selected value $\alpha = 0.5$ results in a maximum accuracy decrease of 1.6 percentage points.

5.3 Efficiency Analysis

Construction Cost Trade-off. Unlike standard GraphRAG, F-GraphRAG introduces additional offline extraction cost, amortized over repeated queries (Table 13).

Token and Relationship Efficiency. F-GraphRAG achieves substantial token savings over standard GraphRAG as query volume grows, driven by using 33.8% fewer relationships per query (196 vs. 296). Table 14 reports cumulative token consumption; by the 100th query, F-GraphRAG has consumed 1,872.4k tokens vs. 2,294.5k for GraphRAG—a reduction of 18.4%. The reduction stems from the granularity-aware extractor assigning higher priority to attribute-carrying relations and the information-preserving filter selecting edges based on query-adaptive, multi-dimensional scoring. Despite using fewer edges, the retained relations carry more fine-grained attributes, as confirmed by the diagnostic metrics in Table 9.

5.4 Case Study

To illustrate the FGFR cascade and the effect of F-GraphRAG, consider the query *"How long did the Jia family prepare for the funeral ceremony, and what materials were used?"* The source text states: *"Jia family spent 49 days preparing a funeral ceremony with imperial-grade materials."* Table 15 contrasts the graph representations and answers.

Table 11: Sensitivity of accuracy to coefficient perturbations in the information-preserving filter.

Coefficient	–20%	–10%	Optimal	+10%	+20%
$\lambda_{a1} = 0.55$	68.0	68.4	**68.7**	68.5	68.1
$\lambda_{a2} = 0.45$	68.1	68.5	**68.7**	68.4	68.0
$\lambda_{l1} = 0.35$	68.3	68.5	**68.7**	68.6	68.2
$\lambda_{l2} = 0.40$	68.2	68.5	**68.7**	68.5	68.3
$\lambda_{l3} = 0.25$	68.4	68.6	**68.7**	68.5	68.2
$\lambda_{n1} = 0.65$	68.1	68.4	**68.7**	68.4	68.0
$\lambda_{n2} = 0.35$	68.3	68.6	**68.7**	68.5	68.2

Note: All values under full F-GraphRAG configuration, evaluated on the development set.

Table 12: Sensitivity to the priority coefficient α.

α	Dev Accuracy	R_{rep}	Graph size
0.2	65.0 (0.9)	71	38.2
0.3	66.5 (0.8)	73	41.5
0.4	67.1 (0.8)	76	44.7
0.5	**68.7 (0.7)**	**77**	**48.9**
0.6	68.5 (0.9)	78	50.0

Note: Graph size is in relations per 1k tokens, measured after global relation deduplication. Changing α alters which relations survive local Top-K selection and therefore changes the number of unique relations after merging. Top-K is a hard upper bound of 50; average size may be slightly below it due to chunks with fewer candidates. $\alpha = 0.5$ is selected for its best accuracy and high R_{rep}; although $\alpha = 0.6$ has a higher R_{rep}, it lowers accuracy and enlarges the graph.

5.5 Global Reasoning Analysis

FGFR primarily affects atomic fact preservation, but graph-based RAG is also designed for global synthesis. We therefore conduct a lightweight additional analysis to assess whether preserving fine-grained evidence can benefit higher-level narrative understanding. We use 1,000 queries from *Dream of the Red Chamber* and a pairwise comparison protocol. Two domain experts independently judge anonymized method outputs on five dimensions: Coverage, Diversity, Insightfulness, Consistency, and Inspirativeness. These dimensions are grounded in prior work: Coverage and Diversity follow the GraphRAG evaluation framework (Edge et al., 2024); Consistency draws on FActScore (Min et al., 2023); Inspirativeness aligns with MT-Bench (Zheng et al., 2023); Insightfulness addresses deeper interpretation as in long-form summarization. These dimensions are also aligned with our FGFR analysis: Coverage and Consistency mainly reflect whether fine-grained evidence survives extraction, whereas Insightfulness and Inspirativeness evaluate whether semantically rich

Table 13: Computational cost comparison.

Metric	GraphRAG	F-GraphRAG
Offline build time (h)	37.5	45.0
Offline build tokens (M)	32.5	38.75
Online latency	baseline	baseline + 3–5%

Note: The additional offline cost originates from attribute extraction (approx. 30K extra LLM calls) and one-time Bayesian optimization.

Table 14: Cumulative token consumption on FGQA (k tokens), sampled every 10 queries.

Num. Queries	F-GraphRAG	GraphRAG	Savings
10	179.9	218.2	38.3
20	366.5	453.9	87.4
30	551.2	690.5	139.3
40	742.3	904.2	161.9
50	919.4	1,131.8	212.4
60	1,117.8	1,367.5	249.7
70	1,306.7	1,590.5	283.8
80	1,489.2	1,828.8	339.6
90	1,683.9	2,057.4	373.5
100	1,872.4	2,294.5	**422.1**

Note: Savings = GraphRAG − F-GraphRAG (k tokens).

but low-frequency relations are effectively utilized during reasoning. Judges are forced to select a winner for each pair (no ties); the final win-rate matrix aggregates over all queries and experts (Cohen's $\kappa = 0.84$). Figure 4 visualizes representative win-rates. The full evaluation protocol with dimension definitions, a case analysis (Table G.5), and complete pairwise matrices are in Appendix G. We emphasize that these qualitative results provide preliminary evidence that richer factual grounding can benefit reasoning evaluation, rather than serving as the primary contribution.

5.6 Evidence Reliability

Although all methods use the same generation backbone, LLM-based generation may produce outputs under stochastic decoding. We examine whether higher-quality retrieved evidence can provide more reliable grounding across generations. Using a diagnostic subset of 500 queries, we run each system ten times with random seeds and measure factual consistency, defined as the percentage of queries whose answers remain factually equivalent across runs, as judged by the same LLM-based factual evaluator used in our main evaluation. As shown in Table 16, F-GraphRAG achieves higher factual consistency than GraphRAG, because the attribute-preserving graph construction supplies the generator with more complete and less ambiguous evidence, reducing reliance on incomplete grounding.

