RELIABLE EVIDENCE PATH SELECTION IN GRAPH-BASED RAG VIA FUZZY EVIDENCE-GUIDED REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-augmented generation (RAG) grounds large language models in external knowledge, while graph-based RAG organizes retrieved information into entities, relations, and paths to support complex reasoning. However, graph connectivity and semantic relevance do not guarantee evidential reliability. Candidate paths may be relevant yet weakly supportive, conflicting, uncertain, or redundant. These conditions vary in degree and may coexist, while limited context budgets amplify the cost of admitting unreliable evidence. To address this problem, we propose Fuzzy Evidence-Guided Reasoning (FEGR), a graph-based RAG system for reliable evidence path selection. Specifically, FEGR models relevance, supportiveness, conflict, and uncertainty as fuzzy memberships, aggregates them at the path level through noncompensatory inference, and selects a compact set of complementary evidence paths under a token budget for answer generation. Across five domains, FEGR achieves Overall win rates above parity against all five representative RAG baselines under a common budget for retrieved context, with consistent gains in Supportiveness and Conflict Avoidance and lower admission of mixed evidence risks.

1 INTRODUCTION

Retrieval-augmented generation (RAG) grounds large language models (LLMs) in knowledge retrieved from external corpora Lewis et al. (2020); Asai et al. (2024). Text-based RAG represents a corpus as independent chunks and ranks them by lexical or semantic similarity Robertson & Zaragoza (2009); Karpukhin et al. (2020). Although effective for many knowledge-intensive queries, this flat organization is less suitable when an answer depends on relations among entities, events, or claims distributed across multiple documents. Graph-based RAG instead organizes knowledge through entities, relations, passages, communities, and paths, thereby preserving dependencies that independent chunk retrieval may obscure Edge et al. (2024); Guo et al. (2025); Gutiérrez et al. (2024); Chen et al. (2026a); Zhu et al. (2025). This organization improves access to related and multi-hop evidence, but leaves open how candidates should be allocated within a limited context.

Existing graph-based RAG methods improve graph construction, candidate retrieval, community organization, relational path discovery, or redundancy control. Recent studies of factual conflict show that relevance matching alone is insufficient when retrieved claims are incompatible or weakly grounded Zhang et al. (2025); Liu et al. (2026). Despite these advances, *graph connectivity and query relevance do not guarantee evidential reliability*. This limitation reflects an objective mismatch. Graph retrieval commonly seeks structurally connected, semantically related, or broadly covering candidates, whereas context construction must allocate limited capacity according to evidential contribution and risk. A candidate path may be related to the query yet provide partial support Joren et al. (2025). It may conflict with comparable evidence in factual content, time, or scope, or rely on ambiguous or weakly grounded statements. Redundant paths consume context capacity without adding new information. Relevance, supportiveness, conflict, and uncertainty may coexist and vary by degree. A path can be partly supportive while remaining mildly conflicting or uncertain, and reducing these conditions to a binary reliability judgment would discard distinctions relevant to selection. As Figure 1 illustrates, admitting a low-value or risky path under a fixed bud-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

get can displace more supportive evidence. The problem is how to account for graded evidence conditions when selecting complementary evidence paths under limited context budgets.

To address this problem, we propose **Fuzzy Evidence-Guided Reasoning (FEGR)**, a graph-based RAG framework for reliable evidence path selection under limited context budgets. FEGR retrieves query-relevant evidence and organizes it into candidate paths. It represents each evidence unit through graded memberships for relevance, supportiveness, conflict, and uncertainty. These conditions are aggregated through noncompensatory path-level inference, limiting the influence of strong positive signals when severe evidence risks are present. Under a token budget, FEGR selects a compact set of reliable and complementary paths by considering path reliability, query-anchor coverage, incremental token cost, and evidence identity overlap. Selection accounts for both the quality of each candidate path and the evidence already admitted. The selected evidence is deduplicated and serialized into the context for answer generation.

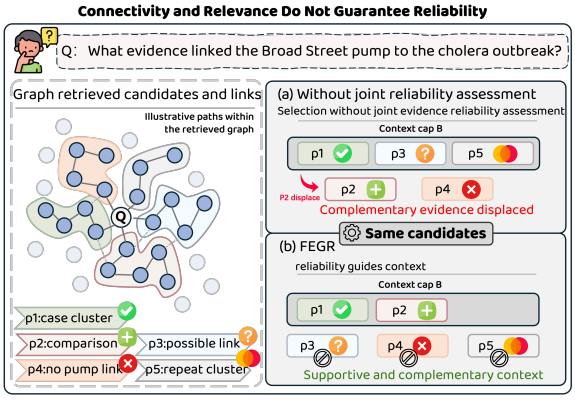


Figure 1: Motivation for FEGR: connected and query-relevant paths may still be weakly supportive, conflicting, uncertain, or redundant, displacing more useful evidence under a fixed context budget.

Across five domains, FEGR obtains Overall point estimates above parity against five text RAG and graph-based RAG systems under a common cap on retrieved context. The observed gains are especially consistent in Supportiveness and Conflict Avoidance, the two dimensions most directly aligned with evidence reliability. Budget and stress analyses indicate larger observed advantages under small and moderate context budgets, together with lower admission of mixed evidence risks. Component ablations support complementary contributions from the principal FEGR components.

The main contributions of this work are as follows:

- We formulate evidence path selection under limited context budgets as a reliability-aware context allocation problem in graph-based RAG.
- We propose FEGR, a graph-based RAG framework that combines graded evidence memberships, noncompensatory path-level reliability inference, and budgeted selection of complementary evidence paths.
- We evaluate FEGR across five domains through comparisons of complete systems and controlled analyses of context budgets, mixed evidence risks, and component contributions.

2 Related Work

2.1 Text RAG and Context Organization

Text RAG systems retrieve passages through sparse lexical matching Robertson & Zaragoza (2009) or dense representations Karpukhin et al. (2020). Later work explores long-context compression Jiang et al. (2024). Self-RAG learns when to retrieve and critique outputs Asai et al. (2024), while RAPTOR retrieves recursive summaries Sarthi et al. (2024). RankRAG jointly instruction-tunes an LLM for context ranking and answer generation Yu et al. (2024). MemoRAG uses global memory for long contexts Qian et al. (2025), and GraphReader explores graph-structured documents with limited context Li et al. (2024a). Long-context analyses show that evidence use remains sensitive to position and context composition Liu et al. (2024b); Cuconasu et al. (2024).

2.2 Graph-Based RAG

Graph-based RAG organizes knowledge through entities, passages, concepts, communities, and relations. GraphRAG builds knowledge graphs and community summaries Edge et al. (2024), Ligh-

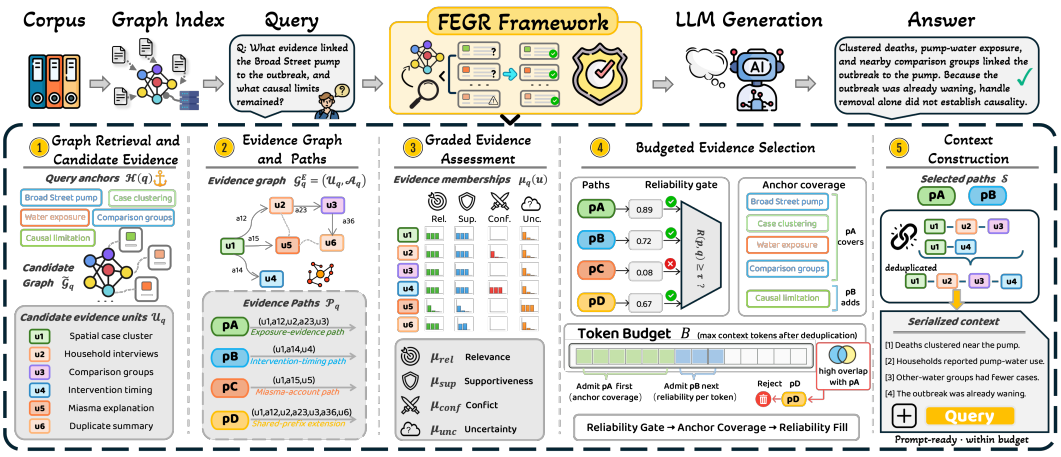


Figure 2: Overview of the FEGR framework for reliability-aware graph-based retrieval-augmented generation. The framework integrates graph retrieval, evidence path organization, fuzzy reliability inference, budgeted evidence selection, and final context construction before answer generation.

tRAG combines local and global graph retrieval Guo et al. (2025), and HippoRAG variants retrieve connected evidence through associative memory Gutiérrez et al. (2024; 2025). Other work retrieves paths or subgraphs for multidocument and graph reasoning Wang et al. (2024b); He et al. (2024); Liu et al. (2024a); Luo et al. (2024). PathRAG retrieves and ranks relational paths through flow-based pruning Chen et al. (2026a), while recent methods construct query-specific or constrained reasoning graphs Huang et al. (2026); Chen et al. (2026b); Cao et al. (2026). These approaches advance graph retrieval and evidence organization, but leave less explored how graded evidence conditions should guide path selection under limited context budgets.

2.3 Evidence Reliability in RAG

Recent work examines reliability and factual conflict in RAG Niu et al. (2024); Chen et al. (2024b); Park & Lee (2024); Shen et al. (2024). FaithfulRAG models fact-level conflicts between retrieved context and parametric knowledge Zhang et al. (2025), whereas TruthfulRAG uses knowledge graphs and entropy-based filtering to identify and mitigate factual inconsistencies Liu et al. (2026). Both address discrepancies between retrieved information and the model's parametric knowledge. Fuzzy set theory represents graded conditions through membership functions, while t-norms provide conjunctive operators for combining membership degrees Zadeh (1965); Klement et al. (2000); Dubois & Prade (1985). FEGR incorporates locally assessed evidence conflicts alongside graded relevance, supportiveness, and uncertainty in budgeted evidence path selection.

3 Fuzzy Evidence-Guided Reasoning

Figure 2 presents the FEGR framework. FEGR constructs a graph index from the input documents, retrieves query-relevant evidence, organizes it into candidate paths, evaluates four graded evidence conditions, and selects a budget-feasible context for answer generation.

