F-GraphRAG: Breaking the Fine-Grained Fact Retrieval Cascade via Granularity-Aware Extraction and Semantic-Aware Filtering

Anonymous ACL submission

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 introduces a distinct failure mode: semantic compression during triple extraction systematically discards fine-grained facts such as precise quantities, temporal constraints, and entity attributes. We formalize this as the Fine-Grained Fact Retrieval (FGFR) cascade and introduce stage-wise diagnostic metrics—R_{rep}, R_{sel}, and R_{end}—to quantify information loss. To stress-test FGFR, we construct a fine-grained QA benchmark from classic novels, whose long-range narratives, intricate character webs, and dense factual details make them an ideal testbed for exposing semantic compression failures that conventional benchmarks overlook. F-GraphRAG, our framework, features a granularity-aware extractor that preserves atomic details and a semantic-aware filter that counters global importance bias. On this benchmark, F-GraphRAG outperforms GraphRAG by 12.6% in accuracy while using 33.8% fewer relationships, with consistent gains of 5.3–12.6% across nine diverse public benchmarks. These results show the detail-structure trade-off in graph-based RAG is not inevitable; prioritizing informational quality over relational quantity enables more faithful, efficient, and stable RAG.

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., 2021b; Zhang et al., 2024; Liang et al., 2024). 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. Graph-based RAG (GraphRAG) addresses this gap (Edge et al., 2024) by constructing unified knowledge graphs for superior thematic discovery (Pan et al., 2024; Besta et al., 2024). However, this structural shift introduces a critical trade-off: the systematic loss of fine-grained facts (e.g. , numerical values, temporal constraints, attributes).

We formalize this limitation as the **Fine-Grained Fact Retrieval (FGFR)** problem. Unlike conventional recall failures where relevant information is entirely absent, FGFR constitutes a *structural representation failure*: the target facts exist in the source corpus but are discarded or generalized during knowledge graph construction due to semantic compression and statistical bias (Ji et al., 2021a; Xu et al., 2024; Albertoni et al., 2006; Liang et al., 2025; Peng et al., 2024). These two bottlenecks synergistically create a **FGFR cascade** in which generic triples dominate query contexts, systematically eroding fine-grained evidence.

This cascade has gone unnoticed because mainstream QA benchmarks such as HotpotQA and Natural Questions are built on corpora with relatively straightforward narrative structures and rarely pose questions that require precise recall of fine-grained attributes as a primary capability. Our own attempts to repurpose several public datasets for FGFR diagnosis were unsuccessful: their evaluation protocols cannot pinpoint whether information loss occurs at the extraction or retrieval stage, and thus cannot compute our three diagnostic metrics. To bridge this gap, we construct a dedicated fine-grained QA benchmark (FGQA) from ten classic novels, whose long-range narratives, intricate character webs, and dense factual details naturally embed abundant fine-grained facts, providing the first testbed capable of systematic FGFR dissection. On conventional benchmarks, GraphRAG achieves high accuracy (86.5–91.3% across Natural Questions (NQ), and HotpotQA), yet on our FGQA benchmark its performance plummets to 56.1%—a drop of over 30 percentage points. Even with

our proposed F-GraphRAG, accuracy falls from 87.2–92.1% on standard benchmarks to 68.7% on FGQA, still a decline of nearly 20 points. These stark contrasts confirm that FGQA uniquely exposes the FGFR cascade that existing evaluations systematically overlook.[1]

To break this cascade, we propose **F-GraphRAG**, an integrated framework that directly targets both bottlenecks: a granularity-aware extractor to preserve atomic details, and a semantic-aware filter to counteract statistical bias. On our constructed FGQA benchmark, F-GraphRAG outperforms standard GraphRAG by 12.6% in accuracy while using 33.8% fewer relationships. Our contributions are threefold:

1. **Formalizing the FGFR cascade.** We formalize the Fine-Grained Fact Retrieval (FGFR) cascade and introduce stage-wise diagnostic metrics (R_{rep}, R_{sel}, R_{end}) to quantify and diagnose fine-grained fact degradation for the first time.
2. **The FGQA benchmark.** We construct FGQA, a fine-grained QA benchmark spanning ten classic novels with 10,000 questions annotated with query-relevant attribute sets, plus 1,000 global reasoning questions. FGQA exposes FGFR defects and, together with fine-grained subsets we curate from nine public benchmarks to match the FGFR task requirements, validates cross-domain generality.
3. **The F-GraphRAG framework.** We propose F-GraphRAG, which breaks the FGFR cascade through a granularity-aware extractor and semantic-aware filter, achieving 12.6% over GraphRAG using 33.8% fewer relationships.