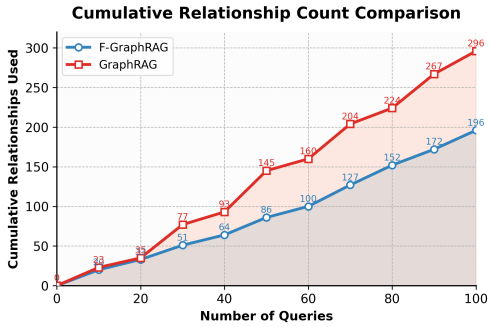


Figure 5: Cumulative relationships over 100 queries.

Table 15: Case study of a fine-grained fact through GraphRAG and F-GraphRAG.

Component	Content
Source text	"Jia family spent 49 days preparing a funeral ceremony with imperial-grade materials."
GraphRAG extraction	(Jia family, held, funeral ceremony)
F-GraphRAG extraction	(Jia family, held, funeral ceremony, {time=49 days, material=imperial-grade, purpose=ritual completion})
GraphRAG answer	Incorrect: cannot answer the duration or material.
F-GraphRAG answer	Correct: 49 days; imperial-grade materials.

Note: This case is representative of the 62% of failures attributed to representation loss in our diagnostic analysis.

Table 16: Factual consistency across ten independent runs on 500 diagnostic queries.

Method	Factual Consistency (%)	Std
GraphRAG	61.2	4.8
F-GraphRAG	**73.5**	**3.2**

This result provides evidence that improving evidence completeness at the representation stage leads to more reliable downstream generation.

6 Conclusion

This work identified fine-grained factual degradation in graph-based RAG as a two-stage FGFR cascade driven by semantic compression and global-importance bias. Our framework mitigates through evidence-preserving extractor that retains query-critical attributes and information-preserving filter that prioritizes evidence completeness over structural prominence, improving accuracy and efficiency without enlarging the graph. Results across literary and cross-domain tests confirm evidence preservation as a direction for graph-based RAG.

7 Limitations

The study establishes the FGFR diagnostic framework and demonstrates F-GraphRAG's effectiveness; several natural extensions can strengthen the framework and broaden its applicability.

Benchmark Contamination. As with many literary benchmarks, some source texts may overlap with the pretraining data of modern LLMs. We mitigate this by evaluating under retrieval-only settings, where answers must be grounded in retrieved evidence rather than parametric memory. Closed-book experiments confirm that retrieval substantially improves performance for both GraphRAG and F-GraphRAG, indicating that reported gains reflect retrieval quality rather than model memory. Future work will extend FGQA to post-training-cutoff or proprietary corpora.

Attribute-Level Evaluation and Context Modeling. The granularity-aware extractor operates at the sentence level, a deliberate design choice that provides high precision and language portability. Incorporating cross-sentence context, for example through coreference-aware windows, is a promising direction for modeling longer factual chains. Additionally, the diagnostic metrics R_{rep}, R_{sel}, and R_{end} rely on expert-annotated attribute sets A_q; automating this annotation would further scale the evaluation to broader benchmarks.

Backbone and Language Generalization. The current implementation uses DeepSeek-V3 for extraction, but the evidence-preserving strategy is agnostic to the backbone. Future work will validate this with rule-based or smaller open models to confirm that gains originate from the strategy rather than the specific LLM. The window mechanism is language-agnostic by design; embedding-based span detection could further broaden deployment across varied linguistic settings.

Incremental Graph Construction. The knowledge graph is constructed offline to ensure reproducibility and evaluation. Extending the pipeline to streaming or incremental indexing is a direction for real-time applications and growing corpora.

References

Circlemind AI. 2025. Fastgraphrag: Lightweight and efficient graph retrieval-augmented generation. Version: 0.4.1.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, and 1 others. 2024. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 3119–3137.

Vittorio Castelli, Rishav Chakravarti, Saswati Dana, Anthony Ferritto, Radu Florian, Martin Franz, Dinesh Garg, Dinesh Khandelwal, J Scott McCarley, Michael McCawley, and 1 others. 2020. The techqa dataset. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 1269–1278.

Boyu Chen, Zirui Guo, Zidan Yang, Yuluo Chen, Junze Chen, Zhenghao Liu, Chuan Shi, and Cheng Yang. 2026. Pathrag: Pruning graph-based retrieval augmented generation with relational paths. In *Proceedings of the AAAI conference on artificial intelligence*, volume 40, pages 30183–30191.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 6279–6292.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378.

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*.

Yifan Feng, Hao Hu, Shihui Ying, Xingliang Hou, Shiquan Liu, Mingyuan Yang, Junchang Li, Shaoyi Du, Nanning Zheng, Han Hu, and 1 others. 2026. Hyperrag: Combating llm hallucinations using hypergraph-driven retrieval-augmented generation. *Nature Communications*.

Kai Guo, Harry Shomer, Shenglai Zeng, Haoyu Han, Yu Wang, and Jiliang Tang. 2025a. Empowering graphrag with knowledge filtering and integration.

In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25450–25464.

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

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

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Haoyu Han, Yu Wang, Harry Shomer, Kai Guo, Jiayuan Ding, Yongjia Lei, Mahantesh Halappanavar, Ryan A Rossi, Subhabrata Mukherjee, Xianfeng Tang, and 1 others. 2024. Retrieval-augmented generation with graphs (graphrag). *arXiv preprint arXiv:2501.00309*.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. Cuad: An expert-annotated nlp dataset for legal contract review. *arXiv preprint arXiv:2103.06268*.

Yuntong Hu, Zhihan Lei, Zheng Zhang, Bo Pan, Chen Ling, and Liang Zhao. 2025. Grag: Graph retrieval-augmented generation. In *Findings of the association for computational linguistics: NAACL 2025*, pages 4145–4157.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 conference of the north American chapter of the association for computational linguistics: Human language technologies*, pages 1419–1436.

Manzong Huang, Chenyang Bu, Yi He, and Xindong Wu. 2025. How to mitigate information loss in knowledge graphs for graphrag: Leveraging triple context restoration and query-driven feedback. *arXiv preprint arXiv:2501.15378*.