3.1 Problem Formulation

Given a document collection $\mathcal{D}$, FEGR constructs a graph index G that organizes textual evidence and its associated entities and relations. For an input query q, FEGR retrieves query-relevant graph elements and their associated textual evidence from G, preserving their structural and relational connections.

Given a retrieved-context budget B, FEGR organizes the retrieved evidence into candidate paths and selects a subset to construct a deduplicated context C_B satisfying $\mathrm{Tok}(C_B) \leq B$. Selection considers path reliability, query-anchor coverage, incremental token cost, and evidence identity overlap. The resulting context is supplied with the query to the answer generator.

3.2 Query-Conditioned Evidence Organization

FEGR converts the retrieved candidates into textual evidence units while retaining the structural links needed for local comparison and path construction. This produces the query-conditioned evidence graph

$$\mathcal{G}_q^E = (\mathcal{U}_q, \mathcal{A}_q), \tag{1}$$

where $\mathcal{U}_q$ contains the textual evidence units derived from the retrieved candidates, and $\mathcal{A}_q$ retains their structural and relational connections. Each unit has a stable identity derived from its source provenance, which supports subsequent deduplication and overlap control.

Within $\mathcal{G}_q^E$, FEGR constructs a finite set of candidate paths $\mathcal{P}_q = \{p_1, \ldots, p_n\}$ through bounded graph traversal. A path $p = (u_1, a_1, u_2, \ldots, a_{m-1}, u_m)$ is an ordered sequence of evidence units $u_i \in \mathcal{U}_q$ connected by links $a_i \in \mathcal{A}_q$. The ordered representation preserves the evidence chain, whereas $\mathcal{U}(p) = \{u_1, \ldots, u_m\}$ denotes its unit set for membership aggregation and overlap control. Path construction follows fixed limits on path length and candidate count, and individual evidence units are retained as valid paths.

3.3 Fuzzy Evidence Modeling

For each evidence unit $u \in \mathcal{U}_q$, FEGR computes four graded memberships representing relevance, supportiveness, conflict, and uncertainty:

$$\boldsymbol{\mu}_q(u) = \left[\mu_{\mathrm{rel}}(u,q), \mu_{\mathrm{sup}}(u,q), \mu_{\mathrm{conf}}(u,q), \mu_{\mathrm{unc}}(u,q)\right]^\top \in [0,1]^4. \tag{2}$$

Relevance and supportiveness describe positive evidence conditions, whereas conflict and uncertainty represent distinct sources of risk. These conditions may coexist within the same unit and are aggregated according to their roles during path-level inference.

Relevance, supportiveness, and uncertainty are computed from fixed weighted combinations of normalized operational signals. For $x \in \{\mathrm{rel}, \mathrm{sup}, \mathrm{unc}\}$, their common aggregation rule is

$$\mu_x(u,q) = \frac{\sum_{j \in \mathcal{I}_x(u)} w_{x,j}\, s_{x,j}(u,q)}{\sum_{j \in \mathcal{I}_x(u)} w_{x,j}}, \tag{3}$$

where $\mathcal{I}_x(u)$ contains the available signals, $s_{x,j}(u,q) \in [0,1]$ is a normalized signal, and $w_{x,j}$ is its fixed weight. Unavailable signals are excluded and the remaining weights are renormalized. Relevance combines retrieval confidence, typed query-anchor coverage, and entity or relation correspondence. Supportiveness estimates how well a unit covers information required by the query, using anchor coverage, relation cues, answer-bearing content, and statement completeness. Unlike the marginal coverage used during path selection, it is assessed for each unit before answer generation. Uncertainty captures hedging, ambiguity, weak grounding, low retrieval confidence, incomplete scope, and structural content risks.

Conflict is assessed through local comparisons between evidence units. Let $\mathcal{N}_q(u)$ contain path-adjacent units and units sharing query-associated entities or relations, and let $\mathcal{K}$ denote the conflict detection rules. Claims are compared only when their subjects, factual attributes, and applicable scopes align. FEGR computes $\mu_{\mathrm{conf}}(u,q) = \max_{v \in \mathcal{N}_q(u),\, k \in \mathcal{K}} d_k(u,v)$, where $d_k(u,v)$ is the weighted response of rule k, including applicable source-provenance adjustments. The rules cover numerical and temporal disagreements, incompatible entity values, explicit corrections, negation mismatches, and mutually exclusive states. The membership is zero when no aligned conflict signal is detected. Complete signal definitions, fixed weights, normalization procedures, and provenance adjustments are provided in Appendix A.1.

3.4 Path-Level Reliability Inference

FEGR aggregates unit memberships into evidence conditions at the path level. Relevance and supportiveness use the geometric mean, $M_x(p,q) = \mathrm{GM}_{u \in \mathcal{U}(p)}\, \mu_x(u,q)$ for $x \in \{\mathrm{rel}, \mathrm{sup}\}$, which remains sensitive to weak constituent units. Conflict and uncertainty use the maximum, $M_x(p,q) = \max_{u \in \mathcal{U}(p)} \mu_x(u,q)$ for $x \in \{\mathrm{conf}, \mathrm{unc}\}$, so that a strong risk signal is not diluted by other units in the path.

A reliable path should combine strong relevance and supportiveness with low conflict and uncertainty. FEGR represents this conjunction using the product t-norm:

$$R(p,q) = M_{\mathrm{rel}}(p,q)M_{\mathrm{sup}}(p,q)\big(1 - M_{\mathrm{conf}}(p,q)\big)\big(1 - M_{\mathrm{unc}}(p,q)\big). \tag{4}$$

Because every factor lies in $[0,1]$, the resulting score cannot exceed any individual factor. A severe deficiency in one condition therefore limits the overall score without additive compensation from the others. FEGR uses $R(p,q)$ as an operational path-selection score rather than a calibrated probability of factual correctness. Path length incurs no separate reliability penalty; its rendered token cost is considered during selection.

3.5 Budgeted Evidence Selection

Ranking paths solely by reliability may repeatedly favor evidence addressing the same aspect of a query, leaving other information requirements uncovered. FEGR therefore applies a reliability gate followed by two greedy phases: query-anchor coverage and reliability-based filling of the remaining budget. The admissible candidate pool is

$$\mathcal{P}_q^{\tau} = \{p \in \mathcal{P}_q \mid R(p,q) \geq \tau_R\}, \tag{5}$$

where τ_R is selected from a fixed descending threshold schedule. FEGR uses the first threshold that retains a preset minimum number of paths, falling back to the final threshold when necessary.

For a selected path set $\mathcal{S} \subseteq \mathcal{P}_q$, let $C(\mathcal{S})$ denote the rendered context after evidence identity deduplication, with token cost $\mathrm{Cost}(\mathcal{S}) = \mathrm{Tok}(C(\mathcal{S}))$, where Tok denotes the token count computed by the context tokenizer. The incremental cost of adding a path p is

$$\mathrm{Tok}(p \mid \mathcal{S}) = \mathrm{Cost}(\mathcal{S} \cup \{p\}) - \mathrm{Cost}(\mathcal{S}). \tag{6}$$

Coverage of query anchors. Let $\mathcal{H}(q)$ denote the typed query anchors, including lexical items, entities, and relation cues, and let $\mathcal{H}(p) \subseteq \mathcal{H}(q)$ contain the anchors matched by path p. For the current selection $\mathcal{S}$, the covered anchors are $\mathcal{H}(\mathcal{S}) = \bigcup_{p' \in \mathcal{S}} \mathcal{H}(p')$. Within a reserved fraction γB of the context budget, FEGR prioritizes feasible paths by uncovered anchor gain per incremental token:

$$g_{\mathrm{cov}}(p \mid \mathcal{S}, q) = \frac{|\mathcal{H}(p) \setminus \mathcal{H}(\mathcal{S})|_{\mathrm{anch}}}{\mathrm{Tok}(p \mid \mathcal{S})}. \tag{7}$$

Here, $|\cdot|_{\mathrm{anch}}$ denotes fixed weighted anchor mass, with generic lexical anchors excluded from core coverage. This phase stops when the reserved capacity is reached or no feasible candidate adds uncovered core anchor mass.

Filling the remaining budget. FEGR allocates the remaining context capacity by path reliability per incremental token:

$$g_{\mathrm{rel}}(p \mid \mathcal{S}, q) = \frac{R(p,q)}{\mathrm{Tok}(p \mid \mathcal{S})}. \tag{8}$$

Both selection objectives are recomputed after each admission because incremental cost and coverage depend on the evidence already selected.

As an additional redundancy check, FEGR measures the Jaccard overlap between the evidence unit identities of two paths:

$$\mathrm{red}(p_i, p_j) = \frac{|\mathcal{U}(p_i) \cap \mathcal{U}(p_j)|}{|\mathcal{U}(p_i) \cup \mathcal{U}(p_j)|}. \tag{9}$$

In both phases, candidates must satisfy the reliability gate, have positive incremental cost, fit the applicable budget, and have overlap no greater than a fixed threshold δ with every selected path. Feasible paths are admitted greedily under deterministic tie breaking. The overlap check uses evidence identities rather than semantic similarity, and selection may terminate below the budget when no feasible candidate remains. The complete algorithm and fixed parameter settings are provided in Appendix A.

3.6 Context Construction

FEGR serializes the selected paths into the final evidence context C_B. Paths are ordered by the earliest retrieval rank of their constituent units, with stable path identifiers used to resolve ties. Unit order within each path is preserved, and shared evidence is emitted only once. The context contains the selected textual evidence, associated entity and relation records, and source references, while excluding internal scores and selection diagnostics. Only relations with resolvable endpoints are retained. Before generation, FEGR checks the final serialization against budget B using the same tokenizer employed during selection. A budget violation triggers an assertion rather than silent truncation. The validated context is then supplied with the query to the answer generator.

4 Experiments

We evaluate FEGR through four research questions. **RQ1:** Does FEGR improve answer quality under a common retrieved-context cap? **RQ2:** How does its relative advantage change with context capacity? **RQ3:** How robust are graph-based RAG systems to controlled mixed evidence risks? **RQ4:** How do the principal components of FEGR contribute to its performance?