2 Related Work

Standard GraphRAG (Edge et al., 2024) suppresses fine-grained details during graph construction. Context-enriched (Huang et al., 2025; Xu et al., 2024) and hypergraph (Feng et al., 2025) methods still operate on compressed triples. Retrieval works optimize either efficiency (e.g. , LightRAG (Guo et al., 2024), FastGraphRAG (AI, 2025), CheapRAG (Zhou et al., 2025)) or reasoning paths (e.g. , PathRAG (Chen et al., 2025), QCG-RAG (Wu et al., 2025), E^{2}GraphRAG (Zhao et al., 2025)); GraphRAG-FI (Guo et al., 2025) adds two-stage filtering. None reward fine-grained informativeness explicitly. Surveys (Pan et al., 2024; Besta et al., 2024) and benchmarks (Liang et al., 2025) lack per-stage diagnostics. We address these gaps with (1) a granularity-aware extractor that preserves attributes during construction, (2) a semantic-aware filter with multi-dimensional, query-adaptive scoring, and (3) stage-wise metrics (R_{rep}, R_{sel}, R_{end}) plus the FGQA benchmark for principled root-cause analysis.

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 discards A, introducing a systematic 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)**.** A GraphRAG system suffers from Fine-Grained Fact Retrieval (FGFR) if, 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] Code and data are available at https://anonymous.4open.science/status/F-GraphRAG-446F.

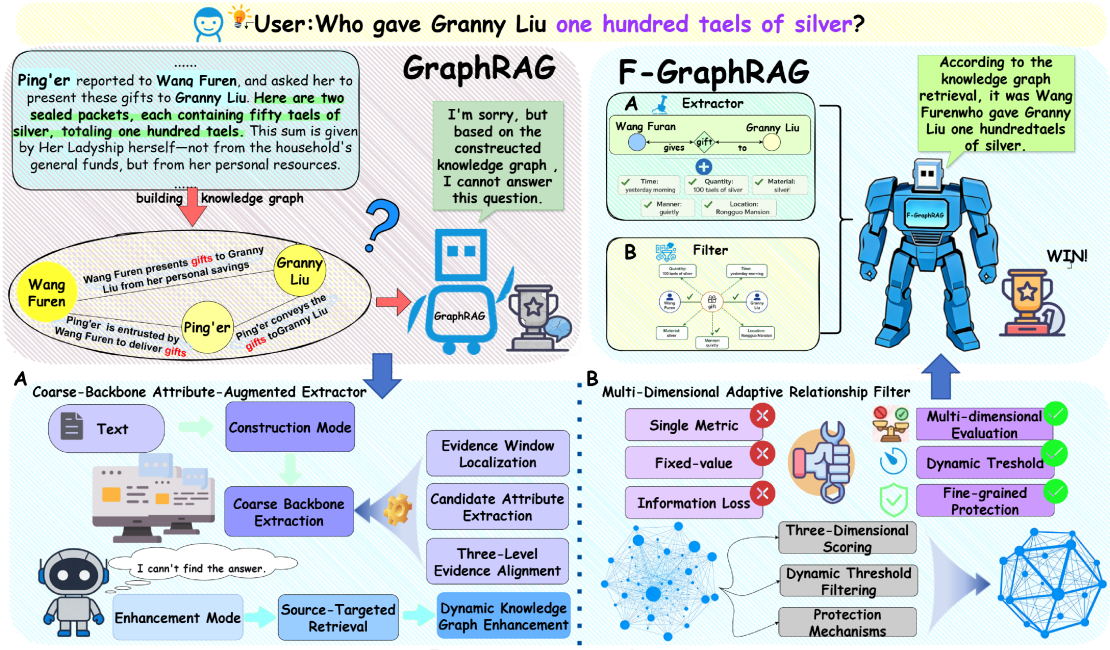

Figure 1: Overcoming FGFR: F-GraphRAG's adaptive components vs. GraphRAG's failure cases.

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.

These two failures are interdependent, forming a reinforcing cascade that systematically erodes fine-grained evidence. 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:

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

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

3.2 Motivation

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:

$$R_{\text{rep}}(q) = \frac{|F_q^G|}{|F_q|}. \quad (3)$$

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

$$R_{\text{sel}}(q) = \begin{cases} \dfrac{|F_q^G \cap C_q^*|}{|F_q^G|}, & F_q^G \neq \emptyset, \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

Here C_q^* denotes the set of facts whose corresponding relation and all query-relevant attributes are present in C_q. When $F_q^G = \emptyset$, a representation failure forces end-to-end retention to zero.

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

$$R_{\text{end}}(q) = R_{\text{rep}}(q) \times R_{\text{sel}}(q). \quad (5)$$

The multiplicative form naturally penalizes any single-stage failure, reflecting the physical reality of cascade collapse. 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 $A_r = \{(\tau_1, \nu_1), \dots\}$.