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

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

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 785–794.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 6449–6464.

Jiacheng Liang, Yuhui Wang, Changjiang Li, Tanqiu Jiang, Rongyi Zhu, Neil Gong, and Ting Wang. 2026. Graphrag under fire. In *2026 IEEE Symposium on Security and Privacy (SP)*, pages 195–212. IEEE.

Xun Liang, Shichao Song, Simin Niu, Zhiyu Li, Feiyu Xiong, Bo Tang, Yezhaohui Wang, Dawei He, Cheng Peng, Zhonghao Wang, and 1 others. 2024. Uhgeval: Benchmarking the hallucination of chinese large language models via unconstrained generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5266–5293.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 12076–12100.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599.

Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. 2025. Graph retrieval-augmented generation: A survey. *ACM Transactions on Information Systems*, 44(2):1–52.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th annual meeting of the association for computational linguistics: system demonstrations*, pages 101–108.

Kartik Sharma, Peeyush Kumar, and Yunqing Li. 2025. Og-rag: Ontology-grounded retrieval-augmented generation for large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32950–32969.

Naftali Tishby, Fernando C Pereira, and William Bialek. 2000. The information bottleneck method. *arXiv preprint physics/0004057*.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, and 1 others. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):138.

Yaxiong Wu, Jianyuan Bo, Yongyue Zhang, Sheng Liang, and Yong Liu. 2025. Query-centric graph retrieval augmented generation. *arXiv preprint arXiv:2509.21237*.

Chengjin Xu, Muzhi Li, Cehao Yang, Xuhui Jiang, Lumingyuan Tang, Yiyan Qi, and Jian Guo. 2024. Move beyond triples: Contextual knowledge graph representation and reasoning. *arXiv preprint arXiv:2406.11160*. Cs.AI.

Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. 2024. Self-alignment for factuality: Mitigating hallucinations in llms via self-evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1965.

Yibo Zhao, Jiapeng Zhu, Ye Guo, Kangkang He, and Xiang Li. 2025. Eˆ 2graphrag: Streamlining graph-based rag for high efficiency and effectiveness. *arXiv preprint arXiv:2505.24226*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Yingli Zhou, Yaodong Su, Youran Sun, Shu Wang, Taotao Wang, Runyuan He, Yongwei Zhang, Sicong Liang, Xilin Liu, Yuchi Ma, and 1 others. 2025. In-depth analysis of graph-based rag in a unified framework. *arXiv preprint arXiv:2503.04338*.

Xiangrong Zhu, Yuexiang Xie, Yi Liu, Yaliang Li, and Wei Hu. 2025. Knowledge graph-guided retrieval augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8912–8924.

A Statistical Analysis Protocol

All models were evaluated with LLM temperature=0.1 for reproducibility. For key comparisons, we performed five independent runs. We applied paired t-tests with Bonferroni correction, one-way ANOVA with Tukey's HSD for ablations, and report effect sizes in terms of Cohen's d categories. All paired t-tests are performed at the query level, pairing each method's correctness on the same question. Per-dataset Bonferroni-adjusted p-values for significant improvements: BioASQ < 0.001, ConvFinQA 0.002, CUAD < 0.001, DROP 0.001, LongBench 0.002. The remaining datasets (GovReport, Qasper, RACE, TechQA) show positive but non-significant trends after correction (adjusted $p > 0.05$). We refrain from reporting exact Cohen's d values for the main comparisons because the controlled evaluation protocol yields very small standard deviations, making raw effect sizes uninformative and potentially misleading; instead, we report descriptive statistics and significance tests.

B Construction and Analysis of Fine-Grained Subsets

B.1 Corpus Selection

FGQA is constructed from ten classic novels spanning Chinese and Western literary traditions. The five Chinese novels are *Water Margin*, *Dream of the Red Chamber*, *Romance of the Three Kingdoms*, *Journey to the West*, and *Strange Tales from a Chinese Studio*. The five Western novels are *The Old Man and the Sea*, *One Hundred Years of Solitude*, *The Adventures of Sherlock Holmes*, *The Moon and Sixpence*, and *To Kill a Mockingbird*. These works feature long-range narratives, intricate character relationships, and abundant fine-grained details such as numerical values, temporal markers, and entity attributes, making them an ideal testbed for systematic FGFR stress-testing.

The selection balances several factors often confounded in single-corpus evaluations. The Chinese and Western subsets differ in narrative conventions, entity reference patterns, and temporal expression styles, which tests whether attribute extraction remains stable across languages and cultural contexts. Moreover, the novels span multiple centuries and literary genres, reducing the risk that a particular authorial style or historical vocabulary dominates the evaluation. Because fine-grained facts are explicitly stated in the original texts, they can be independently verified without relying on model parametric knowledge. These properties make FGQA suitable for isolating attribute-level retention failures rather than general comprehension ability.

Novel corpora in FGQA are compiled and processed by the authors from Chinese and Western

literary texts. The released FGQA suite contains QA pairs, annotations, and derived retrieval indices; it does not redistribute copyrighted texts.

B.2 FGQA Category Distribution

To ensure comprehensive coverage of fine-grained fact types that are most susceptible to semantic compression in GraphRAG, we design six question categories aligned with the eight attribute types defined in our extraction framework (Appendix D.1). Descriptive and Numerical questions target entity attributes and precise quantities—the two most frequently discarded fact types in standard triple extraction. Action, Location, Material, and Temporal questions probe manner, spatial, physical, and chronological details that are often generalized or omitted during graph construction. The eight attribute types are grouped into six question categories because some attributes co-occur naturally in literary contexts: action questions often require manner and purpose, location questions require place and time, and material/descriptive questions require physical or social attributes. The complete mapping is provided in Appendix D.1. Table B.1 reports the distribution of these six categories across the 10,000 fine-grained questions in FGQA.

Table B.1: Distribution of fine-grained question categories in FGQA (10,000 questions).

