SemFlowRAG: Directed Semantic Flow from Abstraction to Evidence for Complex Reasoning

Houyuan Qin[1*], **Rong Wu**[1,2*], **Qinyuan Qin**[1], **Botian Shi**[1]
Jingjing Qu[1], **Yang Sun**[1,3], **Pinlong Cai**[1†]

[1]Shanghai Artificial Intelligence Laboratory
[2]Zhejiang University
[3]Fudan University

Abstract

Retrieval-Augmented Generation (RAG) enhanced by Knowledge Graphs has shown promise in complex multi-hop reasoning tasks. However, existing graph-based retrieval methods typically rely on flat, undirected topologies. During the retrieval process, the probability flow often gets trapped in high-degree abstract concept nodes which we define as "probability black holes", leading to semantic drift and noise accumulation. To address this, we propose SemFlowRAG, a framework that reconstructs the flat retrieval space into a corpus-adaptive semantic gradient graph. This data-driven self-organization enables a hierarchical structure to emerge naturally from the data distribution, capturing the intrinsic semantic granularity of the corpus to suppress structural noise. By quantifying the semantic abstractness of entities through the embedding variance of their associated passages, we transform static undirected edges into directed semantic constraints. Furthermore, we design an abstractness-guided directed PageRank algorithm that forces the retrieval trajectory to follow a "high-to-low semantic abstractness" gradient. This mechanism ensures layer-by-layer evidence convergence, smoothly guiding the retrieval process from abstract concepts to specific document evidence. Extensive experiments on complex QA datasets demonstrate that SemFlowRAG effectively mitigates the "probability black holes" issue, outperforming existing baselines in both retrieval and downstream reasoning performance. Code is available at https://github.com/KnowledgeXLab/SemFlowRAG.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable performance across diverse natural language processing tasks (Achiam et al., 2023; Yang et al., 2025). However, they remain constrained by several inherent limitations, such as hallucination and the inability to access up-to-date or domain-specific knowledge (Huang et al., 2025; Ji et al., 2023). Retrieval-Augmented Generation (RAG) effectively mitigates these issues by integrating external corpora and performing dynamic retrieval during the inference phase (Fan et al., 2024; Arslan et al., 2024). However, while vector-based RAG excels at simple factual queries, it struggles with complex reasoning tasks like multi-hop question answering, as it retrieves isolated semantic snippets that lack the coherent evidence chains and cross-document logical dependencies (Fan et al., 2024; Wu et al., 2025). Consequently, constructing semantically structured knowledge networks from large-scale document corpora to establish systematic knowledge topologies has emerged as a critical research direction for overcoming the current performance bottlenecks in RAG systems.

To bridge this gap, Graph-based RAG converts unstructured corpora into Knowledge Graph (KG) with explicit logical relations, enabling retrieval for reasoning (Edge et al., 2024; Zhang et al., 2026). However, current approaches mainly generate flat and static topological spaces, lacking fine-grained semantic constraints in both knowledge organization and path retrieval (Han et al., 2025, 2024). This oversight manifests in two critical limitations:

First, at the knowledge construction level, self-organizing mechanisms are prone to semantic construction bias. Predominant approaches rely on community detection (e.g., Microsoft GraphRAG, LightRAG) (Edge et al., 2024; Guo et al., 2024) or leverage LLM-driven summarization for hierarchical modeling (e.g., RAPTOR, LeanRAG) (Sarthi et al., 2024; Zhang et al., 2026). However, these strategies over-rely on generic model priors for information aggregation, failing to capture the intrinsic logical structure of domain-specific corpora.

Second, at the query retrieval level, mainstream strategies employ semantic hierarchy-agnostic

*Equal contribution.
†Corresponding author: caipinlong@pjlab.org.cn

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 7094–7109
October 24-29, 2026 ©2026 Association for Computational Linguistics

random-walk traversal (e.g., HippoRAG) (Gutiérrez et al., 2024). Without semantic hierarchical guidance, the walker drifts toward high-centrality abstract nodes (He et al., 2024; Veremyev et al., 2019), which absorb a disproportionate share of retrieval probability at the expense of fine-grained evidence. We term this the *"Probability Black Hole."* Consequently, ensuring accurate, traceable multi-hop reasoning requires semantic hierarchy-aware structure modeling and retrieval alignment.

To address these limitations, we propose SemFlowRAG, which constructs a hierarchical semantic gradient graph in a data-driven manner (Figure 1). The abstractness of each entity is quantified by the embedding variance of its associated passages, allowing the knowledge hierarchy to emerge from the corpus itself without external annotation. We then orient edges top-down along this gradient, establishing directed semantic information flow for complex reasoning.

Tailored to this structure, we introduce a directed Personalized PageRank (PPR) retrieval paradigm with a smooth penalty mechanism that routes probability mass downward (from abstract concepts to concrete evidence) preventing probability trapping at intermediate nodes. A preliminary analysis confirms this: our directed routing reduces top-10 entity abstractness by 73.6% (0.321 $\rightarrow$ 0.085) and boosts query similarity in ranks 60-100 by 11.3%. Unlike macro-level Graph RAG methods (Edge et al., 2024; Luo et al., 2024) that rely on discrete structural clustering and LLM-generated community summaries, SemFlowRAG injects a continuous semantic gradient into the KG, guiding retrieval to naturally converge on specific evidence without explicit graph clustering. Extensive experiments on multiple complex QA benchmarks demonstrate significant gains over existing Graph RAG baselines.

In summary, main contributions are as follows:

- We identify the "probability black hole" issue in existing flat Graph RAG paradigms and propose a dynamic graph reconstruction method that transforms static topologies into semantic gradient graphs based on entity semantic abstractness.
- We design a semantic abstractness-guided directed PPR algorithm equipped with a smooth penalty mechanism to navigate the semantic gradient graph. By constraining the retrieval flow along the "high-to-low abstraction" direction, the algorithm ensures layer-by-layer evidence convergence.
- We conduct extensive experiments across multiple complex QA benchmarks. Empirical results demonstrate that SemFlowRAG outperforms existing Graph RAG baselines, validating the effectiveness of integrating semantic gradients into structural retrieval.

2 Related Work

2.1 Structured Knowledge Representation

Early RAG systems typically organize documents into flat text chunks and rely on embedding-based similarity for retrieval (Lewis et al., 2020). To better capture inter-document relationships, subsequent works have explored graph-structured knowledge representations. For instance, GraphRAG (Edge et al., 2024) organizes text into community-based knowledge graphs, while HippoRAG (Gutiérrez et al., 2024) constructs personalized page graphs to simulate associative memory during retrieval. These methods represent a meaningful step beyond flat chunk retrieval by explicitly encoding relational structure.