To break the FGFR cascade identified, we propose F-GraphRAG with two core coordinated components (Figure 1). Specifically, the **granularity-aware extractor** targets the extraction-stage semantic compression by preserving fine-grained attributes, thereby improving R_{rep}. The **semantic-aware filter** counters the filtering-stage statistical bias by prioritizing informational richness, thereby improving R_{sel}. When the main pipeline yields insufficient context, a lightweight fallback dense retrieval step is triggered as a robustness safeguard; this is not a core architectural component and its contribution is examined in the ablation study.

4.2 Granularity-Aware Relationship Extraction

To directly mitigate the semantic compression bottleneck, we propose a two-stage granularity-aware extractor, illustrated in Figure 2. First, a coarse backbone of triples is extracted to ensure structural coverage. Second, an evidence-grounded attribute augmentation step enriches each triple with fine-grained attributes from a local evidence window. The window is determined through a pluggable language interface (e.g., using the lowest common ancestor in a dependency parse, or an entity-based span), ensuring language-agnostic applicability. Inside this window, we extract candidate attributes of eight types (time, quantity, location, manner, material, purpose, descriptor, appearance) and validate them via a three-level alignment (exact, normalized, semantic similarity). Full details of the extraction pipeline and evaluation (F1, hallucination rate) are provided in Appendix D.2.

To balance informativeness and graph size, each relation is assigned a priority score $\text{priority}(r) = 1 + \alpha \cdot |A|$ ($\alpha = 0.4$, chosen via sensitivity analysis). This lightweight design avoids introducing additional computational overhead at extraction; complex ranking decisions are deferred to the semantic-aware filter. The top-K relations per 1,000 tokens are retained. Algorithm 1 formalizes the process.

4.3 Semantic-Aware Relationship Filtering

To address the statistical bias bottleneck in filtering, we design a multi-dimensional semantic-aware scoring framework (illustrated in Figure 3). We obtain seed entities by prompting the LLM (DeepSeek-V3, temperature 0.1) to extract entity mentions from the query and link them to graph

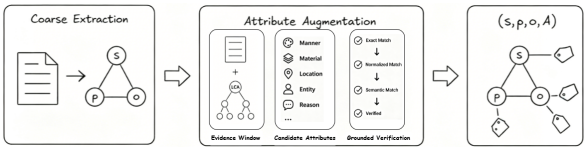

Figure 2: Coarse backbone extraction followed by evidence-grounded attribute augmentation.