4.1 Experimental Setup

Data and baselines. We use five UltraDomain corpora Qian et al. (2025): Legal, History, Biology, Mix, and Politics. Each domain contains 40 validated queries spanning general retrieval, multi-hop reasoning, and cross-document synthesis, for 200 queries in RQ1. We compare FEGR with five representative RAG systems: NaiveRAG Gao et al. (2024), GraphRAG Edge et al. (2024), LightRAG Guo et al. (2025), PathRAG Chen et al. (2026a), and HippoRAG2 Gutiérrez et al. (2025). Each system uses its respective indexing, retrieval, and context-construction procedures, while sharing the corpora, queries, generator, generation configuration, and evaluation protocol. RQ1 uses a common 6K rendered-evidence cap. NaiveRAG appears only in RQ1; RQ2 and RQ3 compare FEGR with the four graph-based baselines. Detailed system configurations are provided in Appendix B.1.

Implementation and evaluation. We use BAAI/bge-m3 embeddings Chen et al. (2024a) and the frozen `gpt-4o-mini-2024-07-18` OpenAI (2024) for generation and blinded pairwise evaluation. In the absence of standardized reference answers, we follow prior graph-based RAG studies in adopting LLM-based pairwise evaluation Edge et al. (2024); Guo et al. (2025); Chen et al. (2026a). Each pair is judged anonymously in both presentation orders, with ties allowed, to mitigate position bias Zheng et al. (2023); Wang et al. (2024a); Li et al. (2024b); Shi et al. (2025). Following criterion-guided evaluation Liu et al. (2023), we assess Relevance, Supportiveness, Conflict Avoidance, Logicality, and Coherence. Tie-aware win rates assign ties half credit; Overall is the unweighted mean across criteria. Budgets cover only rendered evidence, excluding the query, system instructions, and generation template. All FEGR parameters are fixed across experiments; parameter settings and additional evaluation details appear in Appendix A.

4.2 Main Results (RQ1)

Table 1 reports pairwise results under the common 6K retrieved-context cap. FEGR obtains an above-parity Overall point estimate against every baseline in all five domains. Its unweighted averages across domains are 57.3% against NaiveRAG, 73.3% against GraphRAG, 58.5% against LightRAG, 65.1% against PathRAG, and 56.6% against HippoRAG2. HippoRAG2 yields the lowest FEGR-relative Overall point estimate under this protocol, followed closely by NaiveRAG and LightRAG.

The gains are especially consistent in Supportiveness and Conflict Avoidance, the two dimensions most directly aligned with evidence reliability. Against LightRAG, FEGR obtains domain-macro win rates of 61.3% and 59.8% on these dimensions, compared with 55.4% on Relevance. Against HippoRAG2, the corresponding values are 56.5%, 57.1%, and 55.4%. Logicality and Coherence also remain above parity, a pattern consistent with supplying more supportive and less conflicting context. Thus, RQ1 shows that the complete FEGR system achieves favorable point estimates under the common cap, particularly on reliability-related dimensions.

	Legal		History		Biology		Mix		Politics	
	NaiveRAG	**FEGR**	NaiveRAG	**FEGR**	NaiveRAG	**FEGR**	NaiveRAG	**FEGR**	NaiveRAG	**FEGR**
Relevance	43.1	**56.9**	45.0	**55.0**	40.0	**60.0**	48.1	**51.9**	45.0	**55.0**
Supportiveness	43.1	**56.9**	38.7	**61.3**	36.2	**63.8**	49.4	**50.6**	46.2	**53.8**
Conflict Avoidance	38.7	**61.3**	44.4	**55.6**	47.5	**52.5**	39.4	**60.6**	34.4	**65.6**
Logicality	40.6	**59.4**	38.7	**61.3**	42.5	**57.5**	46.9	**53.1**	45.0	**55.0**
Coherence	39.4	**60.6**	40.0	**60.0**	48.7	**51.3**	44.4	**55.6**	41.9	**58.1**
Overall	41.0	**59.0**	41.4	**58.6**	43.0	**57.0**	45.6	**54.4**	42.5	**57.5**
	GraphRAG	**FEGR**	GraphRAG	**FEGR**	GraphRAG	**FEGR**	GraphRAG	**FEGR**	GraphRAG	**FEGR**
Relevance	30.0	**70.0**	21.2	**78.8**	30.0	**70.0**	28.7	**71.3**	31.2	**68.8**
Supportiveness	26.9	**73.1**	18.7	**81.3**	30.0	**70.0**	31.2	**68.8**	25.0	**75.0**
Conflict Avoidance	28.7	**71.3**	16.2	**83.8**	33.7	**66.3**	28.7	**71.3**	30.0	**70.0**
Logicality	30.6	**69.4**	18.7	**81.3**	30.0	**70.0**	30.0	**70.0**	22.5	**77.5**
Coherence	26.9	**73.1**	13.7	**86.3**	31.2	**68.8**	30.0	**70.0**	25.0	**75.0**
Overall	28.6	**71.4**	17.7	**82.3**	31.0	**69.0**	29.7	**70.3**	26.7	**73.3**
	LightRAG	**FEGR**	LightRAG	**FEGR**	LightRAG	**FEGR**	LightRAG	**FEGR**	LightRAG	**FEGR**
Relevance	40.6	**59.4**	46.9	**53.1**	46.2	**53.8**	41.9	**58.1**	47.5	**52.5**
Supportiveness	37.5	**62.5**	45.0	**55.0**	37.5	**62.5**	33.7	**66.3**	40.0	**60.0**
Conflict Avoidance	36.9	**63.1**	41.2	**58.8**	41.2	**58.8**	41.2	**58.8**	40.6	**59.4**
Logicality	44.4	**55.6**	38.7	**61.3**	43.7	**56.3**	43.1	**56.9**	45.0	**55.0**
Coherence	38.7	**61.3**	43.7	**56.3**	41.2	**58.8**	43.7	**56.3**	37.5	**62.5**
Overall	39.6	**60.4**	43.1	**56.9**	42.0	**58.0**	40.7	**59.3**	42.1	**57.9**
	PathRAG	**FEGR**	PathRAG	**FEGR**	PathRAG	**FEGR**	PathRAG	**FEGR**	PathRAG	**FEGR**
Relevance	37.5	**62.5**	46.2	**53.8**	37.5	**62.5**	34.4	**65.6**	33.7	**66.3**
Supportiveness	37.5	**62.5**	43.7	**56.3**	40.0	**60.0**	33.1	**66.9**	31.2	**68.8**
Conflict Avoidance	35.6	**64.4**	42.5	**57.5**	36.9	**63.1**	31.2	**68.8**	21.9	**78.1**
Logicality	36.2	**63.8**	42.5	**57.5**	31.2	**68.8**	34.4	**65.6**	30.0	**70.0**
Coherence	31.2	**68.8**	35.0	**65.0**	31.2	**68.8**	28.1	**71.9**	29.4	**70.6**
Overall	35.6	**64.4**	42.0	**58.0**	35.4	**64.6**	32.2	**67.8**	29.2	**70.8**
	HippoRAG2	**FEGR**	HippoRAG2	**FEGR**	HippoRAG2	**FEGR**	HippoRAG2	**FEGR**	HippoRAG2	**FEGR**
Relevance	43.1	**56.9**	**51.2**	48.8	45.0	**55.0**	41.2	**58.8**	42.5	**57.5**
Supportiveness	40.0	**60.0**	45.0	**55.0**	46.2	**53.8**	45.0	**55.0**	41.2	**58.8**
Conflict Avoidance	44.4	**55.6**	48.7	**51.3**	36.9	**63.1**	42.5	**57.5**	41.9	**58.1**
Logicality	36.2	**63.8**	41.2	**58.8**	45.0	**55.0**	44.4	**55.6**	48.7	**51.3**
Coherence	38.1	**61.9**	43.7	**56.3**	46.2	**53.8**	43.7	**56.3**	41.2	**58.8**
Overall	40.4	**59.6**	46.0	**54.0**	43.9	**56.1**	43.4	**56.6**	43.1	**56.9**

Table 1: Domain-wise pairwise tie-aware win rates (%) under a 6K retrieved-context cap. Boldface marks each pair's higher score.

4.3 Budget Sensitivity (RQ2)

Using a fixed stratified subset of 100 queries with 20 per domain, we compare FEGR with LightRAG, GraphRAG, PathRAG, and HippoRAG2 under five fixed token caps of 3K, 6K, 9K, 12K, and 16K and under a Full setting that permits rendered evidence up to the common generator input limit. All other settings remain fixed. Figure 3(a) shows that the unweighted average over four baselines and five domains rises from 62.8% at 3K to its highest observed value of 66.0% at 9K, before declining to 59.0% at Full. This nonmonotonic trend indicates that increasing context capacity does not uniformly improve FEGR's relative performance and is consistent with an intermediate budget providing sufficient supportive evidence while preserving meaningful competition for limited context capacity.

The change from 6K to Full varies across baselines. Figure 3(b) shows decreases of 7–11 percentage points against LightRAG and 10–15 points against PathRAG across domains, whereas the results against GraphRAG and HippoRAG2 remain comparatively stable. Figure 3(c) further shows that the narrowing at Full reflects fewer wins rather than increased ties alone. Overall, RQ2 indicates that the largest observed gains occur under low to moderate context caps, with different degrees of narrowing at Full across the four baseline comparisons. This narrowing reflects a reduced relative advantage and does not necessarily imply that FEGR's absolute answer quality decreases under Full.