To further enable multi-granularity access to knowledge, several recent approaches have introduced hierarchical organization. RAPTOR (Sarthi et al., 2024) builds a tree structure via recursive clustering and summarization of text chunks. GraphRAG (Edge et al., 2024) generates community summaries at multiple levels, effectively creating a multi-resolution view of the corpus. HiRAG (Jiao et al., 2025), LightRAG (Guo et al., 2024) and LeanRAG (Zhang et al., 2026) similarly adopt hierarchical architectures to support retrieval at varying levels of abstraction. However, these methods primarily treat the hierarchy as a multi-granularity index: they organize knowledge into layers of increasing abstraction but do not explicitly model the *semantic gradient* between abstract concepts and concrete evidence. In other words, they lack a principled mechanism to quantify how much semantic abstractness a given node contains or to distinguish abstract, high-frequency concepts from specific, low-frequency details. Our work addresses this gap by introducing abstractness-based hierarchy modeling, which provides an explicit structural prior to differentiate abstract and concrete knowledge during retrieval.

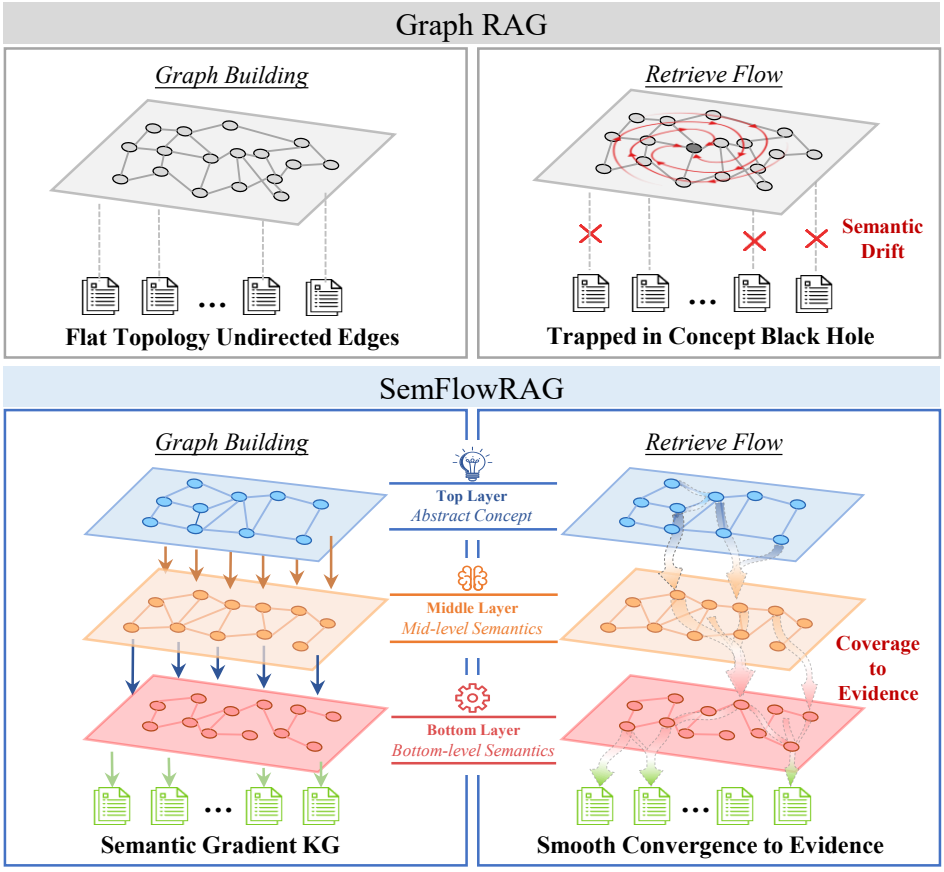

Figure 1: **Conceptual comparison.** (Top) Standard Graph RAG uses flat, undirected graphs; random walks suffer from "probability black holes" at high-degree abstract nodes. (Bottom) SemFlowRAG enforces top-down directed flow along semantic gradients, routing probability from abstract concepts to specific evidence.

2.2 Graph-based RAG

Beyond knowledge representation, how to effectively retrieve the constructed graph to get relevant evidence is a central challenge in GraphRAG. Existing graph retrieval approaches mainly include two paradigms: GNN-based message passing and random-walk-based diffusion. GNN-RAG (Mavromatis and Karypis, 2025) employs graph neural networks to propagate information across neighboring nodes, producing context-aware representations for downstream generation. While effective for localized reasoning, the receptive field of GNNs is constrained by the number of message-passing layers, and training such models introduces extra computational overhead. More recently, HippoRAG (Gutiérrez et al., 2024) adopted PPR (Haveliwala, 2002) to simulate human-like associative retrieval, using undirected random walks over personalized page graphs to identify evidence nodes in an unsupervised manner.

Despite their effectiveness, both paradigms share a common limitation: they treat the retrieval graph as undirected and overlook difference in semantic hierarchy. During diffusion, the walker's transition is governed primarily by structural connectivity and local semantic similarity, with no mechanism to distinguish abstract, high-centrality concept nodes from specific evidence nodes. In densely connected knowledge graphs, this causes probability mass to accumulate at high-degree concept nodes leading to semantic drift and diluting the retrieval signal for fine-grained evidence passages. To address this, we propose a top-down directed routing mechanism. By quantifying node-level semantic abstractness and incorporating an abstractness penalty into the transition weights, we explicitly regularize the random walk to flow from abstract concepts toward specific evidence passages, effectively mitigating probability trapping at intermediate, non-evidential nodes.

3 Methodology

3.1 Overview

The proposed SemFlowRAG framework operates in a two-stage paradigm: offline graph construction and online directed retrieval, as illustrated in Figure 2. In the offline stage, we upgrade a standard flat KG into a *Semantic Gradient KG* by explicitly quantifying the semantic abstractness of entities based on their associated passages. This step transforms undirected, static KG into a knowledge representation with an inherent top-down semantic hierarchy. Built upon this representation, the online stage introduces a semantic information-guided directed retrieval mechanism. By employing a directionally constrained PPR, the retrieval trajectory is explicitly guided to follow the semantic gradient, smoothly penetrating from generalized concepts and ultimately converging on the passage nodes that contain concrete and relevant evidence.

3.2 Offline Stage: Semantic Gradient KG Construction

The primary objective of offline stage is to reconstruct a static, flat KG into a knowledge representation network with an intrinsic semantic hierarchy. This is achieved through three sequential steps: basic graph extraction, semantic abstractness quantification and semantic gradient edge construction.

Entity and Fact Extraction. Given a corpus of documents, we first segment them into retrievable passages and cache their dense embeddings. For each passage, we apply Open Information Extraction (OpenIE) to extract named entities and factual relational triplets (i.e., subject-relation-object). The extracted subjects and objects are normalized and mapped to entity nodes in the graph. We establish unweighted *entity-entity* edges based on the co-occurrence of entities within the same triplet. Simultaneously, to preserve the source of the knowledge, we record the occurrences of each entity and construct bidirectional *passage-entity* edges, explicitly linking entities to the specific passages they originate from.