Category	Count	%
Descriptive	5,070	50.7%
Numerical	1,830	18.3%
Action	1,040	10.4%
Location	750	7.5%
Material	660	6.6%
Temporal	650	6.5%
Total	10,000	100%

B.3 Construction of Fine-Grained Subsets from Public Benchmarks

For each of the nine public benchmarks, we construct two subsets. The random subset contains 1,000 questions uniformly sampled from the original benchmark. The high-density subset is sampled from the top quartile of attribute density, where attribute density is defined as the number of numerical values, temporal expressions, and entity-specific modifiers appearing in the ground-truth supporting context. For each selected question, we concatenate the question with all paragraphs identified as supporting evidence in the original benchmark annotations, producing long-context retrieval scenarios. Two annotators independently label the query-relevant attribute set A_q for each question following the protocol in Appendix C (Cohen's $\kappa = 0.85$). The resulting subsets serve

B.4 Cross-Benchmark Attribute Density Comparison

Table B.2: Attribute density and cross-sentence dependency comparison across benchmarks.

Benchmark	Focus	Avg. Attr./Q	Cross-sent
HotpotQA	Low	0.2	61%
NQ	Low	0.3	55%
Podcast	Low	0.2	30%
News	Low	0.2	25%
DROP	Medium	1.2	48%
Qasper	Medium	1.1	45%
FGQA	**High**	**1.8**	**31%**

B.5 Summary

Together, these analyses characterize why FGFR has remained largely undetected in prior evaluations. The six-category distribution ensures systematic coverage of fine-grained fact types that are most frequently discarded by standard GraphRAG, while the cross-benchmark comparison demonstrates that existing benchmarks lack sufficient attribute density to expose the resulting failures. FGQA thus fills a critical evaluation gap, providing a testbed where FGFR-related degradation can be systematically diagnosed across diverse fact types and narrative contexts.

C Annotation Protocol for Diagnostic Metrics

To instantiate the query-relevant attribute set A_q for each diagnostic query, two domain experts independently annotate each of the 500 diagnostic queries. We randomly sample 500 diagnostic queries from the test set. All 500 queries are annotated first. For the subsequent R_{end} quintile analysis, the annotated queries are partitioned into five equal-sized quintiles of 100 queries each. For each query, annotators identify all fine-grained attributes (time, quantity, location, manner, material, purpose, descriptor, appearance) whose values are necessary to produce a correct answer. The annotation is conducted at the **attribute-span level**: each attribute key–value pair is treated as an independent unit, and annotators assign a binary label indicating whether the attribute is required for answering q. Inter-annotator agreement is measured at this

attribute-span level (Cohen's $\kappa = 0.88$), and all disagreements are resolved through adjudication by a third expert. To avoid annotation leakage, A_q is annotated exclusively from the source text and before any retrieval or generation output is inspected. The matching procedure is implemented automatically after extraction and uses the same pre-defined alias and normalization dictionaries. The resulting A_q sets are used to compute $R_{\text{rep}}(q)$ and $R_{\text{sel}}(q)$ as defined in Section 3.2.

D Implementation Details

D.1 Attribute Type Definitions and Ontology

We define eight attribute types for fine-grained fact extraction. **Time** captures temporal points or durations (e.g., "yesterday morning"). **Quantity** represents numerical values with units (e.g., "100 taels"). **Location** denotes physical or abstract places (e.g., "Rongguo Mansion"). **Manner** describes the way or style of an action (e.g., "quietly"). **Material** specifies substances or fabrics (e.g., "silver"). **Purpose** indicates goals or reasons (e.g., "to ask for help"). **Descriptor** captures static social or hierarchical attributes (e.g., "eldest daughter"). **Appearance** records visually perceptible features (e.g., "silver-red flower jacket"). The mapping from attribute types to the six FGQA question categories is as follows: Descriptive questions require Descriptor and Appearance; Numerical questions require Quantity; Action questions require Manner and Purpose; Location questions require Location and often Time; Material questions require Material; Temporal questions require Time.

Table D.3 provides representative positive and negative examples for each type.

Table D.3: Positive and negative examples for each attribute type.

Type	Positive	Negative
time	"yesterday morning"	"often", "long ago"
quantity	"100 taels"	"many", "a few"
location	"Rongguo Mansion"	"there", "here"
manner	"quietly"	"then", "thus"
material	"silver"	"thing", "object"
purpose	"to ask for help"	"because", "so"
descriptor	"eldest daughter"	"beautiful", "tall"
appearance	"silver-red jacket"	"wearing", "dressed"

D.2 Attribute Extraction Evaluation

Table D.4 shows per-category attribute extraction performance. Macro-averaged F1 is 81.8%. Hallucination rate drops from 9.2% to 1.4% (95% Wilson CI: [0.3, 4.1]).

Table D.4: Per-category attribute extraction performance (500 manually annotated sentences).

Category	Precision	Recall	F1
Time	84.1%	82.3%	83.2%
Quantity	87.2%	85.8%	86.5%
Location	80.5%	81.5%	81.0%
Manner	78.9%	80.7%	79.8%
Material	89.3%	86.9%	88.1%
Purpose	71.8%	73.0%	72.4%
Descriptor	86.6%	84.8%	85.7%
Appearance	78.5%	77.3%	77.9%
Macro Avg.	82.1%	81.5%	**81.8%**

D.3 Dependency Parser and Evidence Window Detection

For Chinese text processing, we use Stanza v1.6.0 (Qi et al., 2020), a multilingual neural pipeline trained on Universal Dependencies. The evidence window is determined by finding the lowest common ancestor (LCA) of the subject and object in the dependency tree. Stanza achieves UAS/LAS of 72.9%/69.8% on the Chinese-GSD treebank, providing a reliable foundation for window positioning. For English text, we replace the parser with `spaCy en_core_web_lg`. When the LCA subtree cannot be constructed (e.g., due to poetic or classical language), we fall back to a fixed window of ± 10 tokens around the subject and object; if the subject or object cannot be located, we use the full sentence. On 500 manually annotated sentences, the primary parser succeeds in 94.2% of cases; in the remaining 5.8%, the fallback mechanism maintains 81.5% attribute recall.