4.4 Mixed-Risk Stress Test (RQ3)

We stress each graph-based RAG system after retrieval by replacing 0%–50% of eligible candidates in its fixed clean pool with a length-matched, balanced mixture of weakly supportive, conflicting,

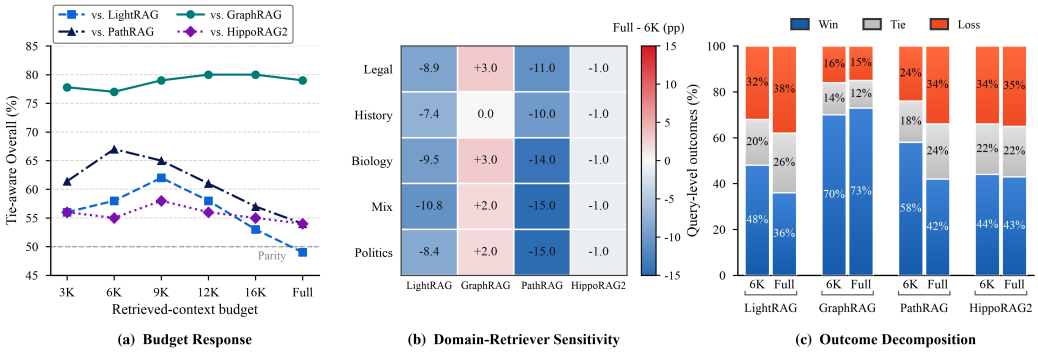


Figure 3: Budget sensitivity of FEGR against four graph-based baselines: (a) Overall win rates across budgets, (b) changes from 6K to Full, and (c) win, tie, and loss proportions at 6K and Full.

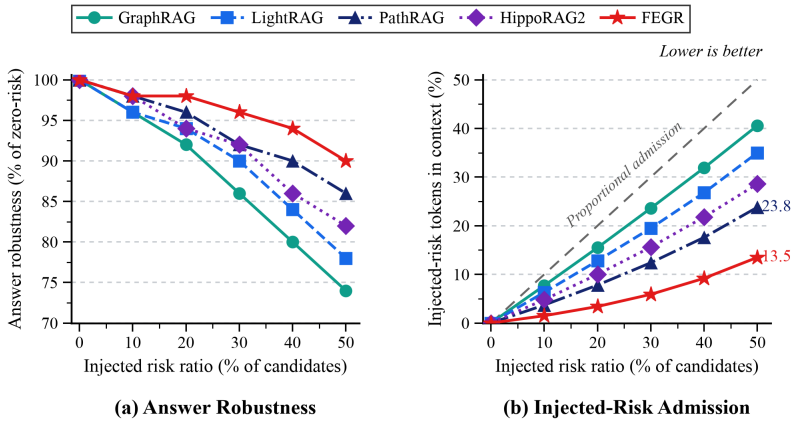


Figure 4: Controlled mixed-risk stress test: (a) answer robustness and (b) risk admission. The gray diagonal in (b) marks proportional admission.

uncertain, and redundant evidence. Length matching keeps the candidate-level injection ratio comparable to token-weighted risk admission. The query set, pool size, 6K cap, generator, and evaluation protocol remain fixed across stress levels. Because clean pools are system-specific, the experiment measures within-system degradation under a common perturbation rather than selector performance on an identical pool. Risk labels are used only for perturbation construction and diagnostic accounting; FEGR memberships may still coexist. Answer robustness measures the proportion of stressed answers judged at least as good as the corresponding zero-risk answers from the same system, while risk admission is the proportion of final-context tokens originating from injected candidates.

Figure 4(a) shows that all systems degrade as the mixed-risk ratio increases, but FEGR declines more slowly. At 50% injected risk, FEGR retains 90% answer robustness, compared with 86% for PathRAG, 82% for HippoRAG2, 78% for LightRAG, and 74% for GraphRAG. Figure 4(b) shows the corresponding ordering in risk admission: injected evidence accounts for 13.5% of the context selected by FEGR, compared with 23.8% for PathRAG, 28.6% for HippoRAG2, 35.0% for LightRAG, and 40.6% for GraphRAG. Because the perturbations are length matched, the gap between the 50% candidate-level injection ratio and the 13.5% token-level admission rate suggests that injected evidence is preferentially excluded during context construction.

All systems operate under the same 6K cap, so the observed differences arise under a matched maximum context capacity. Relative to PathRAG, the strongest baseline under both diagnostics at this stress level, FEGR retains 4.0 percentage points more answer robustness while reducing risk admission by 10.3 percentage points. Across the two panels, lower risk admission consistently coincides with higher answer robustness. This inverse ordering provides an evidence-level explanation for the answer-level pattern and supports the interpretation that FEGR's robustness advantage is associated with limiting how much injected risk reaches the generator.

4.5 COMPONENT ABLATION (RQ4)

We finally examine the principal components of FEGR using a fixed subset of 100 queries. All variants use the same retrieved candidates, evidence identifiers, 6K cap, context renderer, generator, and position-balanced evaluation protocol as full FEGR. Each membership ablation neutralizes only its target membership: μ_{rel} or μ_{sup} is set to one, while μ_{conf} or μ_{unc} is set to zero. The compensatory variant replaces the product t-norm with the arithmetic mean of M_{rel}, M_{sup}, $1 - M_{\mathrm{conf}}$, and $1 - M_{\mathrm{unc}}$. Removing anchor coverage makes the entire budget available for reliability-based selection per incremental token while retaining the reliability gate and overlap filter, whereas removing the overlap filter disables the evidence identity overlap constraint.

Table 2 shows targeted effects across the four memberships. Removing relevance or supportiveness produces the largest degradation in its corresponding evaluation dimension, while removing conflict modeling most strongly reduces Conflict Avoidance. Within the uncertainty ablation, the largest degradations appear in Logicality and Coherence. These dimension-specific effects are consistent with the intended semantic roles of the four memberships. The compensatory mean and removal of conflict modeling yield the two largest Overall degradations, a pattern consistent with the roles of noncompensatory inference and explicit conflict assessment. Removing anchor coverage most strongly affects Supportiveness, whereas disabling the overlap filter has its largest effect on Coherence. Taken together, the distribution of dimensional degradations across different ablations is consistent with complementary contributions from graded evidence modeling, noncompensatory path inference, query-anchor coverage, and evidence identity overlap control, rather than a single dominant component.

Variant	Rel.	Sup.	Conf.	Log.	Coh.	Overall
Membership ablations						
w/o μ_{rel}	**61.5**	55.8	53.5	55.0	54.5	56.0
w/o μ_{sup}	54.8	**64.5**	55.2	58.5	57.5	58.1
w/o μ_{conf}	53.5	57.0	**66.8**	60.5	58.5	59.2
w/o μ_{unc}	53.2	55.5	60.8	**63.2**	61.5	58.8
Inference and selection variants						
Compensatory Mean	56.2	59.0	61.5	60.8	59.8	**59.4**
w/o Anchor Coverage	56.0	62.0	56.8	58.2	58.8	58.4
w/o Overlap Filter	53.5	57.2	55.8	57.0	**64.0**	57.5

Table 2: Position-balanced pairwise tie-aware win rates (%) of full FEGR against its ablated variants under a 6K retrieved-context cap. Values above 50 favor full FEGR; bold values indicate the largest degradation in each dimension.

To further isolate selection effects from upstream retrieval and path construction, we compare FEGR with four alternative selection strategies on identical candidate paths (Appendix C). FEGR achieves Overall win rates ranging from 58.4% to 62.6%, with above-parity Supportiveness and Conflict Avoidance in every comparison.

5 CONCLUSION

This paper introduced FEGR, a graph-based RAG system for reliable evidence admission under limited context budgets. FEGR models graded relevance, supportiveness, conflict, and uncertainty, performs noncompensatory path-level reliability inference, and selects a reliable and complementary evidence subset within the available budget. Across five domains, FEGR obtains above-parity Overall win-rate point estimates against all five baselines under a common 6K retrieved-context cap, with consistent gains in Supportiveness and Conflict Avoidance. Budget, stress, and ablation analyses further indicate larger observed advantages under low-to-moderate context budgets, lower admission of mixed evidence risks, and complementary contributions from its principal components. These findings highlight the importance of jointly considering evidence reliability and complementarity when constructing retrieval contexts under limited budgets.

AI USE STATEMENT

Generative AI tools were used for limited English-language editing to improve the readability and clarity of the manuscript. All AI-assisted text was reviewed and approved by the authors, who take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The paper specifies the evidence representation, fuzzy membership construction, path-level inference, and budgeted selection procedure. The appendix provides the complete algorithm, fixed method settings, corpus and query counts, baseline configurations, prompts, output schemas, and budget-accounting protocol. Together, these materials document the assumptions and implementation choices needed to reproduce the reported comparisons and controlled analyses.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, volume 2024, pp. 9112–9141, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/25f7be9694d7b32d5cc670927b8091e1-Paper-Conference.pdf.

Yukun Cao, Zirui Xu, Dongyang Li, Zhihao Guo, Luobin Huang, and Lisheng Wang. DCTR: Dual-constraint subgraph optimization for Knowledge Graph-based Retrieval-Augmented Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(36):30156–30164, 2026. doi: 10.1609/aaai.v40i36.40265. URL https://ojs.aaai.org/index.php/AAAI/article/view/40265.

Boyu Chen, Zirui Guo, Zidan Yang, Yuluo Chen, Junze Chen, Zhenghao Liu, Chuan Shi, and Cheng Yang. PathRAG: Pruning graph-based retrieval augmented generation with relational paths. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(36):30183–30191, 2026a. doi: 10.1609/aaai.v40i36.40268. URL https://ojs.aaai.org/index.php/AAAI/article/view/40268.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335, Bangkok, Thailand, 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.137.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking Large Language Models in Retrieval-Augmented Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17754–17762, 2024b. doi: 10.1609/aaai.v38i16.29728. URL https://ojs.aaai.org/index.php/AAAI/article/view/29728.

Shengyuan Chen, Chuang Zhou, Zheng Yuan, Qinggang Zhang, Zeyang Cui, Hao Chen, Yilin Xiao, Jiannong Cao, and Xiao Huang. You don't need pre-built graphs for RAG: Retrieval augmented generation with adaptive reasoning structures. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(36):30270–30278, 2026b. doi: 10.1609/aaai.v40i36.40278. URL https://ojs.aaai.org/index.php/AAAI/article/view/40278.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. The power of noise: Redefining retrieval for RAG systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 719–729, New York, NY, USA, 2024. Association for Computing Machinery. doi: 10.1145/3626772.3657834.

Didier Dubois and Henri Prade. A review of fuzzy set aggregation connectives. *Information Sciences*, 36(1-2):85–121, 1985. doi: 10.1016/0020-0255(85)90027-1.