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)
    R_c ← E_c(t)
    R ← ∅
    for (s, p, o) ∈ R_c do
        window ← getEvidenceWindow(t, s, o)
        candidates ← E_a(window)
        A ← ∅
        for (τ, ν) ∈ candidates do
            if aligned(ν, window) then
                A ← A ∪ {(τ, ν)}
            end if
        end for
        R ← R ∪ {(s, p, o, A)}
    end for
    priority(r) ← 1 + 0.4 · |A| for each r ∈ R
    Sort R by priority(r) descending
    Keep top K = 50 relations per 1k tokens
    return R
end function
```

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 semantic alignment scoring in the subsequent filtering stage. Then, three complementary scores are computed for each candidate relation: S_{attr} (attribute richness and endpoint importance), S_{ling} (semantic alignment, information density, syntactic complexity), and S_{net} (local connectivity and bridge centrality). These scores are defined formally in Appendix D.5.

We adopt an adaptive weighting strategy based on the coefficient of variation (CV) of each score distribution, ensuring that dimensions with higher discriminative power influence the final ranking more. The comprehensive score $S_{\text{total}}(r) = \sum_i w_i \cdot S_i(r)$ 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.4.1).

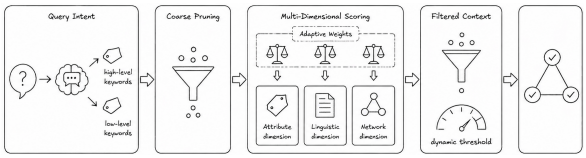

Figure 3: Semantic-aware filtering pipeline with three-dimensional scoring and adaptive thresholding.

4.4 Integrated Pipeline

The complete F-GraphRAG pipeline integrates the granularity-aware extractor and the semantic-aware filter as its two core components (Figure 1). During offline construction, documents are processed through the granularity-aware extractor to build an attribute-augmented knowledge graph. During online querying, the semantic-aware filter strategically selects the most informative relationships from this graph under the given token budget.

In the rare case that the main pipeline yields no edges within the k-hop candidate subgraph (fewer than 3% of queries in our experiments), a lightweight fallback is triggered: the query is dispatched to the dense retriever used in our comparative baselines (`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 rather than a core architectural contribution. Removing the fallback (retaining the extractor and filter) reduces accuracy from 68.7% to 66.8%, confirming that the primary gains originate from the two core components.

5 Experiments

5.1 Experimental Settings

Benchmarks. Benchmarks. We evaluate on our Fine-Grained QA Dataset (FGQA), which 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 (category distribution) and Appendix C (annotation), with key statistics in Table 1. As a baseline, we also include Podcast and News, two datasets constructed following the evaluation protocol of GraphRAG. As shown in Table 1, GraphRAG achieves high scores on standard benchmarks (86.5–91.3%) but drops sharply to 56.1% on FGQA, a decline of over 30 percentage points that directly exposes the FGFR cascade. F-GraphRAG narrows this performance gap, falling from 87.2–92.1% on standard benchmarks to 68.7% on FGQA, confirming that its design effectively targets the root causes of FGFR.

Table 1: Performance drops sharply on FGQA compared to standard QA benchmarks.

Dataset	GraphRAG	F-GraphRAG	ΔGR	ΔFGR
HotpotQA	91.3%	92.1%	35.2%↑	23.4%
NQ	89.8%	90.5%	33.7%↑	21.8%
Podcast	88.2%	88.9%	32.1%↑	20.2%
News	86.5%	87.2%	30.4%↑	18.5%
FGQA	56.1%	68.7%	–	–

Note: ΔGR and ΔFGR denote the score drop from the public dataset to FGQA for GraphRAG and F-GraphRAG, respectively. ↑All ΔGR values exceed 30%.

Public benchmark subsets. For cross-domain validation, we construct fine-grained subsets (1,000 questions each) from nine public datasets spanning biomedical, financial, legal, news, governmental, historical, scientific, educational, and technical domains. These subsets select questions with the highest attribute density and are re-annotated with A_q (Appendix B.3). Unlike FGQA, the original benchmarks lack expert-labeled A_q and thus cannot compute our three diagnostic metrics; they serve exclusively for generalization testing.

Experimental Configuration. Main results on FGQA are reported in Table 2. All methods use the identical document corpus, embedding model (`text-embedding-3-large`), generation model (DeepSeek-V3, temperature=0.1), and context budget of $B = 4096$ tokens. Results are averaged over five independent runs (seeds: 42, 2024, 1024, 2048, 3072). We employ Bayesian optimization (Snoek et al., 2012) for hyperparameter tuning (Appendix D.6) and compare against 7 graph-based RAG baselines (GraphRAG, LightRAG, FastGraphRAG, E^{2}GraphRAG, PathRAG, QCG-RAG, CheapRAG) and 16 dense retrieval models; the complete list is in Appendix F.

5.2 Ablation Study and Component Synergy

We isolate the contributions of F-GraphRAG's two core components—the granularity-aware extractor and the semantic-aware filter —by removing each and measuring the resulting degradation.

Diagnostic analysis. Using our stage-wise metrics on a 500-query annotated set, we find that the extractor counteracts semantic compression, the filter overcomes global importance bias, and only when both operate in tandem is the FGFR cascade fully broken. These findings align with

Table 2: Fine-Grained QA Accuracy on FGQA.

Category	Model	Accuracy
Graph Retrieval	**F-GraphRAG (Ours)**	**68.7% ± 0.8%**†
	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%
	bge-large	60.3% ± 0.8%

Note: †$p < 0.001$ vs. GraphRAG (paired t-test).

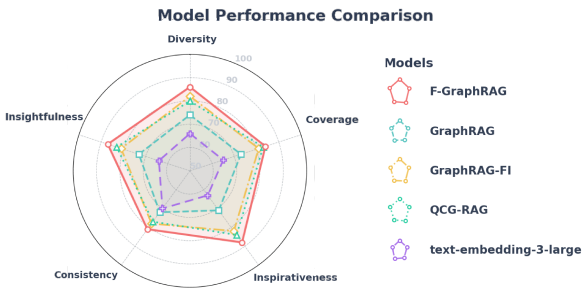

Figure 4: Radar chart comparing five representative methods on five global reasoning quality dimensions.