D.4 Cross-Lingual Adaptation

The eight attribute categories remain identical across languages; the keyword list and synonym dictionary are translated versions with approximately 120 English entries each. All extraction prompts are language-agnostic by design, requiring no structural modification. We observe comparable attribute extraction quality across Chinese and English texts, with macro-averaged F1 scores within 2 percentage points of each other.

D.5 Fallback Fact Retrieval (FFR)

In rare cases where the information-preserving filter yields no edges within the k-hop candidate subgraph, a fallback is triggered. The query is dispatched to the dense retriever

(`text-embedding-3-large`), and the top-10 retrieved chunks are serialized into the context. This fallback operates outside the graph and functions as a safeguard. In our main experiments, the fallback is triggered in fewer than 3% of queries and contributes +1.9 points to accuracy, confirming that the primary gains originate from the two components.

D.6 Algorithm Details

D.6.1 Information-Preserving Filtering Algorithm

Algorithm 2 details the complete information-preserving filtering procedure. For a query q and seed entities E_s, the algorithm first retrieves a candidate edge set E_{cand} via k-hop expansion. It then computes three dimension scores (S_{attr}, S_{ling}, S_{net}) for each edge and derives adaptive weights w_{dim} from the coefficient of variation (σ_{dim}/μ_{dim}) of each score distribution. The final selection combines an adaptive threshold $\theta = \mu_{\mathcal{S}} + 1.5 \cdot \sigma_{\mathcal{S}}$ with explicit protection for structural bridges and semantic outliers.

Algorithm 2 Information-Preserving Relationship Filtering

Require: Knowledge graph G, query q, seed entities E_s, hop limit $k = 2$, retention ratio $\rho = 0.7$
Ensure: Filtered edge set E'

function INFORMATIONPRESERVINGFILTER(G, q, E_s, k, ρ)
  $E_{\text{cand}} \leftarrow$ GetKCandidateRelationships(G, E_s, k)
  for each $r \in E_{\text{cand}}$ **do**
    Compute $S_{\text{attr}}(r)$, $S_{\text{ling}}(r)$, $S_{\text{net}}(r)$
  end for
  for $dim \in \{\text{attr}, \text{ling}, \text{net}\}$ **do**
    $\mu_{dim} \leftarrow \text{mean}(\{S_{dim}(r) \mid r \in E_{\text{cand}}\})$
    $\sigma_{dim} \leftarrow \text{std}(\{S_{dim}(r) \mid r \in E_{\text{cand}}\})$
    $CV_{dim} \leftarrow \sigma_{dim}/(\mu_{dim} + \epsilon)$ ▷ $\epsilon = 10^{-6}$
  end for
  if $\sum_d CV_d < \epsilon$ **then**
    $(w_{\text{attr}}, w_{\text{ling}}, w_{\text{net}}) \leftarrow (0.5, 0.3, 0.2)$
  else
    $w_{dim} \leftarrow CV_{dim} / \sum_d CV_d$ for each dim
  end if
  for each $r \in E_{\text{cand}}$ **do**
    $\mathcal{S}[r] \leftarrow \sum_{dim} w_{dim} \cdot S_{dim}(r)$
  end for
  $\mu_{\mathcal{S}} \leftarrow \text{mean}(\mathcal{S})$
  $\sigma_{\mathcal{S}} \leftarrow \text{std}(\mathcal{S})$
  $\theta \leftarrow \mu_{\mathcal{S}} + 1.5 \cdot \sigma_{\mathcal{S}}$
  $E_{\text{keep}} \leftarrow \{r \mid \mathcal{S}[r] > \theta\}$
  $E_{\text{struct}} \leftarrow$ structural bridges and outliers
  $target \leftarrow \lfloor \rho \cdot |E_{\text{cand}}| \rfloor$
  $remaining \leftarrow target - |E_{\text{keep}} \cup E_{\text{struct}}|$
  if $remaining > 0$ **then**
    return $E_{\text{keep}} \cup E_{\text{struct}} \cup \text{TopK}(\mathcal{S}, remaining)$
  else
    return $E_{\text{keep}} \cup E_{\text{struct}}$
  end if
end function

D.7 Formal Definitions of Scoring Functions

- **Evidence completeness** $S_{\text{attr}}(r) = \lambda_{a1} \cdot |A_r| + \lambda_{a2} \cdot \text{endpoint-importance}(r)$, where A_r is the attribute set of relation r and endpoint importance measures the combined centrality of its subject and object.
- **Query compatibility** $S_{\text{ling}}(r) = \lambda_{l1} \cdot \cos(\mathbf{v}_{\text{pattern}}(r), \mathbf{v}_{\text{indicator}}) + \lambda_{l2} \cdot \text{density}(r) + \lambda_{l3} \cdot \text{attribute-keyword-overlap}(r)$, where $\mathbf{v}_{\text{pattern}}$ embeds the relation verbalization, $\mathbf{v}_{\text{indicator}}$ averages embeddings of fine-grained indicator terms from the query, density is the ratio of entities and attributes to token length, and attribute keyword overlap measures the fraction of query attribute keywords that appear in the relation's attribute set A_r or its verbalization.
- **Structural continuity** $S_{\text{net}}(r) = \lambda_{n1} \cdot \text{conn}(r) + \lambda_{n2} \cdot \text{bridge}(r)$, where $\text{conn}(r)$ is local connectivity and $\text{bridge}(r)$ is normalized edge betweenness centrality.

D.8 Bayesian Optimization Configuration

We employed Gaussian Process Bayesian optimization (Ray Tune v2.0, Matern 5/2 kernel, Expected Improvement) with 5 random initial points and 50 iterations on a 200-query development set. The optimized coefficients are $\lambda_{a1} = 0.55$, $\lambda_{a2} = 0.45$, $\lambda_{l1} = 0.35$, $\lambda_{l2} = 0.40$, $\lambda_{l3} = 0.25$, $\lambda_{n1} = 0.65$, $\lambda_{n2} = 0.35$. The fallback weights are only used when the separability-based normalization is degenerate, and are chosen by development-set optimization. For reproducibility, we report the fallback weights obtained on the development set ($w_{\text{attr}} = 0.5$, $w_{\text{ling}} = 0.3$, $w_{\text{net}} = 0.2$). These weights are used only when the coefficient-of-variation normalization is degenerate (i.e., $\sum_d CV_d < \epsilon$); in normal online operation, the weights are recomputed per query as described in Section 4.3. The optimized threshold parameters are $k = 2$ and $\rho = 0.7$.