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

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. LightRAG: Simple and fast retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 10746–10761, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.568.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 59532–59569. Curran Associates, Inc., 2024. doi: 10.52202/079017-1902.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From RAG to memory: Non-parametric continual learning for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 21497–21515. PMLR, 2025. URL https://proceedings.mlr.press/v267/gutierrez25a.html.

Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh V. Chawla, Thomas Laurent, Yann LeCun, Xavier Bresson, and Bryan Hooi. G-Retriever: Retrieval-augmented generation for textual graph understanding and question answering. In *Advances in Neural Information Processing Systems*, volume 37, pp. 132876–132907. Curran Associates, Inc., 2024. doi: 10.52202/079017-4224. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/efaf1c9726648c8ba363a5c927440529-Abstract-Conference.html.

Manzong Huang, Chenyang Bu, Yi He, Xingrui Zhuo, and Xindong Wu. Relink: Constructing query-driven evidence graph on-the-fly for GraphRAG. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(37):31202–31210, 2026. doi: 10.1609/aaai.v40i37.40382. URL https://ojs.aaai.org/index.php/AAAI/article/view/40382.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1658–1677, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.91.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient context: A new lens on retrieval augmented generation systems. In *International Conference on Learning Representations*, volume 2025, pp. 20310–20334, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/33dffa2e3d2ab74a783d1a8c292f66d9-Paper-Conference.pdf.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 6769–6781, 2020. doi: 10.18653/v1/2020.emnlp-main.550.

Erich Peter Klement, Radko Mesiar, and Endre Pap. *Triangular Norms*, volume 8 of *Trends in Logic*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000. ISBN 978-0-7923-6416-0. doi: 10.1007/978-94-015-9540-7.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Shilong Li, Yancheng He, Hangyu Guo, Xingyuan Bu, Ge Bai, Jie Liu, Jiaheng Liu, Xingwei Qu, Yangguang Li, Wanli Ouyang, Wenbo Su, and Bo Zheng. GraphReader: Building graph-based agent to enhance long-context abilities of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 12758–12786, Miami, Florida, USA, 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.746. URL https://aclanthology.org/2024.findings-emnlp.746/.

Zongjie Li, Chaozheng Wang, Pingchuan Ma, Daoyuan Wu, Shuai Wang, Cuiyun Gao, and Yang Liu. Split and merge: Aligning position biases in LLM-based evaluators. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11084–11108, Miami, Florida, USA, 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.621.

Haochen Liu, Song Wang, Yaochen Zhu, Yushun Dong, and Jundong Li. Knowledge graph-enhanced large language models via path selection. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6311–6321, Bangkok, Thailand, 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.376. URL https://aclanthology.org/2024.findings-acl.376/.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024b. doi: 10.1162/tacl_a_00638.

Shuyi Liu, Yu-Ming Shang, and Xi Zhang. TruthfulRAG: Resolving factual-level conflicts in retrieval-augmented generation with knowledge graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(38):32168–32176, 2026. doi: 10.1609/aaai.v40i38.40489. URL https://ojs.aaai.org/index.php/AAAI/article/view/40489.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153.

Linhao Luo, Yuan-Fang Li, Reza Haffari, and Shirui Pan. Reasoning on graphs: Faithful and interpretable large language model reasoning. In *International Conference on Learning Representations*, volume 2024, pp. 14400–14423, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/3e2aeb66481dd63a32421bf032b70384-Paper-Conference.pdf.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10862–10878, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.585.

OpenAI. GPT-4o mini model. https://developers.openai.com/api/docs/models/gpt-4o-mini, 2024. Snapshot: gpt-4o-mini-2024-07-18; Accessed: 2026-07-29.

Seong-Il Park and Jay-Yoon Lee. Toward robust RALMs: Revealing the impact of imperfect retrieval on retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 12:1686–1702, 2024. doi: 10.1162/tacl_a_00724.

Hongjin Qian, Zheng Liu, Peitian Zhang, Kelong Mao, Defu Lian, Zhicheng Dou, and Tiejun Huang. MemoRAG: Boosting long context processing with global memory-enhanced retrieval augmentation. In *Proceedings of the ACM on Web Conference 2025*, pp. 2366–2377, New York, NY, USA, 2025. Association for Computing Machinery. doi: 10.1145/3696410.3714805.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher Manning. RAPTOR: Recursive abstractive processing for tree-organized retrieval. In

International Conference on Learning Representations, volume 2024, pp. 32628–32649, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/8a2acd174940dbca361a6398a4f9df91-Paper-Conference.pdf.

Xiaoyu Shen, Rexhina Blloshmi, Dawei Zhu, Jiahuan Pei, and Wei Zhang. Assessing "implicit" retrieval robustness of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8988–9003, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.507.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 292–314, Mumbai, India, 2025. Asian Federation of Natural Language Processing and Association for Computational Linguistics. doi: 10.18653/v1/2025.ijcnlp-long.18.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, Bangkok, Thailand, 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.511.

Yu Wang, Nedim Lipka, Ryan A. Rossi, Alexa Siu, Ruiyi Zhang, and Tyler Derr. Knowledge graph prompting for multi-document question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19206–19214, 2024b. doi: 10.1609/aaai.v38i17.29889. URL https://ojs.aaai.org/index.php/AAAI/article/view/29889.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. RankRAG: Unifying context ranking with retrieval-augmented generation in LLMs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 121156–121184, 2024. doi: 10.52202/079017-3850. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/db93ccb6cf392f352570dd5af0a223d3-Abstract.html.

Lotfi A. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338–353, 1965. doi: 10.1016/S0019-9958(65)90241-X.

Qinggang Zhang, Zhishang Xiang, Yilin Xiao, Le Wang, Junhui Li, Xinrun Wang, and Jinsong Su. FaithfulRAG: Fact-level conflict modeling for context-faithful retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21863–21882, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1062. URL https://aclanthology.org/2025.acl-long.1062/.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623. Curran Associates, Inc., 2023. doi: 10.52202/075280-2020.

Xiangrong Zhu, Yuexiang Xie, Yi Liu, Yaliang Li, and Wei Hu. 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)*, pp. 8912–8924, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.449.

A Additional Method Details

A.1 Operational Membership Definition

For each query q and evidence unit u, FEGR computes relevance, supportiveness, conflict, and uncertainty memberships in $[0, 1]$. Table 3 summarizes their operational signal families.

Membership	Operational Signal Families	Interpretation
μ_{rel}	Upstream retrieval confidence, typed query-anchor correspondence, and entity or relation correspondence	Degree to which the unit is semantically and structurally related to the query
μ_{sup}	Required-information coverage, predicate and relation coverage, answer-bearing content, and assertive completeness	Degree to which the unit contributes evidence needed for an answer
μ_{conf}	Scope-aligned numerical, temporal, entity–value, correction, negation, and status incompatibilities	Degree to which the unit conflicts with locally comparable evidence
μ_{unc}	Hedging, ambiguous attribution, weak grounding, incomplete scope, low confidence, and content-form risk	Degree to which the unit is evidentially indeterminate or weakly grounded

Table 3: Operational signal families used to instantiate the four fuzzy memberships. The signal mappings and weights are fixed across domains and experimental conditions.

Signal Aggregation and Normalization. Different retrieval channels expose different subsets of retrieval and structural metadata. Relevance, supportiveness, and uncertainty follow the weighted aggregation rule in Eq. (3). For each membership $x \in \{\mathrm{rel}, \mathrm{sup}, \mathrm{unc}\}$, $\mathcal{I}_x(u)$ contains the available normalized signals $s_{x,j}(u, q)$ with fixed weights $w_{x,j}$. Missing signals are omitted, the remaining weights are renormalized, and the result is clipped to $[0, 1]$.

Relevance and Supportiveness. Relevance combines normalized retrieval confidence with typed anchor, entity, and relation correspondence. Retrieval scores are normalized within the query-specific candidate pool, with a deterministic fallback when all scores are equal. Supportiveness measures whether a unit contributes information needed to answer the query. Its signals include non-generic query-anchor coverage, predicate and relation cues, query-associated entities and relations, answer-bearing content, and statement completeness. Tables of contents, indexes, bibliographies, reference lists, and incomplete fragments receive lower scores. Marginal anchor coverage is handled later during set construction.

Conflict and Uncertainty. Conflict is computed over locally comparable claims, as defined in Section 3.3. For each rule $k \in \mathcal{K}$, the weighted rule response is

$$d_k(u, v) = \lambda_k c_k(u, v), \tag{10}$$

where λ_k is the fixed rule importance and $c_k(u, v)$ denotes the scope-aligned incompatibility score, including applicable source-provenance adjustments. The rules cover numerical and temporal disagreement, entity–value incompatibility, explicit correction, local negation mismatch, and mutually exclusive states. Claims are compared only when their subject, factual attribute, and applicable scope align. Same-source comparisons are attenuated to account for local clarification and progressive refinement. The conflict membership is zero when no aligned conflict signal is detected.

Uncertainty is estimated from hedging, ambiguous attribution, weak grounding, low retrieval confidence, incomplete query scope, and fragmentary or structurally weak content.

A.2 Evidence Representation and Fixed Settings

The graph retriever returns passages, graph-associated evidence, and the structural links exposed by its index. FEGR converts these outputs into a query-conditioned evidence graph $\mathcal{G}_q^E = (\mathcal{U}_q, \mathcal{A}_q)$, with evidence units in $\mathcal{U}_q$ and retained links in $\mathcal{A}_q$.

Each evidence unit records its identity, type, text, provenance, associated entities and relations, retrieval metadata, matched anchors, token cost, and four memberships:

$$u = \langle i_u, t_u, x_u, \pi_u, \mathcal{E}_u, \mathcal{R}_u, r_u, \mathcal{H}_u, \boldsymbol{\mu}_u, c_u \rangle. \tag{11}$$

Object	Core Fields	Role in FEGR
EvidenceUnit	Stable identity and type; evidence text; provenance; graph and retrieval metadata; anchors; memberships; token cost	Atomic object for evidence scoring, path construction, deduplication, and rendering
EvidenceLink	Source and target units; link type; upstream relation or shared graph identity; provenance	Preserves the local structural organization exposed by graph retrieval
EvidencePath	Ordered units and links; unit-identity set; aggregated evidence conditions; reliability; rendered cost	Candidate evidence chain considered during budgeted selection
SelectedContext	Selected paths; deduplicated evidence records; rendered graph, passage, and reference sections	Budget-feasible evidence context supplied to the answer generator

Table 4: Core data structures used by FEGR between upstream graph retrieval and answer generation.