the root-cause attribution in our diagnostic experiments, where 62% of errors stem from representation failure and 24% from selection failure. Detailed numeric results are given in Table 4.

End-to-end accuracy. Table 3 translates diagnostic gains into QA performance. Removing filter costs 3.4 points, removing extractor costs 7.6 points, and removing both costs 12.6 points. The fallback dense retrieval safeguard contributes +1.9 points on queries with insufficient graph context.

Table 4 further reveals how the two components individually repair the FGFR cascade: the extractor predominantly lifts R_{rep} from 58% to 76%, while the filter primarily boosts R_{sel} from 52% to 73%.

Hyperparameter sensitivity. All fixed coefficients in the semantic-aware filter were optimized via Bayesian optimization (Snoek et al., 2012) on a 200-query development set (see Appendix D.6 for the full configuration). To assess stability, we perturb each coefficient by $\pm 20\%$ around its optimal value. Table 5 reports the accuracy on the development set, showing that end-to-end accuracy varies by less than one percentage point across

Table 3: Accuracy under component ablation.

Configuration	E	F	FFR	Accuracy	Δ
F-GraphRAG (full)	✓	✓	✓	**68.7 (0.8)**	–
− FFR only	✓	✓	×	66.8 (0.9)	−1.9
− Filter only	✓	×	✓	65.8 (0.9)	−2.9
− Extractor only	×	✓	✓	61.4 (1.0)	−7.3
− Filter & FFR	✓	×	×	65.3 (0.9)	−3.4
− Extractor & FFR	×	✓	×	61.1 (1.1)	−7.6
− Extractor & Filter	×	×	✓	58.2 (1.1)	−10.5
Standard GraphRAG	×	×	×	56.1 (1.2)	−12.6

Note: ✓= present; × = removed. Δ = drop from full model.

Table 4: 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.

all perturbations, confirming the robustness of the semantic-aware filter. Similarly, the priority coefficient α in the granularity-aware extractor was tuned via a sensitivity grid; as shown in Table 6, accuracy remains stable within $\pm 0.4\%$ under $\pm 20\%$ perturbation around the chosen value of $\alpha = 0.4$.

5.3 Global Reasoning Evaluation

We evaluate global reasoning on 1,000 queries from *Dream of the Red Chamber* using a pairwise comparison protocol. Two domain experts independently judge anonymized method outputs on five dimensions: Coverage, Diversity, Insightfulness, Consistency, and Inspirativeness. 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, and complete pairwise matrices are in Appendix E.

5.4 Diagnostic Analysis

We sample 500 queries from the FGQA test set and annotate their query-relevant attribute sets A_q following the protocol in Appendix C. On this diagnostic subset, baseline GraphRAG achieves only $\overline{R_{\text{rep}}} = 0.58$, $\overline{R_{\text{sel}}} = 0.52$, and $\overline{R_{\text{end}}} = 0.30$. Root-cause attribution reveals that 62% of errors stem from representation failure and 24% from selection failure, motivating our two-pronged design: the granularity-aware extractor targets the domi-

Table 5: Sensitivity of accuracy to coefficient perturbations in the semantic-aware 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.

Table 6: Sensitivity to the priority coefficient α.

α	Accuracy	R_{rep}	Graph size	Δ Acc.
0.2	66.1 (0.9)	71	38.2	-2.4
0.3	69.2 (0.8)	73	41.5	-1.3
0.4	**68.7 (0.8)**	**76**	**44.7**	–
0.5	70.3 (0.7)	77	48.9	-0.2
0.6	70.1 (0.9)	78	53.1	-0.4

Note: Δ Acc. = Accuracy drop from $\alpha = 0.4$. Graph size in relations per 1k tokens.

Table 7: Correlation between $R_{\text{end}}(q)$ and accuracy.

Quintile	Mean R_{end}	Accuracy
1 (Lowest)	0.15	12.4% $\pm$ 3.1%
2	0.35	31.2% $\pm$ 3.8%
3	0.52	47.8% $\pm$ 4.2%
4	0.73	64.1% $\pm$ 3.5%
5 (Highest)	0.89	85.6% $\pm$ 2.8%

Table 8: 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).

nant representation bottleneck, while the semantic-aware filter addresses the secondary selection failure. These two components reinforce each other—the extractor supplies the filter with a richer pool of attribute-carrying relations, enabling more effective semantic ranking than would be possible over a standard graph. The ablation results in Table 4 confirm this synergy: the extractor lifts R_{rep} from 58% to 76%, the filter boosts R_{sel} from 52% to 73%, and their combination yields $R_{\text{end}} = 55\%$, higher than either component alone. The fallback dense retrieval mechanism addresses the remaining 3% of queries where the main pipeline yields insufficient context, serving as an effective safeguard. A monotonic relationship between $R_{\text{end}}(q)$ and accuracy validates our framework (Table 7; logistic regression odds ratio 4.2, $p < 0.001$).

5.5 Practical Benefits

Token efficiency. F-GraphRAG achieves substantial token savings over standard GraphRAG as query volume grows. Table 8 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%.

Fewer relationships, higher informational quality. The token savings stem from using 33.8% fewer relationships per query (196 vs. 296). This reduction is driven by the granularity-aware extractor assigning higher priority to attribute-carrying relations and the semantic-aware 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 4.

Answer stability. The framework demonstrates stability (73% vs. 61% across 10 runs; Figure 6). This improvement originates from the granularity-aware extractor, which produces a knowledge graph where fine-grained attributes are preserved as first-class citizens. When the semantic-aware filter retrieves from this graph, the selected evidence is more determinate and complete, reducing the LLM's interpretive variance. Consequently, numerical errors drop from 40% to 16% and attribute errors from 32% to 11%, confirming accuracy gains reflect a stronger evidential foundation.