E Extended Discussion of Diagnostic Metrics

As noted in Section 3.2, R_{sel} is a conditional diagnostic quantity that isolates selection-stage loss given successful representation. Here we provide additional details on this metric and its relationship to absolute retention.

Why R_{sel} is not a standalone performance metric. R_{sel} normalizes by the number of represented facts F_q^G, so a method that represents very few facts could still achieve a high R_{sel} if it retrieves

those few facts perfectly. To avoid misleading interpretations, we also define an **absolute selection retention** metric:

$$R_{\text{sel}}^{\text{abs}}(q) = \frac{|F_q^G \cap C_q^*|}{|F_q|}. \tag{10}$$

Unlike R_{sel}, $R_{\text{sel}}^{\text{abs}}$ measures the fraction of all required facts that are both represented and successfully retrieved. This metric reflects the end-to-end impact of selection, penalizing methods that represent few facts even if they retrieve them perfectly.

Mathematical equivalence with R_{end}. It is important to note that R_{end} is mathematically equivalent to $R_{\text{sel}}^{\text{abs}}$, since

$$\begin{aligned} R_{\text{end}} &= R_{\text{rep}} \times R_{\text{sel}} \\ &= \frac{|F_q^G|}{|F_q|} \times \frac{|F_q^G \cap C_q^*|}{|F_q^G|} \\ &= \frac{|F_q^G \cap C_q^*|}{|F_q|} = R_{\text{sel}}^{\text{abs}}. \end{aligned} \tag{11}$$

This equivalence reinforces that R_{end} captures absolute retention of essential facts rather than a conditional measure, making it the primary metric for end-to-end evaluation.

Probabilistic interpretation. R_{rep} estimates the probability that a required fact is fully represented in G, and R_{sel} estimates the conditional survival probability of that fact being retrieved given successful representation, i.e., $P(\text{retrieved} \mid \text{represented})$. Therefore, R_{end} corresponds to the joint probability $P(\text{represented}) \times P(\text{retrieved} \mid \text{represented})$, capturing the end-to-end retention of fine-grained evidence throughout the pipeline.

F Experimental Setup Details

This appendix provides detailed information on data split, hyperparameter tuning, and random seeds used in the main experiments.

Data Split. FGQA contains 10,000 fine-grained QA pairs and 1,000 global reasoning queries. We randomly partition 9,700 fine-grained questions as the test set, and set aside 300 questions as a development set for hyperparameter tuning. Among the 300 development questions, 200 are used for Bayesian optimization and 100 are held out as an early-stopping/dev-evaluation set. The 500 diagnostic queries used for stage-wise analysis are randomly sampled from the 9,700 test questions and are disjoint from the development set. No diagnostic query is used for hyperparameter tuning.

Random Seeds. Results are averaged over five independent runs with seeds 42, 2024, 1024, 2048, and 3072. The LLM generation temperature is set to 0.1 for all methods unless otherwise specified.

Bayesian Optimization. Hyperparameters for the information-preserving filter and the priority coefficient α were tuned via Bayesian optimization on the development set (see Appendix D.8 for complete configuration). The optimal values are those reported in the main text.

G Global Reasoning Evaluation

G.1 Evaluation Protocol

Two domain experts independently judged all pairwise method comparisons on the 1,000 global reasoning queries. The evaluation dimensions were defined as follows:

- **Coverage:** Does the answer identify and discuss all key events, characters, and locations relevant to the query, without omitting any essential element?
- **Diversity:** Does the answer draw supporting evidence from multiple distinct aspects (e.g., social hierarchy, ritual formality, emotional tone, symbolic meaning) rather than reiterating a single perspective?
- **Insightfulness:** Does the answer move beyond description to offer non-obvious interpretations, causal connections, or thematic observations?
- **Consistency:** Is the answer internally coherent, free of logical contradictions, and consistent with established facts from the source text?
- **Inspirativeness:** Would the answer stimulate further inquiry—for example, by raising implicit questions, suggesting comparative angles, or revealing underexplored narrative patterns?

Judges received fully anonymized response pairs and were forced to select a winner for each dimension (no ties permitted). Win-rates were computed as the fraction of comparisons won by a method across all 1,000 queries. Inter-annotator agreement reached Cohen's $\kappa = 0.84$.

G.2 Case Study with Dimension-Wise Analysis

The full evaluation protocol with dimension definitions, a case analysis (Table G.5), and complete pairwise matrices are in Appendix G. In the case analysis, green highlights indicate correctly retrieved or well-supported aspects, while red highlights mark omissions or inaccuracies.

Table G.5: Case Study: Funeral events in *Dream of the Red Chamber*

Query	Analyze the differences in descriptions of three important funerals (Qin Keqing, Jia Jing, Grandmother Jia) in the novel.
F-GraphRAG	**Comprehensive Analysis:** Provides complete coverage of all three funerals with detailed comparative analysis. Qin Keqing's grand 49-day ceremony with imperial-grade materials and aristocratic participation symbolizes the family's peak prosperity. Jia Jing's modest Taoist ceremony reflects social decline and spiritual retreat. Grandmother Jia's funeral, marred by financial strain and servant infighting, emphasizes the collapse of internal family cohesion. The analysis demonstrates systematic comparison across ritual scale, social hierarchy, and symbolic meanings.
GraphRAG	**Limited Analysis:** Provides detailed description of Qin Keqing's funeral but explicitly states inability to compare other funerals due to perceived information gaps. The response acknowledges having only "partial information" about Jia Jing and Grandmother Jia, and incorrectly claims the novel lacks sufficient details—when the source text devotes multiple chapters to Grandmother Jia's funeral, resulting in an incomplete and misleading analytical framework.
Human Expert Evaluation	**Winner: F-GraphRAG** • **Comprehensive**: Complete coverage of all three funerals with specific ritual elements that GraphRAG omits • **Diverse**: Multi-dimensional analysis spanning ritual, social, and symbolic dimensions • **Insightful**: Reveals the narrative arc from prosperity to decline across the three events • **Consistent**: Maintains balanced analytical depth grounded in verifiable textual details • **Inspiring**: Opens avenues for further literary analysis and interpretation