Semantic Abstractness Quantification. In existing flat KGs (Liang et al., 2025; Jiao et al., 2025; Gutiérrez et al., 2025), all entities are treated as topologically equivalent, ignoring the semantic hierarchical distinction between broad and specific concepts. To explicitly model this hierarchy, we propose a metric to quantify the "semantic abstractness" of each entity. Let $\mathcal{P}(v)$ denote the set of passages containing entity v, and $\mathbf{h}_p$ be the dense embedding of passage p. We define the semantic abstractness d_v of entity v as the trace of the covariance matrix of its associated passage embeddings:

$$d_v = \mathrm{tr}\left(\mathrm{Var}\left(\{\mathbf{h}_p \mid p \in \mathcal{P}(v)\}\right)\right). \quad (1)$$

Mathematically, the covariance matrix $\mathrm{Var}(\cdot)$ captures the dispersion of the entity's contextual embeddings across the semantic space. Taking the trace $\mathrm{tr}(\cdot)$ sums the variances along all principal dimensions, effectively computing the total spatial variance. The fundamental insight behind this formulation is that abstract, high-level concepts (e.g., "country", "science") naturally appear in highly diverse and semantically disconnected contexts, resulting in a widely scattered distribution and a large total variance. Conversely, specific, fine-grained entities (e.g., "the 2024 Paris Olympics") are typically constrained to a narrow, well-defined semantic region, yielding a highly concentrated distribution and a small variance. Thus, d_v serves as an unsupervised proxy for corpus-level semantic breadth, rather than universal ontological abstractness.

To mitigate the influence of extreme outliers, we apply a truncated min-max normalization to the abstractness scores of all entities, clipping the values at the 1st and 99th percentiles. This scales d_v into a standardized range of $[0, 1]$.

Semantic Gradient Edge Construction. Based on the quantified semantic abstractness, we establish a semantic gradient across the graph. Instead of treating the entity-entity connections as purely undirected co-occurrence links, we compute the semantic span for every connected pair of entities u and v as $\Delta_{u,v} = d_v - d_u$.

By explicitly recording semantic gradient priors ($\Delta_{u,v}$) offline, we transform the flat topological space into a directed, hierarchical knowledge representation. This offline gradient builds a "high-to-low abstraction" scaffold within the graph, which serves as the fundamental constraint for guiding the online probability flow and preventing the retrieval trajectory from converging into abstract nodes.

3.3 Online Stage: Semantic Information-Guided Directed Retrieval

With the Semantic Gradient KG constructed offline, the online stage aims to navigate this hierarchical topology based on a given query. We formulate this process as a dynamically weighted, directed

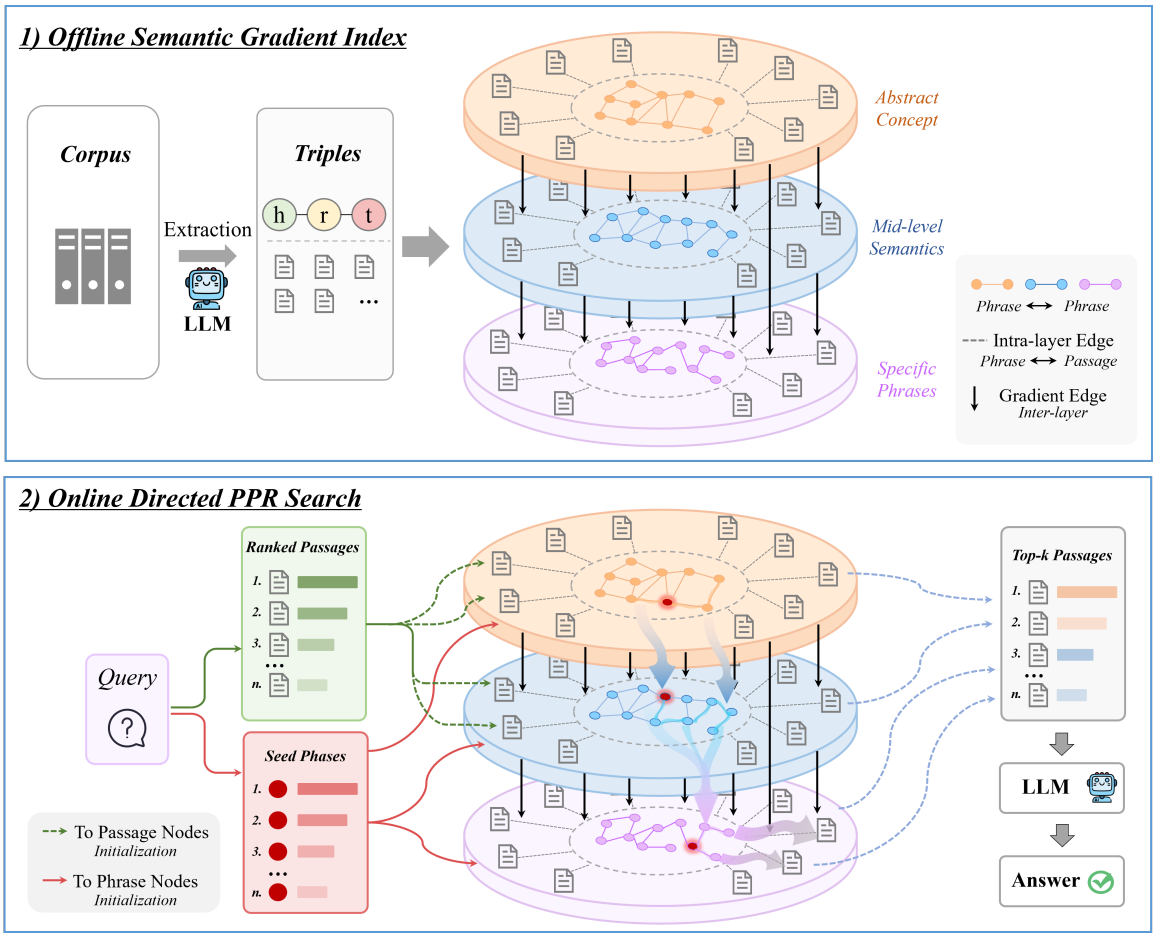

Figure 2: **Overview of SemFlowRAG. 1) Offline:** We construct a hierarchical graph with top-down directed edges, stratifying nodes by semantic abstractness. **2) Online:** A directed PPR search routes probability mass from query seed nodes downward along semantic gradients to retrieve fine-grained evidence passages.

PPR algorithm, which explicitly forces the retrieval trajectory to smoothly converge upon fine-grained and relevant evidence.

Seed Retrieval and Initialization. Given a query q, the system first identifies the initial entry points (seeds) for the graph random walk. We compute the similarity between q and all extracted facts to retrieve the top-k candidate facts. An optional LLM-based filter can be applied to retain only high-quality, query-relevant facts. If no facts pass the filter, the system falls back to standard Dense Passage Retrieval (DPR) without graph propagation.