Component	Setting	Fixed Value
Query anchors	Maximum retained anchors; minimum anchor length	24; 3 characters
Anchor weighting	Quoted title or named entity; relation cue; temporal or numeric expression; content term; generic lexical cue	1.00; 0.80; 0.70; 0.50; 0.15
Path construction	Maximum path length; seed units; neighbors per unit; beam width; candidate paths	3; 24; 8; 12; 128
Candidate policy	Singleton paths; semantic near-duplicate merging	Included; disabled
Reliability gate	Descending threshold schedule $\boldsymbol{\tau}$; minimum admissible pool $K_{\min}$	$(0.10, 0.075, 0.05, 0.025, 0)$; 8
Budgeted admission	Coverage-first budget fraction γ; maximum path overlap δ	0.35; 0.80
Representation	Embedding model; context tokenizer	`BAAI/bge-m3`; `o200k_base`

Table 5: Fixed method settings used in all experiments.

Here, i_u, t_u, and x_u denote identity, type, and text; π_u is provenance; $\mathcal{E}_u$ and $\mathcal{R}_u$ are associated entity and relation identifiers; r_u stores retrieval metadata; $\mathcal{H}_u$ contains matched anchors; $\boldsymbol{\mu}_u$ contains the four memberships; and c_u is token cost. Evidence links store endpoint units, link type, and upstream provenance. Relation-supported links connect evidence associated with entities joined upstream, while shared-relation and shared-entity links preserve common graph associations. Paths retain ordered units and links together with aggregated conditions, reliability, and rendered cost. The selected context consists of admitted paths and identity-deduplicated entity, relation, passage, and source-reference records. Table 4 gives the corresponding data structures, and Section A.5 shows them in one query-level trace.

Query anchors are extracted deterministically from quoted titles, named entities, relation or predicate cues, temporal and numerical expressions, and non-generic content terms. Generic instructional or discourse tokens may be retained for diagnostics but do not count toward core coverage. Stable source provenance defines evidence identity, so exact duplicates are removed without merging similar records from different sources. Table 5 lists the fixed settings used for path construction, reliability gating, and budgeted admission; lexical cues and detector weights are also fixed throughout.

A.3 Path Construction and Budgeted Selection

Candidate paths are built by bounded graph expansion from query-relevant seeds. Units cannot repeat within a path, reverse-equivalent undirected paths are canonicalized, singleton units are retained, and duplicates are removed using ordered unit and source identities. After scoring, FEGR applies the descending threshold schedule $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_L)$ and uses the first level that retains at least $K_{\min}$ paths, falling back to τ_L when necessary. For a current selection $\mathcal{S}$, a path's incremental cost is the increase in fully rendered context length after identity deduplication and serialization overhead.

Algorithm 1 Fuzzy Evidence-Guided Path Selection

Require: Query q, evidence graph $\mathcal{G}_q^E$, budget B
Require: Path settings Θ_{path}, threshold schedule $\boldsymbol{\tau}$
Require: Minimum pool size $K_{\min}$, coverage fraction γ, overlap limit δ
Ensure: Rendered evidence context C_B

$\mathcal{H}(q) \leftarrow \text{ExtractAnchors}(q)$
$(\mathcal{U}_q, \mathcal{A}_q) \leftarrow \mathcal{G}_q^E$
$\mathcal{P}_q \leftarrow \text{Paths}(\mathcal{G}_q^E; \Theta_{\text{path}})$
for all $u \in \mathcal{U}_q$ **do**
    Compute $\mu_{\text{rel}}(u, q)$ and $\mu_{\text{sup}}(u, q)$
    Compute $\mu_{\text{conf}}(u, q)$ and $\mu_{\text{unc}}(u, q)$
end for
for all $p \in \mathcal{P}_q$ **do**
    Aggregate evidence conditions and compute $R(p, q)$
end for
$\tau^* \leftarrow \text{Gate}(\mathcal{P}_q, \boldsymbol{\tau}, K_{\min})$
$\mathcal{P}_q^* \leftarrow \{p \in \mathcal{P}_q \mid R(p, q) \geq \tau^*\}$
$\mathcal{S} \leftarrow \emptyset$
while $\text{Cost}(\mathcal{S}) < \gamma B$ **do**
    $\mathcal{F} \leftarrow \text{FeasiblePaths}(\mathcal{P}_q^*, \mathcal{S})$
    if $\mathcal{F} = \emptyset$ **then**
        break
    end if
    $p^* \leftarrow \text{BestCoverage}(\mathcal{F}, \mathcal{S}, q)$
    if p^* adds no uncovered anchor mass **then**
        break
    end if
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{p^*\}$
end while
while $\text{Cost}(\mathcal{S}) < B$ **do**
    $\mathcal{F} \leftarrow \text{FeasiblePaths}(\mathcal{P}_q^*, \mathcal{S})$
    if $\mathcal{F} = \emptyset$ **then**
        break
    end if
    $p^* \leftarrow \text{BestReliability}(\mathcal{F}, \mathcal{S}, q)$
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{p^*\}$
end while
$C_B \leftarrow \text{Render}(\mathcal{S})$
assert $\text{Tok}(C_B) \leq B$
return C_B

The first phase reserves at most γB tokens and ranks feasible paths by newly covered type-weighted anchor mass per incremental token. It stops when the reserved capacity is reached or no path adds core-anchor coverage. The second phase allocates the remaining budget by reliability per incremental token. Cost, coverage, and overlap are recomputed after each admission. A path must pass the active gate, add positive cost, fit the residual budget, and remain below the overlap limit δ; selection stops when no feasible path remains.

Redundancy Control. Redundancy is measured by Jaccard overlap between the stable evidence identities in two paths. Candidates above the overlap limit are removed during both selection phases.

A.4 Algorithm and Context Serialization

Algorithm 1 summarizes the selection procedure. Its symbolic parameters correspond to the fixed settings in Table 5.

The gate operator implements the threshold schedule above. FeasiblePaths removes paths already selected, below threshold, over budget, above the overlap limit, or with non-positive incremental cost; the two ranking operators implement coverage-per-token and reliability-per-token admission.

Selected paths are ordered by the earliest upstream rank of their units, with stable path identity as the tie-breaker. Intra-path order is preserved and duplicate units are emitted once. The renderer

Query and typed anchors. The query yields ten typed anchors. The principal anchors are `No Trees in the Street` [quoted title, 1.00], `No Trees` [named entity, 1.00], and the content terms `critical`, `reception`, and `film` [0.50 each].

Representative evidence units.

Unit	Evidence role and matched anchors	Membership tuple
u_1	The retrieved passage contains the film description and the reviews from *TV Guide*, *Time Out*, and *Variety*. Anchors: title, critical, reception, film	$(0.383,\ 0.794,\ 0.650,\ 0.249)$
u_2	The entity record identifies Sylvia Syms and her role as Hetty, linking a praised performance to the retrieved film graph. Anchors: title, film	$(0.374,\ 0.622,\ 0.000,\ 0.339)$
u_3	The entity record identifies Tommy as the film's impoverished teenage protagonist. Anchors: title	$(0.321,\ 0.596,\ 0.000,\ 0.349)$
u_4	The relation record states that Wilkie influences Tommy's descent into crime. Anchors: title	$(0.328,\ 0.554,\ 0.000,\ 0.345)$
u_5	The entity record identifies Wilkie as the local racketeer who controls the slum community. Anchors: title	$(0.321,\ 0.596,\ 0.000,\ 0.349)$

Gate and two-stage admission. At the initial reliability threshold $\tau^* = 0.10$, 46 paths remain admissible, exceeding the minimum admissible-pool size of eight. Coverage-first admission selects three paths and consumes 1,385 tokens. Reliability fill admits 19 additional paths, including the multi-unit path $u_3 \rightarrow u_4 \rightarrow u_5$. After identity deduplication and canonical rendering, the selected context contains 22 paths and occupies 5,213 of 6,000 tokens.

Generated answer excerpt. The critical reception of *No Trees in the Street* is characterized by mixed assessments. *TV Guide* criticizes the artificial plot, dialogue, and stereotyped characterization, while recognizing Sylvia Syms's sensitive performance and the documentary-style camerawork. *Time Out* regards the film as an unremarkable kitchen-sink chronicle, whereas *Variety* praises Ted Willis's sympathy for working-class problems and Syms's performance. Overall, the reviews value the naturalistic depiction of slum life but criticize the plot and character development [C1].
Uncertainty. None.

Figure 5: A recorded end-to-end execution of FEGR for a Mix-domain query. The overview shows the principal anchors, representative evidence units and memberships, reliability gating, two-stage path admission, and the evidence-grounded answer.

outputs four sections—*Knowledge Graph Data (Entity)*, *Knowledge Graph Data (Relationship)*, *Document Chunks*, and *Reference Document List*—and includes only entity and relation records associated with selected evidence; relations are serialized with resolvable endpoints. The completed serialization is recounted with the same tokenizer before it is sent to the answer generator.

A.5 Illustrative End-to-End Case

The following Mix-domain query provides a compact trace of the full FEGR pipeline:

> *How does the Mix document characterize the critical reception of the film "No Trees in the Street"?*

We reran this query with trace logging under the same 6K budget, `o200k_base` tokenizer, gate schedule, and two-stage selector. The trace is used only for illustration and is excluded from aggregate results. Figure 5 shows the main anchors, evidence units, selection stages, and answer; Table 6 lists representative path decisions.

Coverage-first admits p_1, the passage containing the answer-bearing reviews. Its supportiveness is high because it directly answers the query, while its conflict score reflects the mixture of favorable and unfavorable assessments. Reliability fill then adds entity and relation records from the film graph, including the three-unit Tommy–influence–Wilkie path p_3. The final answer is grounded mainly in passage `C1`; the additional records retain the local graph structure around that passage.

The serialized context contains 5,213 tokens under `o200k_base`. Passage `C1` supports the answer's summary of a mixed reception: criticism of plotting and characterization appears alongside praise for performance, camerawork, and social observation. The remaining selected records add graph context without displacing that passage under the 6K cap.