The dual evidence of token efficiency and improved diagnostic metrics reveals an insight: informational quality takes precedence over relational quantity in graph-based RAG. By shifting the design focus from structural prominence to fine-grained richness, F-GraphRAG achieves gains in accuracy, efficiency, and stability.

5.6 Cross-Domain Validation

To verify the generalizability of the FGFR cascade and our proposed mitigations, we construct fine-grained evaluation subsets from nine public benchmarks spanning biomedical (BioASQ (Tsatsaronis

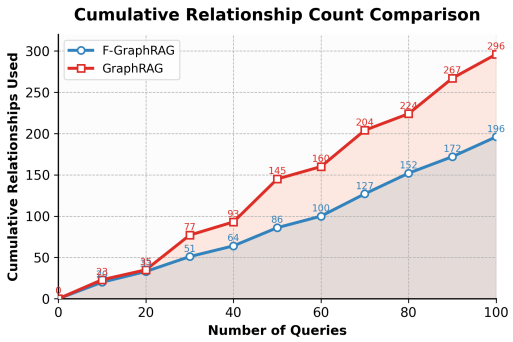

Figure 5: Cumulative relationships over 100 queries.

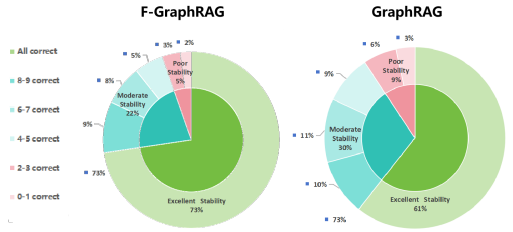

Figure 6: Answer consistency across 10 runs.

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 (HistoryQA (Bai et al., 2024)), scientific (Qasper (Dasigi et al., 2021)), educational (RACE (Lai et al., 2017)), and technical (TechQA (Castelli et al., 2020)) domains. It is important to note that these subsets are not the original benchmarks used off-the-shelf; rather, for each benchmark we deliberately select the top 1,000 questions with the highest attribute density, concatenate each question with its full supporting context to form long-context retrieval scenarios that mirror the complex contextual dependencies of our FGQA benchmark, and re-annotate them with query-relevant attribute sets A_q, ensuring they align with the FGFR task and can meaningfully stress-test fine-grained fact retention (see Appendix B.3 for detailed construction protocol; annotation protocol in Appendix C).

As shown in Table 9, F-GraphRAG outperforms GraphRAG across all domains, with improvements ranging from +5.3% to +12.6%. Larger gains are observed in domains with higher attribute density. A Pearson correlation ($r = 0.78$, $p = 0.008$) exists between average attribute density and improvement magnitude, confirming that the FGFR cascade is a domain-general phenomenon whose severity scales with fine-grained fact density, and that F-GraphRAG's mitigation strategies transfer effectively across diverse contexts.

Table 9: 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	HistoryQA	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: All improvements are significant at $p < 0.05$ by paired t-test (†).

5.7 Discussion

Coordinated intervention on FGFR's root causes yields an insight: informational quality takes precedence over relational quantity. This quality-over-quantity paradigm—achieving higher accuracy with fewer edges—is validated by our diagnostic analysis. Cross-domain results reveal FGFR not as a binary failure condition but as a spectrum: the denser the facts, the more pronounced the cascade. Every GraphRAG deployment operates somewhere on this spectrum, and our framework provides mitigation along its span—from low-attribute encyclopedic QA to high-attribute narrative reasoning. These findings challenge the triple-centric paradigm in GraphRAG: we advocate moving from triple-only representations toward attributed quadruples, where fine-grained details are first-class citizens, realigning graph construction with informational demands of complex reasoning.

6 Conclusion

Building on the FGFR formalization and the FGQA benchmark, we proposed F-GraphRAG, which breaks the cascade through a granularity-aware extractor and a semantic-aware filter. Experiments show 12.6% improvement over GraphRAG on fine-grained QA with 33.8% fewer relationships, and cross-domain gains of 5.3–12.6%. These results establish that prioritizing informational quality over relational quantity is a paradigm for faithful and efficient RAG. This diagnostic-to-intervention pipeline shows that FGFR failures are repairable, shifting focus decisively upstream to graph construction quality and providing a robust blueprint for RAG systems where fine-grained grounding is not sacrificed for structural abstraction.

7 Limitations

The present study establishes the FGFR diagnostic framework and demonstrates F-GraphRAG's effectiveness; several natural extensions define promising directions for future investigation.

Attribute Propagation Across Sentences. The granularity-aware extractor operates at the sentence level, a deliberate design choice that ensures high precision and applicability across languages. Extending this to capture attribute dependencies that span multiple sentences or discourse structures—for instance, by enriching windows with coreference information—constitutes a natural next step for modeling long, complex factual chains.

Incremental Graph Construction. The knowledge graph is constructed offline to ensure reproducibility and controlled evaluation. Adapting the pipeline to streaming or incremental indexing scenarios would broaden its applicability to growing corpora and real time knowledge bases.

Support for Diverse Languages. The window mechanism employs a pluggable language interface that performs robustly with high quality pluggable dependency parsers and then degrades gracefully to fallbacks at the entity or sentence level. Fully agnostic strategies, such as span detection based on alignment or embeddings that work across languages, represent a promising open research direction for deployment in varied linguistic settings.

References

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

Riccardo Albertoni, Elena Camossi, Monica De Martino, Franca Giannini, and Marina Monti. 2006. Semantic granularity for the semantic web. In *On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops*, volume 4278 of *Lecture Notes in Computer Science*, pages 1863–1872. Springer Berlin Heidelberg.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 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.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690.