For the retained facts, their subjects and objects are mapped to the corresponding entity nodes to serve as seeds. To mitigate the topological bias of high-frequency entities, the initial weight of each seed entity is determined by its query-fact similarity normalized by its occurrence frequency in the corpus. Furthermore, we incorporate the DPR scores of passage nodes as a weak prior, controlled by a weight coefficient. Consequently, the final PPR reset distribution (the restart vector) is a hybrid formulation comprising both the normalized entity seed weights and the passage prior weights. More details are provided in the Appendix A.2.

Semantic Information-Guided Layer-by-Layer Convergence. The core of our directed retrieval lies in the dynamic rewriting of transition probabilities. During the random walk, the algorithm dynamically balances the online *query relevance* with the offline *semantic gradient*. For entity v, we define its query relevance as the maximum similarity between q and its associated passages:

$$Rel(v|q) = \max_{p \in \mathcal{P}(v)} \text{sim}(q, p), \tag{2}$$

which is min-max normalized to $[0, 1]$. A higher $Rel(v|q)$ indicates that traversing through v is more likely to yield query-relevant contexts.

To determine the transition score $s(u, v)$ for a directed edge $u \to v$, we inject a smooth penalty mechanism:

$$s(u, v) = Rel(v \mid q) - \lambda|\Delta_{u,v}|, \tag{3}$$

where $\Delta_{u,v} = d_v - d_u$ is the pre-computed semantic gradient from the offline stage, and λ is a penalty coefficient, we set $\lambda = 1$ as default. This formulation explicitly rewards paths leading to query-relevant nodes while penalizing excessively large jumps in semantic abstractness. By penalizing large semantic gaps, we enforce a smooth gradient, ensuring that the retrieval trajectory follows a step-by-step reasoning chain and mitigating the risk of topic drift during the random walk.

To enforce the top-down semantic information flow, we apply an asymmetric directional routing strategy to the graph edges. Specifically, out of the probability mass allocated for edge transitions, the vast majority (e.g., 90%) is distributed to *down-edges* (transitions from higher to lower semantic abstractness), while a marginal fraction (e.g., 10%) is reserved for *up-edges*. This design ensures that the probability mass smoothly penetrates downwards layer by layer, effectively avoiding the "probability black holes" associated with high-frequency abstract nodes. Meanwhile, we preserve a small amount of upward connectivity to simulate the associative recall process in human memory, where reasoning does not proceed strictly downward but occasionally returns to broader concepts. Note that passage-entity edges are kept bidirectional and unweighted to allow flexible information interaction between entity nodes and passage nodes.

Passage Ranking and Output. Driven by the directionally constrained PPR, the probability flow trickles down the semantic gradient and naturally accumulates on the specific passage nodes. Once the PPR reaches its steady-state distribution, we directly extract the PageRank scores of all passage nodes. The top-k scoring passages are retrieved as the final supporting context, which is subsequently fed into the downstream LLM to generate answer.

4 Experiments

In this section, we empirically evaluate SemFlowRAG. Our experiments are designed to answer the following core research questions (RQs):

- **RQ1 (Main Results):** Can SemFlowRAG achieve state-of-the-art performance on complex multi-hop reasoning tasks compared to existing RAG baselines?
- **RQ2 (Ablation on Topology):** Does enforcing a directed walk along the semantic gradient improve performance compared to standard undirected traversal?
- **RQ3 (Ablation on Weighting):** Is the abstractness penalty mechanism in the transition probability more effective than relying solely on query similarity?

4.1 Experimental Setup

Datasets. We evaluate SemFlowRAG on a diverse suite of question answering benchmarks covering factual retrieval, open-domain QA, multi-hop reasoning and long-context evidence aggregation, including NaturalQuestions (NQ) (Kwiatkowski et al., 2019), PopQA (Mallen et al., 2023), MuSiQue (Trivedi et al., 2022), 2WikiMultiHopQA (Ho et al., 2020), HotpotQA (Yang et al., 2018), LV-Eval (Yuan et al., 2024) and NarrativeQA (Kočiskỳ et al., 2018). These datasets collectively test whether a system can move beyond shallow semantic matching and identify evidence chains that support complex reasoning. We evaluate each dataset on a selected subset of queries, with the corresponding dataset statistics summarized in Table 1.

Baselines. We compare SemFlowRAG with three groups of baselines. First, we include unstructured knowledge representation augmented baselines, covering a no-retrieval setting for QA, the sparse retriever BM25 (Robertson and Walker, 1994) and so on. Second, we compare embedding-based knowledge representation augmented baselines, including GTE-Qwen2-7B-Instruct (Li et al., 2023), GritLM-7B (Muennighoff et al., 2025) and so on. Third, we include structured knowledge representation augmented methods, including RAPTOR (Sarthi et al., 2024), HippoRAG 2 (Gutiérrez et al., 2025) and so on. Baseline details are provided in the Appendix A.1. For fair comparison, reproduced structure-augmented baselines are evaluated under the same LLM, embedding model, retrieval budget and generation setting whenever applicable. Additional implementation details are provided in the Appendix A.2.

Metrics. For retrieval, we report Passage Recall@5, defined as the fraction of gold evidence passages successfully retrieved within the top-5 results. For answer generation, we compute F1 based on exact match (EM) between the generated answers and the ground-truth answers. All results are reported under identical top-5 retrieval and generation settings to ensure fair comparison.

	NQ	PopQA	MuSiQue	2Wiki	HotpotQA	LV-Eval	NarrativeQA
Num of queries	1,000	1,000	1,000	1,000	1,000	124	293
Num of passages	9,633	8,676	11,656	6,119	9,811	22,849	4,111

Table 1: Dataset statistics used in our evaluation.

Method	Simple QA		Multi-Hop QA				Discourse	Avg.
	NQ	PopQA	MuSiQue	2Wiki	HotpotQA	LV-Eval	NarrativeQA	
Unstructured Knowledge Representation								
Llama-3.3-70B-Instruct w/o Retrieval	54.9	32.5	26.1	42.8	47.3	6.0	12.9	38.4
BM25	59.0	49.9	28.8	51.2	63.4	5.9	18.3	47.7
Contriever	58.9	53.1	31.3	41.9	62.3	8.1	19.7	46.9
GTR	59.9	56.2	34.6	52.8	62.8	7.1	19.9	50.4
Embedding-based Knowledge Representation								
GTE-Qwen2-7B-Instruct	62.0	**56.3**	40.9	60.0	71.0	7.1	21.3	54.9
GritLM-7B	61.3	55.8	44.8	60.6	73.3	9.8	23.9	56.1
NV-Embed-v2 (7B)	61.9	55.7	45.7	61.5	75.3	9.8	25.7	57.0
Structured Knowledge Representation								
RAPTOR	50.7	56.2	28.9	52.1	69.5	5.0	21.4	48.8
GraphRAG	46.9	48.1	38.5	58.6	68.6	11.2	23.0	49.6
LightRAG	16.6	2.4	1.6	11.6	2.4	1.0	3.7	6.6
HippoRAG	55.3	55.9	35.1	71.8	63.5	8.4	16.3	53.1
HippoRAG 2	**63.3**	56.2	48.6	71.0	75.5	**12.9**	25.9	59.8
HippoRAG 2 (reproduced)	62.1	55.7	48.6	72.5	75.5	9.6	26.2	59.6
SemFlowRAG (Ours)	62.3	56.2	**51.4**	**76.9**	**75.8**	10.6	**26.6**	**61.2**

Table 2: Reasoning performance (F1 score) on seven benchmarks. **Best** and second-best are highlighted.