B Additional Experimental Details

Unless otherwise stated, all systems use the same corpora, query sets, answer generator, generation prompt, and pairwise evaluation procedure, using the indexing and initial retrieval configurations described below.

Path	Units	ΔTok	Admission stage	Decision and evidential role
p_1	Critical-reception passage	1,232	Coverage-first	Selected; supplies the answer-bearing reviews and the main query coverage
p_2	Sylvia Syms	346	Reliability fill	Selected; preserves the performer associated with a favorable review
p_3	Tommy → influence relation → Wilkie	100	Reliability fill	Selected; retains a three-unit entity–relation–entity path in the film graph
p_4	Hetty–Tommy sibling relation	85	Reliability fill	Selected; adds complementary relational context
p_5	Low-reliability passage candidate	—	Reliability gate	Rejected below the active threshold

Table 6: Representative incremental costs and admission decisions in the recorded case. The table shows one coverage-first admission, three reliability-fill admissions, and one gate rejection.

Domain	Candidate Bank	RQ1	RQ2–RQ4 / Control
Legal	50	40	20
History	50	40	20
Biology	50	40	20
Mix	50	40	20
Politics	50	40	20
Total	250	200	100

Table 7: Domain-specific query counts used in the experiments.

B.1 Corpora, Queries, and Baselines

We use the Legal, History, Biology, Mix, and Politics corpora from UltraDomain. For each domain, a 50-query candidate bank covers direct retrieval, multi-hop reasoning, evidence synthesis, conflict and uncertainty, redundancy, and cross-document contamination. We retain queries whose required evidence is present and whose information need is unambiguous, revising or removing empty-retrieval, unanswerable, and corpus-independent cases. RQ1 uses 40 validated queries per domain (200 total); RQ2–RQ4 and the shared-candidate control experiment in Appendix C use the same fixed stratified subset of 20 per domain (100 total), as listed in Table 7.

FEGR. FEGR is evaluated as a standalone graph-based RAG system. It constructs a graph index from the input corpus, performs query-conditioned graph retrieval, and organizes retrieved evidence into candidate paths. It then applies fuzzy reliability inference and budgeted evidence selection to construct the final context for answer generation.

NaiveRAG. NaiveRAG retrieves and ranks independent passages without graph-specific structure. It appears in RQ1 but not in RQ2 or RQ3, which compare graph-based pipelines.

GraphRAG. GraphRAG uses its native index and local retrieval procedure to return entities, relationships, community reports, and source passages.

LightRAG. LightRAG combines local entity-centered retrieval with broader graph retrieval and constructs its context using its native retrieval and context-organization procedures.

PathRAG. PathRAG retrieves relational paths and applies its own redundancy-pruning procedure before generation.

HippoRAG2. HippoRAG2 retrieves connected evidence through its associative-memory architecture and supplies its native context to the shared generator.

Criterion	Operational Definition
Relevance	The answer directly addresses the information requested by the query using pertinent retrieved evidence.
Supportiveness	The principal factual claims in the answer are traceable to the evidence context supplied to the corresponding system.
Conflict Avoidance	The answer identifies or appropriately handles incompatible evidence and avoids presenting disputed claims as settled.
Logicality	The conclusions follow consistently from the supplied evidence without unsupported inferential steps or internal contradiction.
Coherence	The answer is clearly organized, readable, and internally consistent.

Table 8: Operational definitions of the five pairwise evaluation dimensions.

Each system is evaluated as a complete pipeline using its specified indexing, retrieval, and context-construction procedures. All systems share the same corpora, queries, rendered-evidence cap, answer generator, decoding settings, and anonymous position-balanced evaluation protocol. RQ1 therefore compares complete RAG systems rather than isolated retrieval or selection components operating on a common candidate pool.

B.2 Generation, Evaluation, and Budget Accounting

All answers use the frozen `gpt-4o-mini-2024-07-18` generator and one decoding configuration. Each method supplies its rendered context, the canonical query, and the common answer policy; citations use the identifiers exposed in that context. The pairwise judge receives the query and both context–answer pairs under anonymous A/B labels. Ties are allowed, and each comparison is run in both presentation orders. Table 8 defines the five criteria.

For dimension d, query q, and one presentation order, the score assigned to FEGR is

$$z_{q,d} = \begin{cases} 1, & \text{FEGR wins,} \\ 0.5, & \text{the result is a tie,} \\ 0, & \text{FEGR loses.} \end{cases} \tag{12}$$

Let $z^A_{q,d}$ and $z^B_{q,d}$ denote the scores when FEGR is presented as systems A and B. The position-balanced query-level score is

$$\bar{z}_{q,d} = \frac{z^A_{q,d} + z^B_{q,d}}{2}. \tag{13}$$

For query set $\mathcal{Q}$, the tie-aware win rate on dimension d is

$$\mathrm{WR}_d = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \bar{z}_{q,d}, \tag{14}$$

and the Overall score is the unweighted mean over the five dimensions,

$$\mathrm{Overall} = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \mathrm{WR}_d \,. \tag{15}$$

The context budget counts only rendered evidence; it excludes the query, system instruction, answer policy, and reserved output capacity. It is enforced after method-specific rendering, deduplication, and source-reference construction, with whole records admitted. A context may stop below the cap when no remaining record fits. The Full condition includes all method-specific evidence that fits after reserving space for non-evidence input and output. Table 9 lists the shared settings.

Experiment	Setting	Value
Answer generation	Model; temperature; top-p; output limit	`gpt-4o-mini-2024-07-18`; 0; 1; 512 tokens
Pairwise evaluation	Judge model; temperature; presentation orders	`gpt-4o-mini-2024-07-18`; 0; A/B and B/A
Evaluation criteria	Dimensions	Relevance, Supportiveness, Conflict Avoidance, Logicality, and Coherence
RQ1	Queries; baselines; context cap	200; five; 6K tokens
RQ2	Queries; graph-based baselines; context settings	100; four; 3K, 6K, 9K, 12K, 16K, and Full
RQ3	Queries; injected-risk levels; context cap	100; 0–50% in 10% increments; 6K tokens
RQ4	Queries; ablated variants; context cap	100; seven; 6K tokens
Shared-candidate control	Queries; comparators; context cap	100; four; 6K tokens

Table 9: Shared generation, evaluation, and experimental settings. Complete-system comparisons retain method-specific indexing and initial retrieval, whereas the shared-candidate control uses fixed upstream outputs from FEGR.

Main comparison (RQ1). RQ1 compares FEGR with the five baselines on 200 queries under a 6K rendered-evidence cap. Each query–baseline pair is judged in both orders, yielding 2,000 evaluation instances. We average the two positions at the query level before aggregating by criterion, domain, and baseline.

B.3 Budget, Stress-Test, and Ablation Protocols

Budget sensitivity (RQ2). RQ2 compares FEGR with GraphRAG, LightRAG, PathRAG, and HippoRAG2 on the fixed 100-query subset and the budget settings in Table 9. For a given method and query, the retrieval result, evidence identities, ordering, renderer, and generation protocol remain fixed; only the context capacity changes. Contexts are not padded when the retrieved pool is exhausted. We report results by baseline and domain, with win, tie, and loss proportions at 6K and Full.

Controlled mixed-risk stress test (RQ3). RQ3 evaluates FEGR and the four graph-based baselines on the same 100 queries under the 6K cap. Perturbations are applied after retrieval and before each method selects or packs its context. Because each system keeps its own clean candidate pool, the comparison measures within-system degradation under a common injection protocol.

For method m and query q, let $\mathcal{C}^0_{m,q} = \{c_1, \ldots, c_N\}$ be the clean retrieved pool. At risk ratio ρ, the experiment replaces

$$K_\rho = \lfloor \rho N \rfloor \tag{16}$$

eligible candidates while preserving pool size and candidate positions. Replacement sets are nested across risk levels, balanced across the four categories, and length matched to the original candidates.

Weakly supportive evidence. Weakly supportive evidence remains topically related to the query while omitting the factual relation, constraint, comparison side, or answer-bearing content required to resolve it.

Conflicting evidence. Conflicting evidence preserves the relevant subject, factual attribute, and applicable scope while introducing an incompatible value, date, entity assignment, state, or conclusion. Differences in topic, temporal scope, or granularity are handled separately.

Uncertain evidence. Uncertain evidence remains relevant while weakening evidential determinacy through hedging, ambiguous attribution, incomplete scope, indirect reporting, or weak grounding.

Redundant evidence. Redundant evidence restates an answer-bearing contribution already present in the clean pool without adding a new fact. It receives a distinct evidence identity and may be realized as a controlled paraphrase, repeated summary, or locally overlapping passage.

Risk labels are used only to construct and analyze the perturbations; they are not shown to any selector, generator, or judge.

For answer-level analysis, each stressed answer is compared with the same method's clean answer in both presentation orders. Let $\bar{z}^{\rho}_{m,q}$ be their position-balanced Overall score. Answer robustness is

$$\mathrm{Robust}_m(\rho) = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathbb{I}\big[\bar{z}^{\rho}_{m,q} \geq 0.5\big]\,. \tag{17}$$

A tie with the clean answer counts as preserved performance. Let $\mathcal{E}^{\rho}_{m,q}$ denote the evidence records rendered into the stressed context and let $y(e)$ indicate whether record e originates from an injected candidate. Token-weighted risk admission is

$$\mathrm{RiskAdm}_{m,q}(\rho) = \frac{\sum_{e \in \mathcal{E}^{\rho}_{m,q}} \mathrm{Tok}(e) y(e)}{\sum_{e \in \mathcal{E}^{\rho}_{m,q}} \mathrm{Tok}(e)}. \tag{18}$$

We average this quantity over queries; category-specific admission restricts $y(e)$ to one risk type.