G.3 Complete Pairwise Comparison Matrices

Tables G.6–G.10 present the full win-rate matrices for all seven methods across the five evaluation dimensions. Methods are abbreviated as: **FG** (F-GraphRAG), **GR** (GraphRAG), **LR** (LightRAG), **E²G** (E²GraphRAG), **FastG** (FastGraphRAG), **D4** (DashScope-text-embedding-v4), **T3L** (text-embedding-3-large). All pairs satisfy the no-tie constraint, and symmetric entries sum to 100%.

Table G.6: Coverage Comparison for Global Reasoning

A \ B	FG	GR	LR	E²G	FastG	D4	T3L
FG	50.00	55.56	53.42	54.58	54.37	66.96	64.71
GR	44.44	50.00	49.39	48.73	47.98	60.47	59.87
LR	46.58	50.61	50.00	50.43	51.79	61.57	59.31
E²G	45.42	51.27	49.57	50.00	53.74	62.89	60.17
FastG	45.63	52.02	48.21	46.26	50.00	61.04	59.74
D4	33.04	39.53	38.43	37.11	38.96	50.00	50.41
T3L	35.29	40.13	40.69	39.83	40.26	49.59	50.00

Table G.7: Diversity Comparison for Global Reasoning

A \ B	FG	GR	LR	E²G	FastG	D4	T3L
FG	50.00	55.89	56.12	55.17	54.83	67.32	64.82
GR	44.11	50.00	48.43	47.47	50.27	59.35	58.18
LR	43.88	51.57	50.00	48.57	51.87	62.67	60.76
E²G	44.83	52.53	51.43	50.00	55.73	64.53	64.78
FastG	45.17	49.73	48.13	44.27	50.00	58.23	56.97
D4	32.68	40.65	37.33	35.47	41.77	50.00	47.39
T3L	35.18	41.82	39.24	35.22	43.03	52.61	50.00

Table G.8: Insightfulness Comparison for Global Reasoning

A \ B	FG	GR	LR	E²G	FastG	D4	T3L
FG	50.00	55.31	57.18	58.73	59.79	69.67	67.83
GR	44.69	50.00	46.57	44.78	42.89	59.34	54.22
LR	42.82	53.43	50.00	50.78	54.68	64.31	64.27
E²G	41.27	55.22	49.22	50.00	55.90	66.30	64.32
FastG	40.21	57.11	45.32	44.10	50.00	63.56	59.77
D4	30.33	40.66	35.69	33.70	36.44	50.00	53.28
T3L	32.17	45.78	35.73	35.68	40.23	46.72	50.00

Table G.9: Consistency Comparison for Global Reasoning

A \ B	FG	GR	LR	E²G	FastG	D4	T3L
FG	50.00	53.57	55.43	52.41	56.79	51.76	50.86
GR	46.43	50.00	50.57	46.52	48.05	48.67	48.39
LR	44.57	49.43	50.00	52.40	55.84	50.73	49.11
E²G	47.59	53.48	47.60	50.00	53.64	50.32	48.37
FastG	43.21	51.95	44.16	46.36	50.00	49.99	47.42
D4	48.24	51.33	49.27	49.68	50.01	50.00	49.53
T3L	49.14	51.61	50.89	51.63	52.58	50.47	50.00

Table G.10: Inspirativeness Comparison for Global Reasoning

A \ B	FG	GR	LR	E²G	FastG	D4	T3L
FG	50.00	57.76	56.92	57.53	58.43	66.98	63.76
GR	42.24	50.00	44.24	46.72	48.82	64.48	62.92
LR	43.08	55.76	50.00	48.31	56.08	66.47	61.34
E²G	42.47	53.28	51.69	50.00	57.31	65.57	63.36
FastG	41.57	51.18	43.92	42.69	50.00	60.08	57.90
D4	33.02	35.52	33.53	34.43	39.92	50.00	52.07
T3L	36.24	37.08	38.66	36.64	42.10	47.93	50.00

H Failure Cases of F-GraphRAG

While F-GraphRAG substantially reduces FGFR, it is not immune to all fine-grained failures. Below we present three representative failure modes identified during manual error analysis on the FGQA development set. These cases illustrate the remaining challenges and motivate future work.

- **Cross-paragraph attribute propagation.** When a fact's subject, object, and attribute span across multiple paragraphs, the sentence-level evidence window may fail to capture the complete relation. For example, the attribute "49 days" may appear in a paragraph far from the funeral event it modifies, leading to incomplete extraction.
- **Coreference resolution ambiguity.** In narratives with many pronouns and aliases, the entity linking step may map a pronoun to the wrong antecedent, causing attribute assignment to an incorrect entity. This is especially prevalent in Chinese classical texts where a single character may be referred to by multiple titles.
- **Ambiguous temporal relations.** Expressions like "a few days later" are relative and require discourse context to resolve. When the context is not captured in the evidence window, temporal attributes may be extracted as underspecified or incorrectly anchored.

These failure modes account for approximately 15% of the remaining errors in F-GraphRAG. The error categories were obtained by manually inspecting incorrect responses on the FGQA development set and grouping them according to the primary failure cause. They mainly arise from limitations of local evidence windows rather than the retrieval ranking mechanism itself, indicating directions for further improvement in cross-sentence extraction and coreference-aware context modeling.

I Answer Stability Visualization

Detailed analysis of answer consistency under stochastic generation is provided in Section 5.6. For completeness, Figure I.1 visualizes the consistency across ten independent runs for GraphRAG and F-GraphRAG. The visualization complements the quantitative variance analysis by showing run-level consistency rather than only aggregate standard deviations.