4.2 Main Results (RQ1)

Table 2 and Table 3 present the downstream QA and evidence retrieval results. SemFlowRAG achieves the highest overall performance, yielding an average F1 of 61.2 and Recall@5 of 79.3.

The gains are most pronounced on multi-hop reasoning tasks (MuSiQue, 2Wiki and HotpotQA). SemFlowRAG outperforms the strongest baseline, HippoRAG 2, by 4.4 and 2.8 F1 points on 2Wiki and MuSiQue, respectively, while reaching a 94.8 Recall@5 on 2Wiki. This validates that constraining retrieval along a semantic gradient provides reliable evidence paths for complex queries, effectively mitigating noise accumulation and semantic drift. On Simple QA (NQ, PopQA) and Discourse (NarrativeQA) tasks, SemFlowRAG remains highly competitive (e.g., achieving the highest F1 of 26.6 on NarrativeQA). This demonstrates that our design aids long-context synthesis without degrading single-hop retrieval efficiency.

Notably, while standard undirected or tree-based methods (e.g., GraphRAG, RAPTOR) sometimes underperform the dense retriever (NV-Embed-v2) due to graph-induced structural noise, SemFlowRAG consistently improves upon NV-Embed-v2. This confirms that our directed edges effectively filter noise and accurately guide the reasoning trajectory.

4.3 Ablation Study

To understand the contribution of each component in SemFlowRAG, we conduct careful ablation studies focusing on the graph topology and the transition weights. Table 4 presents the ablation results of SemFlowRAG on multi-hop QA benchmarks. We evaluate the contribution of direction control, query similarity and abstractness difference in semantic-aware graph retrieval.

4.3.1 Effect of Directed Topology (RQ2)

To answer RQ2, we evaluate the "w/o Direction Control" variant by disabling the asymmetric directional retrieval. In this setting, instead of forcing the probability mass to flow primarily downwards (from high to low), we allow it to distribute symmetrically across both down-edges and up-edges.

As shown in Table 4, removing this directional constraint reduces the average F1 score from 68.0 to 67.4. Without the constraint, the transition probabilities rely primarily on the edge weight $s(u, v)$. Because the abstractness penalty $|\Delta_{u,v}|$ in the edge weight is symmetric, the random walk can easily move upward toward broader concepts that also exhibit high query relevance. Consequently, the retrieval trajectory tends to wander among abstract

Method	Simple QA		Multi-Hop QA			Avg.
	NQ	PopQA	MuSiQue	2Wiki	HotpotQA	
Unstructured Knowledge Representation Augmented						
BM25	56.1	35.7	43.5	65.3	74.8	55.1
Contriever	54.6	43.2	46.6	57.5	75.3	55.4
GTR	63.4	49.4	49.1	67.9	73.9	60.7
Embedding-based Knowledge Representation Augmented						
GTE-Qwen2-7B-Instruct	74.3	50.6	63.6	74.8	89.1	70.5
GritLM-7B	76.6	50.1	65.9	76.0	92.4	72.2
NV-Embed-v2 (7B)	75.4	51.0	69.7	76.5	94.5	73.4
Structured Knowledge Representation Augmented						
RAPTOR	68.3	48.7	57.8	66.2	86.9	65.6
HippoRAG (reproduced)	44.4	**53.8**	53.2	90.4	77.3	63.8
HippoRAG 2	**78.0**	51.7	74.7	90.4	**96.3**	78.2
HippoRAG 2 (reproduced)	77.0	51.7	73.8	90.7	**96.3**	77.9
SemFlowRAG (Ours)	77.5	52.8	**75.2**	**94.8**	96.1	**79.3**

Table 3: Retrieval performance (passage recall@5) on five benchmarks. **Best** and second-best are highlighted.

Method	Components			Multi-Hop QA			
	Direction Control	Query Similarity	Delta Abstractness	MuSiQue	2Wiki	HotpotQA	Avg.
SemFlowRAG	✓	✓	✓	**51.4**	**76.9**	**75.8**	**68.0**
w/o Direction Control		✓	✓	50.8	76.0	75.5	67.4
w/o Query Similarity	✓		✓	50.9	76.5	75.3	67.6
w/o Delta Abstractness	✓	✓		51.2	76.4	75.3	67.6

Table 4: Ablation study of SemFlowRAG on multi-hop QA benchmarks.

entities instead of actively penetrating down to the specific passages where the actual evidence resides. The performance drop is most notable on 2Wiki, confirming that for scenarios requiring longer reasoning chains, an explicit downward constraint is necessary to prevent semantic drift and ensure the walk converges on concrete evidence.

4.3.2 Effect of Semantic Weighting (RQ3)

To answer RQ3, we evaluate the two components that formulate the transition weights in Equation 3: query similarity and the abstractness penalty. In the "w/o Query Similarity" variant, we remove the relevance score $Rel(v \mid q)$, making the transition probabilities depend solely on the abstractness difference between nodes. Conversely, in the "w/o Delta (abstractness)" variant, we set the penalty coefficient $\lambda = 0$, relying entirely on the semantic similarity between the query and the target node to guide the random walk.

As shown in Table 4, removing either component reduces the average F1 score. Without query similarity, the random walk blindly follows the structural gradient to lower-level nodes, risking the retrieval of evidence that is structurally reachable but irrelevant to the initial question. On the other hand, removing the abstractness penalty causes the walk to rely only on local textual similarity. This can lead to abrupt semantic jumps to nodes that are textually similar to the query but disrupt the continuous, smooth retrieval chain. The equal performance drop indicates that both signals are complementary and indispensable: query similarity ensures task-specific relevance, while the abstractness penalty acts as a regularization term to maintain smooth transitions along the semantic hierarchy.

5 Conclusion

In this paper, we introduced SemFlowRAG, a framework that organizes a corpus into a directed semantic hierarchy and guides multi-hop retrieval via a top-down random walk. By enforcing asymmetric directional routing and combining query similarity with a abstractness penalty in the transition weights, SemFlowRAG steers the retrieval process from abstract concepts toward specific evidential passages. Extensive experiments on multiple complex QA benchmarks demonstrate consistent improvements over competitive baselines. Ablation studies confirm that both the directional constraint and the semantic weighting scheme contribute meaningfully to retrieval quality.