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. 2025. PathRAG: Pruning graph-based retrieval augmented generation with relational paths. *arXiv preprint arXiv:2502.14902*. Cs.CL.

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*. Cs.IR.

Yifan Feng, Hao Hu, Xingliang Hou, Shiquan Liu, Shihui Ying, Shaoyi Du, Han Hu, and Yue Gao. 2025. Hyper-rag: Combating llm hallucinations using hypergraph-driven retrieval-augmented generation. *arXiv preprint arXiv:2504.08758*.

Kai Guo, Harry Shomer, Shenglai Zeng, Haoyu Han, Yu Wang, and Jiliang Tang. 2025. Empowering graphrag with knowledge filtering and integration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25450–25464.

Zhicheng Guo, Lei Xia, Yi Yu, Tong Ao, and Chao Huang. 2024. LightRAG: Simple and fast retrieval-augmented generation. *arXiv preprint arXiv:2410.05779*. Cs.CL.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*. Cs.CL.

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*.

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*. Cs.IR.

Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. 2021a. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2):494–514.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2021b. Survey of hallucination in natural language generation. *Comput. 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, and Tim Rocktäschel. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Jiacheng Liang, Yuhui Wang, Changjiang Li, Tanqiu Jiang, Rongyi Zhu, Neil Gong, and Ting Wang. 2025. GraphRAG under fire. *arXiv preprint arXiv:2501.14050*. Cs.IR.

Xun Liang, Shichao Song, Simin Niu, Zhiyu Li, Feiyu Xiong, Buzhou Tang, Yang Wang, Dawei He, Cheng Peng, Zhonghao Wang, and Haiying Deng. 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.

Shirui Pan, Lei Luo, Yuxuan Wang, and 1 others. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599.

Boci Peng, Yutao Zhu, Yang Liu, and 1 others. 2024. Graph retrieval-augmented generation: A survey. *ACM Transactions on Information Systems*. To appear.

Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. 2012. Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc. NeurIPS.

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*. Cs.IR.

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:2406.15197*. 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. *arXiv preprint arXiv:2402.09267*. Cs.CL.

Yibo Zhao, Junchao Zhu, Yan Guo, Kaimin He, and Xiang Li. 2025. E^2GraphRAG: Streamlining graph-based RAG for high efficiency and effectiveness. *arXiv preprint arXiv:2505.24226*. Cs.IR.

Yingli Zhou, Yaodong Su, Youran Sun, Shu Wang, Taotao Wang, Runyuan He, Yongwei Zhang, Sicong Liang, Xilin Liu, Yuchi Ma, and Yixiang Fang. 2025. In-depth analysis of graph-based rag in a unified framework (with cheaprag extension). *arXiv preprint arXiv:2503.04338*. Cs.IR.

A Statistical Analysis Protocol

All models were evaluated with LLM temperature=0.1 for strict 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 Cohen's d effect sizes. Per-dataset p-values: BioASQ < 0.001, ConvFinQA 0.002, CUAD < 0.001, DROP 0.001, GovReport 0.03, HistoryQA 0.002, Qasper 0.001, RACE < 0.001, TechQA 0.04.

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.

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. 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 a fine-grained subset through a three-step pipeline. First, we compute the attribute density of each question as the number of numerical values, temporal expressions, and entity-specific modifiers appearing in its ground-truth supporting context. We then select the top 1,000 questions with the highest attribute density, ensuring that each subset contains sufficient fine-grained facts to stress-test the FGFR cascade. Second, we concatenate each question with all paragraphs identified as supporting evidence in the original benchmark annotations, producing long-context retrieval scenarios with an average of 1,200 tokens per query. Third, 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 as the cross-domain evaluation sets in Section 5.6.

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 the first testbed where the FGFR cascade 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 (100 per quintile). 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. 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").

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 Cross-Lingual Adaptation

For English datasets, we replace the Chinese dependency parser with `spaCy en_core_web_lg` for evidence window detection. The eight attribute categories remain identical; 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.4 Algorithm Details

D.4.1 Semantic-Aware Filtering Algorithm

Algorithm 2 details the complete semantic-aware 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 Semantic-Aware 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 SemanticAwareFilter(G, q, E_s, k, ρ)
    E_cand ← GetKCandidateRelationships(G, E_s, k)
    for each r ∈ E_cand do
        Compute S_attr(r), S_ling(r), S_net(r)
    end for
    for dim ∈ {attr, ling, net} do
        μ_dim ← mean({S_dim(r) | r ∈ E_cand})
        σ_dim ← std({S_dim(r) | r ∈ E_cand})
        CV_dim ← σ_dim/(μ_dim + ϵ)          ▷ ϵ = 10^-6
    end for
    w_dim ← CV_dim / Σ_d CV_d               ▷ Handling zero variances
    for each r ∈ E_cand do
        S[r] ← Σ_dim w_dim · S_dim(r)
    end for
    μ_S ← mean(S)
    σ_S ← std(S)
    θ ← μ_S + 1.5 · σ_S
    E_keep ← {r | S[r] > θ}
    E_struct ← structural bridges and outliers
    target ← ⌊ρ · |E_cand|⌋
    remaining ← target − |E_keep ∪ E_struct|
    if remaining > 0 then
        return E_keep ∪ E_struct ∪ TopK(S, remaining)
    else
        return E_keep ∪ E_struct
    end if
end function
```