Component ablation (RQ4). RQ4 compares full FEGR with seven ablations on the fixed 100-query subset. All variants share the retrieved candidate graph, path construction, 6K cap, renderer, generator, and position-balanced evaluation. The four membership ablations neutralize one condition before path aggregation:

$$\begin{aligned} \text{w/o Relevance:} &\quad \mu_{\mathrm{rel}}(u,q) = 1, \\ \text{w/o Supportiveness:} &\quad \mu_{\mathrm{sup}}(u,q) = 1, \\ \text{w/o Conflict:} &\quad \mu_{\mathrm{conf}}(u,q) = 0, \\ \text{w/o Uncertainty:} &\quad \mu_{\mathrm{unc}}(u,q) = 0. \end{aligned} \tag{19}$$

The compensatory variant replaces the noncompensatory reliability conjunction with the arithmetic mean

$$R_{\mathrm{mean}}(p,q) = \frac{1}{4}\Big[M_{\mathrm{rel}}(p,q) + M_{\mathrm{sup}}(p,q) \quad + \big(1 - M_{\mathrm{conf}}(p,q)\big) + \big(1 - M_{\mathrm{unc}}(p,q)\big)\Big]. \tag{20}$$

The remaining variants remove coverage-first admission or the path-overlap constraint. Without coverage-first admission, the full budget is allocated by reliability per incremental token while retaining the gate and overlap control. Without redundancy control, exact identity deduplication remains active during rendering. Every variant regenerates its context and answer and is compared with the full model in both presentation orders.

C Controlled Evidence Selection under Shared Candidates

We conduct a controlled comparison to assess evidence scoring and selection under fixed upstream conditions. Unlike the complete-system comparisons in RQ1, all strategies operate on an identical candidate path pool produced by the FEGR pipeline.

Shared candidate protocol. We use the same fixed stratified subset of 100 queries as RQ4, comprising 20 queries from each of the five domains. For each query, we freeze the upstream retrieval output, query-conditioned evidence graph, evidence units, stable evidence identities, structural links, and candidate paths produced by FEGR. Every selection strategy receives this identical candidate pool without additional retrieval or path construction.

All strategies share the same 6K rendered-evidence cap, `o200k_base` tokenizer, incremental token accounting, identity-deduplicated context renderer, generator, decoding configuration, and evaluation protocol. Only the evidence scoring and selection strategy varies.

Comparison strategies. Let $\mathcal{P}_q$ denote the shared candidate path pool, $\mathcal{S}$ the currently selected paths, and $\Delta T(p \mid \mathcal{S})$ the incremental rendered token cost of admitting path p. We consider four alternatives.

Relevance Greedy selects feasible paths using

$$g_{\mathrm{RG}}(p \mid \mathcal{S}, q) = \frac{M_{\mathrm{rel}}(p, q)}{\Delta T(p \mid \mathcal{S})}.$$

It uses FEGR's path relevance scores without reliability gating or the additional path-overlap constraint.

MMR balances path relevance against semantic redundancy:

$$g_{\mathrm{MMR}}(p \mid \mathcal{S}, q) = 0.5 M_{\mathrm{rel}}(p, q) - 0.5 \max_{s \in \mathcal{S}} \mathrm{sim}(p, s).$$

Path representations are computed by applying the fixed BAAI/bge-m3 model to the ordered evidence text of each path. Similarity is the cosine similarity clipped to $[0, 1]$, and the redundancy term is zero when $\mathcal{S}$ is empty. MMR uses neither FEGR's reliability gate nor its hard path-overlap constraint. The relevance–diversity coefficient is fixed before evaluation.

Reliability Greedy retains FEGR's path reliability scores, descending reliability gate, and path-overlap constraint, but removes coverage-first admission. It allocates the entire budget using

$$g_{\mathrm{RGreedy}}(p \mid \mathcal{S}, q) = \frac{R(p, q)}{\Delta T(p \mid \mathcal{S})}.$$

Joint Greedy retains FEGR's reliability gate and path-overlap constraint, but replaces its two-stage admission procedure with a single objective:

$$g_{\mathrm{Joint}}(p \mid \mathcal{S}, q) = \frac{0.5 \widehat{R}(p, q) + 0.5 \widehat{\Delta A}(p \mid \mathcal{S}, q)}{\Delta T(p \mid \mathcal{S})}.$$

Here, $\widehat{R}(p, q)$ normalizes reliability by the maximum score in the gated candidate pool, and $\widehat{\Delta A}$ is the newly covered core-anchor mass divided by the query's total core-anchor mass. A normalized term is set to zero if its denominator is zero. Both coefficients are fixed at 0.5 before evaluation.

All methods enforce the same 6K budget, positive incremental token cost, identity deduplication, context serialization, and deterministic tie breaking. The two reliability-based alternatives additionally retain FEGR's reliability gate and path-overlap constraint. No method modifies the frozen candidate pool. Following the evaluation protocol in Appendix B.2, we report position-balanced, tie-aware win rates across all five evaluation dimensions and their unweighted Overall mean.

Under identical candidate paths, FEGR achieves Overall win rates of 62.6% against Relevance Greedy, 61.4% against MMR, 58.4% against Reliability Greedy, and 58.8% against Joint Greedy. All comparisons yield above-parity point estimates in Supportiveness and Conflict Avoidance. These results provide controlled evidence for the effectiveness of FEGR's reliability-aware evidence selection under fixed upstream conditions, complementing the complete-system comparisons in RQ1 and the component ablations in RQ4.

Table 10: Controlled evidence selection over identical candidate paths under a common 6K retrieved-context cap. Values are FEGR's position-balanced, tie-aware win rates (%) against each comparator.

Comparator	Rel.	Sup.	Conf.	Log.	Coh.	Overall
Relevance Greedy	60.8	65.0	63.2	60.5	63.5	62.6
MMR	69.8	61.2	58.2	59.2	58.5	61.4
Reliability Greedy	56.0	62.0	56.8	58.2	58.8	58.4
Joint Greedy	65.8	56.2	53.2	61.5	57.5	58.8

```
Retrieved evidence:

{{CONTEXT}}

Question:

{{QUESTION}}
```

Figure 6: User-input template used to provide the rendered evidence context and canonical query.

D Prompts and Output Schemas

D.1 Answer-Generation Input and Policy

For each query, the runtime fills Figure 6 with the rendered context and canonical query. Figure 7 gives the fixed system policy, and Table 9 reports the decoding settings.

The two markers separate the answer from its uncertainty statement, and citations are checked against identifiers in the rendered context.

D.2 Pairwise Evaluation Prompt and Output Schema

For each comparison, the judge receives the canonical query and both context–answer pairs under anonymous A/B labels. Figure 8 reproduces the prompt and JSON schema.

```
You are given a question and a set of retrieved evidence items.

Answer the question completely.
Use only supported evidence.
Synthesize information across multiple evidence items, relations, and paths when
necessary.
Do not invent unsupported details.

When the evidence supports only part of the answer, provide the supported part and
clearly state what remains uncertain. Fully abstain only when none of the retrieved
evidence is relevant enough to support a meaningful answer.

Target approximately 300-400 output tokens.
Do not exceed 450 output tokens.
Avoid unnecessary background.
Avoid repetitive explanations.
Avoid decorative headings.
Do not expand answers merely to appear more complete.
Prefer concise but sufficient answers.
If a shorter answer completely answers the question, a shorter answer is preferred.

Cite supporting evidence using the supplied identifiers exactly as they appear in the
context; examples may include [1], [E1], or [P1], depending on the context format.

Output format:

Answer:
<answer text with evidence citations>

Uncertainty:
<remaining uncertainty, or "None">
```

Figure 7: Answer-generation policy used for all methods and experimental conditions.

```
You are an expert evaluator of RAG system outputs. Compare two answers (System A and
System B) to the same question through an impartial, dimension-by-dimension pairwise
judgment. This is the J2 calibration variant.

STRICT RULES:
- Judge ONLY from the provided retrieved context. Ignore your own world knowledge
entirely. A claim not grounded in the shown context is unsupported, regardless of
real-world truth.
- "System A" / "System B" are random labels with no identity meaning. Never guess which
model produced which answer.
- Do not reward length, verbosity, formatting, or confident tone. Judge substance
against context only.
- Treat position (which answer is shown first) as irrelevant. If two answers are
genuinely comparable on a dimension, choose "tie".

Question:
{{QUESTION}}

System A Context:
{{CONTEXT_A}}

System A Answer:
{{ANSWER_A}}

System B Context:
{{CONTEXT_B}}

System B Answer:
{{ANSWER_B}}

Dimensions (evaluate each independently, context-only):

1. Relevance -- Directly addresses the question using pertinent retrieved evidence;
on-topic; brevity is fine if the question is fully answered.

2. Supportiveness -- Main factual claims trace to the provided context; faithful use of
evidence; parametric-only correctness earns NO credit.

3. Conflict Avoidance -- Handles contradictions in context; acknowledges/resolves
conflicts; no cherry-picking against better-supported evidence; disputed claims not
presented as settled. Hedging alone is not rewarded.

4. Logicality -- Internally consistent reasoning; conclusions follow from evidence; no
unwarranted leaps; no self-contradiction.

5. Coherence -- Well-structured, clear, internally consistent prose; non-native English
not penalized heavily.

For each dimension pick exactly one winner: "A", "B", or "tie" ("tie" when no clear
winner). Provide one concise rationale (≤500 chars) per dimension.

Output EXACTLY one JSON object, no other text, no markdown, no code fences.
Exact lowercase keys:
{
  "relevance": {"winner": "A|B|tie", "rationale": "..."},
  "supportiveness": {"winner": "A|B|tie", "rationale": "..."},
  "conflict_avoidance": {"winner": "A|B|tie", "rationale": "..."},
  "logicality": {"winner": "A|B|tie", "rationale": "..."},
  "coherence": {"winner": "A|B|tie", "rationale": "..."}
}
```

Figure 8: Context-grounded pairwise evaluation prompt and required output schema. The placeholders are instantiated independently for each presentation order.

A judgment is valid when all five keys are present and every winner belongs to $\{\texttt{A}, \texttt{B}, \texttt{tie}\}$. Valid ties and cross-order disagreements are retained. Each query–comparison pair is evaluated in both presentation orders:

$$J_q^{AB} = J(q, C_A, A_A, C_B, A_B), \tag{21}$$

$$J_q^{BA} = J(q, C_B, A_B, C_A, A_A). \tag{22}$$

Winner labels are mapped back to the method perspective and averaged according to Eq. (13). Rationales are kept for inspection but do not affect the score.