Limitations

Despite its strong performance, SemFlowRAG has several limitations that warrant discussion. First, the effectiveness of our framework depends on the quality of the underlying semantic graph. The abstractness scores and entity relationships are derived from an offline preprocessing stage; errors or omissions in this stage can propagate downstream and degrade retrieval quality. Second, our top-down directed routing makes an explicit structural assumption—that evidence flows from abstract concepts to specific facts. While this holds broadly across the benchmarks we evaluated, it may not generalize to scenarios requiring lateral reasoning or bottom-up inference, where high-abstractness nodes are the desired targets rather than sources. Third, as shown in our sensitivity analysis, the optimal reset probability varies across datasets, indicating that performance is somewhat sensitive to hyperparameter tuning. Developing an adaptive mechanism to dynamically adjust the exploration depth remains an important direction for future work.

Ethical Considerations

The development of SemFlowRAG presents ethical considerations common to RAG systems. On the positive side, our framework enhances the transparency and traceability of LLM outputs. By constraining the retrieval trajectory to converge on concrete passage evidence, it helps mitigate LLM hallucinations and allows users to straightforwardly verify the exact evidence used for generation. However, because SemFlowRAG derives its semantic hierarchy and directed edges entirely in a data-driven manner, it is inherently susceptible to the quality of the source corpus. If the corpus contains biases, misinformation, or toxic content, the directed information flow may systematically retrieve and amplify these flawed narratives. Therefore, deploying this framework in real-world, user-facing applications requires careful curation and rigorous auditing of the knowledge base to ensure safety and fairness.

Acknowledgments

This work was supported by Shanghai Artificial Intelligence Laboratory.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Muhammad Arslan, Hussam Ghanem, Saba Munawar, and Christophe Cruz. 2024. A survey on rag with llms. *Procedia computer science*, 246:3781–3790.

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*.

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

Zirui Guo, Lianghao Xia, Yanhua Yu, Tian Ao, and Chao Huang. 2024. Lightrag: Simple and fast retrieval-augmented generation. *arXiv preprint arXiv:2410.05779*, 2(3).

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

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. 2025. From rag to memory: Non-parametric continual learning for large language models. *arXiv preprint arXiv:2502.14802*.

Haoyu Han, Li Ma, Yu Wang, Harry Shomer, Yongjia Lei, Zhisheng Qi, Kai Guo, Zhigang Hua, Bo Long, Hui Liu, and 1 others. 2025. Rag vs. graphrag: A systematic evaluation and key insights. *arXiv preprint arXiv:2502.11371*.

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

Taher H Haveliwala. 2002. Topic-sensitive pagerank. In *Proceedings of the 11th international conference on World Wide Web*, pages 517–526.

Mutian He, Tianqing Fang, Weiqi Wang, and Yangqiu Song. 2024. Acquiring and modeling abstract commonsense knowledge via conceptualization. *Artificial Intelligence*, 333:104149.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop

qa dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating llm hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843.

Yihan Jiao, Zhehao Tan, Dan Yang, Duolin Sun, Jie Feng, Yue Shen, Jian Wang, Peng Wei, and AntGroup Hangzhou Zhejiang China. 2025. Hirag: Hierarchical-thought instruction-tuning retrieval-augmented generation. *arXiv preprint arXiv:2507.05714*.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025. Nv-embed: Improved techniques for training llms as generalist embedding models. In *International Conference on Learning Representations*, volume 2025, pages 79310–79333.

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

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Lei Liang, Zhongpu Bo, Zhengke Gui, Zhongshu Zhu, Ling Zhong, Peilong Zhao, Mengshu Sun, Zhiqiang Zhang, Jun Zhou, Wenguang Chen, and 1 others. 2025. Kag: Boosting llms in professional domains via knowledge augmented generation. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 334–343.

Linhao Luo, Yuan-Fang Li, Reza Haffari, and Shirui Pan. 2024. Reasoning on graphs: Faithful and interpretable large language model reasoning. In *International Conference on Learning Representations*, volume 2024, pages 14400–14423.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 9802–9822.

Costas Mavromatis and George Karypis. 2025. Gnn-rag: Graph neural retrieval for efficient large language model reasoning on knowledge graphs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16682–16699.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2025. Generative representational instruction tuning. In *International Conference on Learning Representations*, volume 2025, pages 45544–45613.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and 1 others. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855.

Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*, pages 232–241. Springer.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher Manning. 2024. Raptor: Recursive abstractive processing for tree-organized retrieval. In *International Conference on Learning Representations*, volume 2024, pages 32628–32649.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Alexander Veremyev, Alexander Semenov, Eduardo L Pasiliao, and Vladimir Boginski. 2019. Graph-based exploration and clustering analysis of semantic spaces. *Applied Network Science*, 4(1):104.

Rong Wu, Pinlong Cai, Jianbiao Mei, Licheng Wen, Tao Hu, Xuemeng Yang, Daocheng Fu, and Botian Shi. 2025. Kg-traces: Enhancing large language models with knowledge graph-constrained trajectory reasoning and attribution supervision. *arXiv preprint arXiv:2506.00783*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2369–2380.

Tao Yuan, Xuefei Ning, Dong Zhou, Zhijie Yang, Shiyao Li, Minghui Zhuang, Zheyue Tan, Zhuyu Yao, Dahua Lin, Boxun Li, and 1 others. 2024. Lv-eval: A balanced long-context benchmark with 5 length levels up to 256k. *arXiv preprint arXiv:2402.05136*.

Yaoze Zhang, Rong Wu, Pinlong Cai, Xiaoman Wang, Guohang Yan, Song Mao, Ding Wang, and Botian Shi. 2026. Leanrag: Knowledge-graph-based generation with semantic aggregation and hierarchical retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 34862–34869.

A Appendix

A.1 Baseline Details

We compare SemFlowRAG with three groups of baselines. First, we include unstructured knowledge representation augmented baselines, covering a no-retrieval setting for QA, the sparse retriever BM25 (Robertson and Walker, 1994) and widely used dense retrievers including Contriever (Izacard et al., 2021) and GTR (Ni et al., 2022). Second, we compare against embedding-based knowledge representation augmented baselines, including GTE-Qwen2-7B-Instruct (Li et al., 2023), GritLM-7B (Muennighoff et al., 2025) and NV-Embed-v2 (Lee et al., 2025), which represent competitive retrieval-only alternatives based on recent large-scale embedding models. Third, we include structured knowledge representation augmented methods, including RAPTOR (Sarthi et al., 2024), GraphRAG (Han et al., 2025), LightRAG (Guo et al., 2024), HippoRAG (Gutiérrez et al., 2024) and HippoRAG 2 (Gutiérrez et al., 2025).

A.2 Implementation Details

Backbone models. We use Llama-3.3-70B-Instruct as the backbone language model and NV-Embed-v2 as the embedding model. The language model is responsible for OpenIE extraction, fact filtering and final answer generation. NV-Embed-v2 is used to encode passages, entities, facts and queries. We serve Llama-3.3-70B-Instruct with vLLM (Kwon et al., 2023) using tensor parallelism on 4 × NVIDIA A100 GPUs.

Indexing and OpenIE. We build the retrieval graph from OpenIE triples extracted from the corpus. To improve the entity coverage of the constructed graph, we use a high-recall OpenIE prompt that encourages extracting explicit factual statements, including aliases, dates, quantities, locations, affiliations, events, roles and descriptive attributes. We show prompt details in A.5.1. Compound statements are split into separate triples when they express multiple facts. Unless otherwise specified, the maximum generation lengths are 512 tokens for named entity recognition (NER) and 2048 tokens for triple extraction.