D.5 Formal Definitions of Scoring Functions

- **Semantic alignment** $R(r) = \cos(\mathbf{v}_{\text{pattern}}(r), \mathbf{v}_{\text{indicator}})$, where $\mathbf{v}_{\text{pattern}}$ embeds the relation verbalization and $\mathbf{v}_{\text{indicator}}$ averages embeddings of fine-grained indicator terms.
- **Information density** $I(r) = (|E_r| + |A_r|)/\text{len}(t)$, where $|E_r|$ is the number of unique entities, $|A_r|$ the number of unique attributes, and $\text{len}(t)$ the token count of the source sentence.
- **Syntactic complexity** $C(r)$ is the maximum depth of the dependency parse tree of the source sentence.
- **Local connectivity** $\text{conn}(r) = (\deg(s) + \deg(o))/(2 \cdot \deg_{\max})$.
- **Bridge centrality** $\text{bridge}(r)$: edge betweenness centrality on the candidate subgraph, normalized to $[0, 1]$.

D.6 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. Optimal parameters: adaptive weights $w_{\text{attr}} = 0.5$, $w_{\text{ling}} = 0.3$, $w_{\text{net}} = 0.2$; $k = 1.5$, $\rho = 0.7$; fixed coefficients $\lambda_{a1} = 0.55$, $\lambda_{a2} = 0.45$, $\lambda_{l1} = 0.35$, $\lambda_{l2} = 0.40$, $\lambda_{l3} = 0.25$, $\lambda_{n1} = 0.65$, $\lambda_{n2} = 0.35$.

E Global Reasoning Evaluation

E.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 surface 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 $\kappa = 0.84$.

E.2 Case Study with Dimension-Wise Analysis

Table E.5 illustrates how the five evaluation dimensions are applied to a representative global reasoning query, comparing F-GraphRAG and standard GraphRAG outputs.

Table E.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

E.3 Complete Pairwise Comparison Matrices

Tables E.6–E.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).

Table E.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 E.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	41.65	38.33	35.47	41.77	50.00	47.39
T3L	35.18	41.82	39.24	35.22	43.03	52.61	50.00

Table E.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 E.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	54.32	49.27	49.68	50.01	50.00	49.53
T3L	49.14	55.61	52.89	51.63	52.58	50.47	50.00

Table E.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	48.76	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	34.02	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

F Full Model Comparison

This table extends the representative subset shown in Table 2 of the main text.

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

Category	Model	Accuracy
Graph Retrieval	**F-GraphRAG (Ours)**	**68.7 (0.8)**
	GraphRAG-FI	64.0 (1.0)
	QCG-RAG	66.1 (1.0)
	E^2GraphRAG	64.3 (1.1)
	GraphRAG	56.1 (1.2)
	FastGraphRAG	58.9 (1.3)
	LightRAG	55.8 (1.4)
	CheapRAG	60.2 (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)

G Effect Sizes for Main Comparisons

All reported effect sizes (Cohen's d) indicate very large effects ($d > 0.8$).

Table G.12: Effect sizes for main comparisons.

Comparison	Cohen's d	Interpretation
F-GraphRAG vs. GraphRAG	8.2	Very large
F-GraphRAG vs. text-embedding-3-large	9.0	Very large
F-GraphRAG vs. DashScope-text-embedding-v4	1.1	Large

H Failure Analysis of Baseline GraphRAG

Numerical and material attributes account for over half of all omissions.

Table H.13: 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")