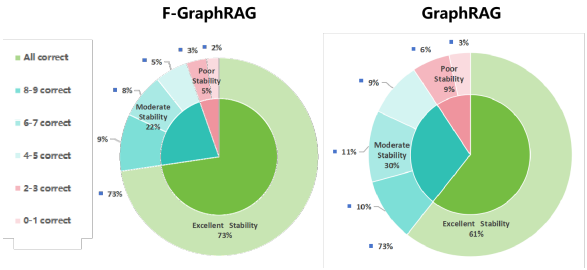


Figure I.1: Answer consistency across ten independent generation runs.

J Full Model Comparison

Table J.11 provides the complete comparison between F-GraphRAG and all evaluated retrieval baselines. The main paper reports a representative subset for readability, while this appendix includes the full set of graph-based and dense retrieval approaches.

The compared methods cover two major retrieval paradigms. The graph retrieval group includes recent GraphRAG variants that construct structured representations for multi-hop reasoning, while the dense retrieval group contains representative embedding-based retrievers covering both general-purpose and Chinese-oriented models. All methods are evaluated under the same generation backbone, prompt configuration, and answer evaluation protocol to ensure a fair comparison.

The complete results further support the effectiveness of F-GraphRAG for fine-grained evidence retrieval. Compared with existing graph-based approaches, F-GraphRAG achieves consistent improvements by preserving attribute-level information during graph construction and reducing retrieval-stage evidence loss. Meanwhile, its performance remains competitive against strong dense retrievers, demonstrating that fine-grained evidence preservation can complement rather than replace conventional semantic retrieval.

The full comparison also reveals several observations. First, graph-based retrieval methods generally outperform lightweight dense retrievers in fine-grained question answering, suggesting that explicit relational structures remain beneficial when queries require attribute-level reasoning. However, conventional graph construction approaches still suffer from information compression during entity-relation abstraction, limiting their ability to preserve detailed evidence. F-GraphRAG improves upon these approaches by introducing evidence-preserving representation and selective

retrieval mechanisms, indicating that graph-based retrieval can achieve stronger fine-grained performance without sacrificing structural reasoning capability.

Table J.11: Fine-Grained QA Accuracy Comparison (Complete Results)

Category	Model	Accuracy
Graph Retrieval	**F-GraphRAG (Ours)**	**68.7 (0.8)**
	F-GraphRAG w/o fallback	66.8 (0.9)
	GraphRAG-FI	64.4 (1.0)
	QCG-RAG	64.3 (1.0)
	E^2GraphRAG	62.1 (1.1)
	GraphRAG	56.1 (1.2)
	FastGraphRAG	59.8 (1.3)
	LightRAG	56.7 (1.4)
	CheapRAG	60.5 (1.1)
Dense Retrieval	text-embedding-3-large	65.9 (0.8)
	DashScope-text-embedding-v4	63.6 (0.7)
	stella-large-v2	61.7 (0.9)
	DashScope-text-embedding-asyn-v1	62.3 (0.8)
	bge-large	60.3 (0.8)
	DashScope-text-embedding-asyn-v2	62.0 (0.8)
	gte-sentence-embedding-chinese-large	58.6 (1.2)
	bge-base-zh	53.2 (1.5)
	gte-sentence-embedding-chinese-base	53.2 (1.5)
	text2vec-large-chinese	52.8 (1.6)
	bloom-1b1-embedding	48.4 (1.8)
	Jina-embedding-v4	58.8 (1.2)
	m3e-base	45.8 (1.9)
	Yuan-embedding-1.0	41.2 (2.0)
	BERT-embedding-chinese-base	41.2 (2.0)
	text2vec-base-chinese	38.2 (2.1)

K Failure Analysis of Baseline GraphRAG

To understand why conventional GraphRAG suffers from FGFR, we manually analyze omitted information categories in baseline responses. Numerical and material attributes account for over half of all omissions.

Table K.12: Fine-Grained Information Loss Analysis in Baseline GraphRAG

Type	Count	Specific Information Lost
Numerical	33%	Amounts (e.g. , "100 taels")
Descriptive	15%	Character (e.g. , "clan leader")
Action	12%	Manners (e.g. , "quietly handed")
Location	11%	Places (e.g. , "Rongguo Mansion")
Material	21%	Substances (e.g. , "silk clothing")
Temporal	8%	Time (e.g. , "midnight")

The analysis indicates that conventional graph construction tends to preserve coarse semantic relations while losing low-frequency attributes that are less important for global graph connectivity. These omitted details are often critical for fine-grained question answering, where the answer depends on exact quantities, materials, temporal markers, or descriptive modifiers. This observation motivates the design of F-GraphRAG, which explicitly treats attribute preservation as a first-class objective during graph construction.

L Random Subset Results for Cross-Domain Validation

As a complement to the high-density stress test, we evaluate on randomly sampled 1,000-question subsets from each public benchmark. Results are shown in Table L.13.

Table L.13: Accuracy on random subsets (original distributions).

Dataset	GraphRAG	F-GraphRAG	Δ
BioASQ	54.2	55.5	+1.3
ConvFinQA	50.1	54.6	+4.5
CUAD	57.8	58.6	+0.8
DROP	63.5	65.8	+2.3
GovReport	59.2	61.3	+2.1
LongBench	55.4	59.7	+4.3
Qasper	61.3	63.6	+2.3
RACE	65.2	68.1	+2.9
TechQA	58.4	59.9	+1.5

Note: Δ = F-GraphRAG − GraphRAG. All differences are positive on the sampled subsets, showing consistent improvements across evaluated domains.

M Cross-Model Validation Standard Deviations

This appendix provides the standard deviations for the cross-model validation results reported in Table 7. All values are averaged over five independent runs using the same seeds as the main experiments.

Table M.14: Cross-model validation with standard deviations.

Generation Model	GraphRAG	F-GraphRAG	Δ
DeepSeek-V3	58.1 (0.6)	69.0 (0.4)	+10.9
GPT-4o	58.3 (0.7)	70.1 (0.5)	+11.8
Qwen2-72B	54.7 (0.8)	64.2 (0.6)	+9.5