PPR reset distribution. The PPR reset distribution defines where the random walk returns when a reset step is triggered. This distribution is constructed as a mixture of two types of seeds: entity seeds and passage seeds. Entity seeds are obtained from the subject and object entities of the retained query-relevant facts. Their weights are determined by the corresponding query-fact similarity.

Passage seeds are introduced from DPR retrieval scores over passage nodes. All passage nodes are assigned reset weights according to their DPR scores, and these weights are scaled by passage reset weight. In this way, highly relevant passages can receive part of the reset probability directly, while the overall reset distribution is still dominated by entity seeds. During PPR propagation, each reset step returns to this mixed distribution over both entity nodes and passage nodes, rather than returning only to entity seeds.

Retrieval configuration. We use directed PPR over a graph containing both entity and passage nodes. We set passage reset weight to 0.08. We use a PPR reset probability of 0.5. For semantic direction control, the transition probability budget is split into 0.9 for abstractness decreasing edges and 0.1 for abstractness increasing edges. We first retrieve the top 12 candidate facts by embedding similarity of query and fact, and then apply an LLM-based fact filter that keeps up to 12 query-relevant facts. We use the entities appearing in the filtered facts to construct the entity part of the PPR reset vector, and retain the top 20 entity seeds.

Answer generation. For answer generation, the LLM uses the top 5 retrieved passages. We use a QA temperature of 0.4 and keep the global temperature for other LLM calls at 0.

Efficiency and computational overhead. We further analyze the computational and storage overhead introduced by the semantic gradient graph and directed PPR retrieval. Table 5 reports the offline construction and storage costs, while Table 6 reports the average online retrieval latency per query. The additional storage mainly consists of directed graph metadata, including entity-level abstractness scores, edge directions, and normalized transition weights. Since this process is performed only once during indexing, the cost can be amortized over subsequent queries. The additional storage mainly consists of directed graph metadata, including entity-level abstractness scores and transition-related edge information. During online retrieval, directed PPR introduces additional semantic directional guidance, resulting in moderately higher retrieval latency.

Dataset	Construction	Storage (MB)
2Wiki	34s + 1m41s	18.2 → 22.5
HotpotQA	1m35s + 5m11s	34.6 → 42.9
MuSiQue	1m45s + 6m21s	37.7 → 47.5
PopQA	1m24s + 4m35s	31.8 → 39.1
NQ	1m09s + 6m33s	31.8 → 44.3
LV-Eval	6m04s + 1h02m	96.7 → 122.4
NarrativeQA	3s + 18s	5.0 → 6.4

Table 5: Offline efficiency comparison between the standard KG retrieval pipeline and SemFlowRAG. Construction reports the standard KG construction time plus the additional semantic-gradient preprocessing time. Storage is reported as KG → SemFlowRAG.

Dataset	KG	SemFlowRAG
2Wiki	1.600	2.008
HotpotQA	1.775	2.579
MuSiQue	1.695	2.509
PopQA	1.508	2.247
NQ	1.941	2.909
LV-Eval	2.071	3.879
NarrativeQA	1.314	1.342

Table 6: Average online retrieval latency per query in seconds. SemFlowRAG introduces additional directional guidance along the semantic gradient during directed PPR retrieval.

A.3 Additional Experimental Results

A.3.1 Sensitivity Analysis

We evaluate the sensitivity of SemFlowRAG to three key hyperparameters: the PPR reset probability, the number of entity seeds and the QA decoding temperature. All experiments are conducted under the default configuration described above, varying only the parameter under test.

PPR reset probability. This parameter controls how often the random walk returns to the initial seed nodes. As shown in Figure 3, increasing it from 0.1 to 0.5 consistently improves the average F1 score from 53.5 to 61.2. A lower reset probability allows deeper graph traversal but risks semantic drift by diluting the probability mass into noisy, irrelevant nodes. The anchoring effect is particularly beneficial for multi-hop datasets: the F1 score on 2Wiki jumps from 64.7 to 76.9, indicating that while complex reasoning requires multi-hop exploration, the necessary evidence typically resides within a few hops of the initial seeds. Although a few datasets (e.g., LV-Eval and PopQA) peak or plateau earlier, a reset probability of 0.5 provides the most robust regularization overall.

Entity seeds. We examine the number of initial entity seeds used in the PPR reset vector (Table 7).

Seeds	NQ	PopQA	MuSiQue	2Wiki	HotpotQA	Avg.
8	62.6	56.0	51.6	76.0	75.7	64.4
16	62.5	56.3	51.2	76.3	75.9	64.4
20	62.3	56.2	51.4	76.9	75.8	64.5

Table 7: Sensitivity analysis of the number of entity seeds used in the PPR reset vector. All values are QA F1 scores.

For this experiment, the QA temperature is fixed at 0.8. Varying the seed count from 8 to 20 yields minimal fluctuation in both per-dataset and average F1 scores (64.4-64.5), with a marginal gain of 0.9 points on 2Wiki at 20 seeds. This stability indicates that our fact-filtering pipeline already supplies a sufficient set of query-relevant entities, and that SemFlowRAG is largely insensitive to this parameter within a reasonable range.

QA decoding temperature. We further assess the impact of the QA generation temperature (Table 8), with the number of entity seeds fixed at 5. Across the range of 0.0 to 0.8, the average F1 varies within a narrow band of 63.8-64.0, with no dataset exhibiting a monotonic upward or downward trend. This confirms that our approach is robust to decoding stochasticity; a moderate temperature (e.g., 0.2-0.6) offers a good balance between determinism and output diversity, while extreme values incur no substantial degradation.

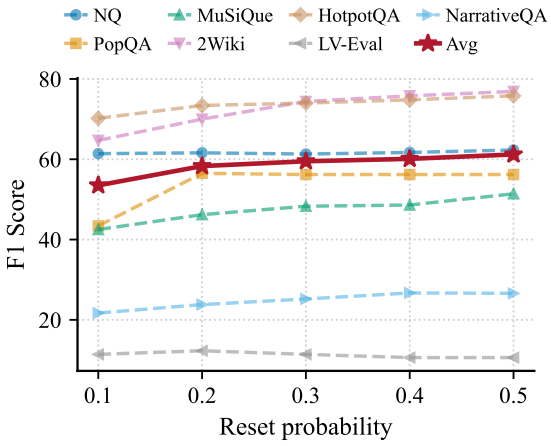

Figure 3: **Reset probability factor.** We report QA performance (F1 scores) with different PPR reset probabilities in our algorithm.

A.4 Case Study

To illustrate the retrieval and reasoning process of SemFlowRAG, we provide a case study in Figure 4. The given query requires two-hop reasoning: first

Case Study: End to End QA Performance of SemFlowRAG

Question: Who is the child of the director of film *An Event*?

Ground Truth: Sergio Mimica-Gezzan

Retrieved Top-5 Passages:

1. **An Event:** An Event is a 1969 Yugoslav feature film directed by Vatroslav Mimica, based on a short story by Anton Chekhov...
2. **Vatroslav Mimica:** Vatroslav Mimica is a Croatian film director and screenwriter. His son Sergio Mimica-Gezzan is an American film and television director...
3. **M. Night Shyamalan:** Manoj Nelliyattu "M. Night" Shyamalan is an American filmmaker and actor...
4. **Mario Bava:** Mario Bava was an Italian cinematographer, director, special effects artist and screenwriter...
5. **Anup Sengupta:** Anup Sengupta is a Bengali film Director and Producer...

Answer: Sergio Mimica-Gezzan.

Figure 4: A case study of SemFlowRAG.

identifying the director of the film "*An Event*", and subsequently finding the director's child.

As shown in the retrieved top-5 passages, SemFlowRAG successfully captures the complete evidence chain. Passage 1 resolves the first hop by identifying Vatroslav Mimica as the director of the film. Building upon this, Passage 2 resolves the second hop by explicitly stating that his son is Sergio Mimica-Gezzan. By securing these structurally connected facts at the very top of the retrieved context, SemFlowRAG provides a clear reasoning evidence. Consequently, the LLM is able to accurately generate the right answer.

A.5 Prompt Details

A.5.1 KG build prompt

We show LLM prompts for NER and triple extraction in Figure 5 and 6, including the system instructions and demonstrations for both tasks. The NER prompt is used to extract named entities from the passage, while the triple extraction prompt further converts the passage and entity list into RDF-style triples.

NER Prompt

System prompt:
Your task is to extract named entities from the given paragraph. Respond with a JSON list of entities.

Demonstration input:
Radio City
Radio City is India's first private FM radio station and was started on 3 July 2001. It plays Hindi, English and regional songs.
Radio City recently forayed into New Media in May 2008 with the launch of a music portal - PlanetRadiocity.com that offers music related news, videos, songs, and other music-related features.

Demonstration output:
{"named_entities":
["Radio City", "India", "3 July 2001", "Hindi", "English", "May 2008", "PlanetRadiocity.com"]
}

Figure 5: Prompts and demonstrations for NER

A.5.2 LLM fact filter prompt

We show LLM prompts for fact filtering in Figure 7. The prompt requires the model to select relevant facts from candidate triples while preserving answer-bearing facts and bridge facts for multi-hop reasoning.

Dataset	0.0	0.2	0.4	0.6	0.8
NQ	0.6235	0.6264	0.6267	0.6290	0.6193
PopQA	0.5573	0.5588	0.5587	0.5598	0.5639
MuSiQue	0.5138	0.5072	0.5023	0.5037	0.5015
2Wiki	0.7445	0.7536	0.7602	0.7526	0.7542
HotpotQA	0.7525	0.7531	0.7541	0.7540	0.7566
Avg.	0.6383	0.6398	0.6404	0.6398	0.6391

Table 8: Sensitivity analysis of QA decoding temperature. All values are QA F1 scores.

Triple Extraction Prompt

System prompt:
Your task is to construct an RDF (Resource Description Framework) graph from the given passages and named entity lists. Respond with a JSON list of triples, with each triple representing a relationship in the RDF graph. Pay attention to the following requirements:

- Extract as many explicit factual statements as possible from the paragraph; aim for high recall rather than a small summary.
- Each triple should contain at least one named entity when possible, but do not omit explicit facts whose useful endpoints are descriptive phrases, dates, quantities, titles, roles, events, locations, affiliations, or other salient noun phrases.
- Split enumerations and compound statements into separate triples when they express multiple facts.
- Preserve factual specificity from the paragraph, including aliases, dates, numbers, locations, occupations, memberships, affiliations, works, events, and descriptive attributes.
- Do not infer facts that are not explicitly supported by the paragraph.
- Clearly resolve pronouns to their specific names to maintain clarity.

Demonstration input:
Convert the paragraph into a JSON dict, it has a named entity list and a triple list.
Paragraph:
Radio City
Radio City is India's first private FM radio station and was started on 3 July 2001. It plays Hindi, English and regional songs.
Radio City recently forayed into New Media in May 2008 with the launch of a music portal - PlanetRadiocity.com that offers music related news, videos, songs, and other music-related features.
{"named_entities":
["Radio City", "India", "3 July 2001", "Hindi", "English", "May 2008", "PlanetRadiocity.com"]
}

Demonstration output:
{"triples": [
["Radio City", "located in", "India"],
["Radio City", "is", "private FM radio station"],
["Radio City", "started on", "3 July 2001"],
["Radio City", "plays songs in", "Hindi"],
["Radio City", "plays songs in", "English"],
["Radio City", "forayed into", "New Media"],
["Radio City", "launched", "PlanetRadiocity.com"],
["PlanetRadiocity.com", "launched in", "May 2008"],
["PlanetRadiocity.com", "is", "music portal"],
["PlanetRadiocity.com", "offers", "news"],
["PlanetRadiocity.com", "offers", "videos"],
["PlanetRadiocity.com", "offers", "songs"]
] }

Figure 6: Prompts and demonstrations for triple extraction

Fact Filtering

System prompt:
Your input fields are: question, which denotes the query for retrieval, and fact_before_filter, which denotes the candidate facts to be filtered. The output field is fact_after_filter, which stores the filtered facts in JSON format. All interactions follow the structure: question, fact_before_filter, and fact_after_filter. The output must be parseable according to the JSON schema: {"fact": [[subject, predicate, object], ...]}
The model should select up to 12 facts that are most relevant to the question. Prefer keeping complementary bridge facts, entity-linking facts, and answer-bearing facts when they may support multi-hop reasoning. If no candidate fact is relevant, it should return {"fact": []}. The model must only use facts from the candidate list and must not generate new facts.

Demonstration 1:
Question: Are Imperial River (Florida) and Amaradia (Dolj) both located in the same country?
Fact Before Filter: {"fact": [
["imperial river", "is located in", "florida"],
["imperial river", "is a river in", "united states"],
["imperial river", "may refer to", "south america"],
["amaradia", "flows through", "roia de amaradia"],
["imperial river", "may refer to", "united states"]
]}
Fact After Filter: {"fact": [
["imperial river", "is located in", "florida"],
["imperial river", "is a river in", "united states"],
["amaradia", "flows through", "roia de amaradia"]
]}

Demonstration 2:
Question: When is the director of film The Ancestor's birthday?
Fact Before Filter: {"fact": [
["jean jacques annaud", "born on", "1 october 1943"],
["sui hark", "born on", "15 february 1950"],
["pablo trapero", "born on", "4 october 1971"],
["the ancestor", "directed by", "guido brignone"],
["benh zeitlin", "born on", "october 14 1982"]
]}
Fact After Filter: {"fact": [["the ancestor", "directed by", "guido brignone"]]}

Figure 7: Fact filtering prompt and demonstrations