Breaking the Static Graph: Context-Aware Traversal for Robust Retrieval-Augmented Generation

Anonymous ACL submission

Abstract

Recent advances in Retrieval-Augmented Generation (RAG) have shifted from simple vector similarity to structure-aware approaches like HippoRAG, which leverage Knowledge Graphs (KGs) and Personalized PageRank (PPR) to capture multi-hop dependencies. However, these methods suffer from a "Static Graph Fallacy": they rely on fixed transition probabilities determined during indexing. This rigidity ignores the query-dependent nature of edge relevance, causing semantic drift where random walks are diverted into high-degree "hub" nodes before reaching critical downstream evidence. Consequently, models often achieve high partial recall but fail to retrieve the complete evidence chain required for multi-hop queries. To address this, we propose CatRAG, Context-Aware Traversal for robust RAG, a framework that builds on the HippoRAG 2 architecture and transforms the static KG into athe query-adaptive navigation structure. We introduce a multi-faceted framework to steer the random walk: **(1) Symbolic Anchoring**, which injects weak entity constraints to regularize the random walk; **(2) Query-Aware Dynamic Edge Weighting**, which dynamically modulates graph structure, to prune irrelevant paths while amplifying those aligned with the query's intent; and **(3) Key-Fact Passage Weight Enhancement**, cost-efficient bias that structurally anchors the random walk to likely evidence. Experiments across four multi-hop benchmarks demonstrate that CatRAG consistently outperforms state-of-the-art baselines. Our analysis reveals that while standard Recall metrics show modest gains, CatRAG achieves substantial improvements in reasoning completeness—the capacity to recover entire evidence path without gaps. These results reveal that our approach effectively bridges the gap between retrieving partial context and enabling fully grounded reasoning.

1 Introduction

Large Language Models (LLMs) have demonstrated transformative capabilities across a spectrum of natural language tasks, ranging from creative composition to complex code generation (Joel et al., 2025; Li et al., 2024; Ren et al., 2024; Touvron et al., 2023; Brown et al., 2020). Despite these advances, the widespread deployment of LLMs still remains restricted by hallucinations(Xu et al., 2025; Liu et al., 2024) in response generation, often caused by outdated training data or lack of domain-specific knowledge, resulting in seemingly plausible but actually incorrect content. Retrieval-Augmented Generation (RAG) (Gao et al., 2024; Fan et al., 2024) has emerged as a feasible solution to mitigate the issue, which incorporates external, reliable documents within LLM prompts for response generation.

Standard dense retrieval methods, which select document chunks based on semantic similarity (Izacard et al., 2022a), frequently fail in multi-hop reasoning scenarios when the answer relies on connecting disjoint facts. To overcome this limitation, recent research has shifted towards Structure-Aware RAG, which organizes information into hierarchical trees (Sarthi et al., 2024) or global knowledge graphs (Guo et al., 2025) to capture long-range dependencies. Among these, the HippoRAG framework (Gutiérrez et al., 2024, 2025) distinguishes itself by leveraging Personalized PageRank (PPR) over Knowledge Graphs. HippoRAG simulates neurobiological memory mechanism, enabling deeper and more efficient knowledge integration that vector similarity alone cannot resolve.

However, a critical bottleneck remains in these graph-based paradigms: reliance on a static graph structure. In standard HippoRAG, the transition matrix guiding the Random Walk is fixed during indexing, determined solely by structural properties or a priori semantic similarity. This rigidity

imposes two limitations. First, edge relevance is treated as context-independent. Consider the query: "*Which university did Marie Curie's doctoral advisor attend?*" This requires a precise two-step traversal: *Marie Curie* → *Gabriel Lippmann* (Advisor) → *École Normale Supérieure* (University). Yet, in a static graph, generic edges like *Marie Curie* → *Radioactivity* often possess dominant weights. Consequently, the random walk often suffers from semantic drift: it effectively retrieves the initial entity but is statistically diverted into irrelevant clusters before reaching the second-hop evidence. This results in a common failure mode where retrieval metrics (like Recall) appear high due to partial matches, yet the reasoning chain is broken. Second, traversal is susceptible to the "hub node" problem, where high-degree entities (e.g., *Nobel Prize*, *French*) act as semantic sinks, disproportionately diluting the probability mass and causing the retrieval to drift into irrelevant documents.

To mitigate the constraints imposed by static graph topologies, we develop CatRAG (**C**ontext-**A**ware **T**raversal for robust **RAG**). This framework extends the HippoRAG 2 (Gutiérrez et al., 2025) paradigm by integrating a novel optimization layer tailored for context-driven navigation. First, we introduce **Symbolic Anchoring**. By injecting explicitly recognized entities as weak topological anchors, we constrain the starting distribution to prevent immediate drift into generic hubs. Second, we introduce **Query-Aware Dynamic Edge Weighting**. By employing an LLM to assess the relevance of outgoing edges from seed entities, we dynamically modulate the graph edge weight, effectively pruning irrelevant paths while amplifying those aligned with the query's intent. Third, we propose **Key-Fact Passage Weight Enhancement**, a cost-efficient method to structurally anchor the random walk to documents containing verified evidentiary triples. It guides the random walk to documents that provide distinct evidence, rather than those containing only superficial mentions of seed entities.

We evaluate CatRAG across multiple multi-hop benchmarks. Results demonstrate that while our approach yields consistent gains in standard retrieval metrics, it achieves a significant breakthrough in reasoning sufficiency. CatRAG substantially improves Full Chain Retrieval—the ability to retrieve complete evidence chains—confirming that dynamic graph steering effectively mitigates semantic drift where static baselines fail.

2 Related Work

2.1 Dense Retriever

The foundational paradigm for RAG matches queries and documents in a shared vector space, evolving from probabilistic term-matching (Robertson and Walker, 1994) and dense bi-encoders (Izacard et al., 2022b) to granular late-interaction mechanisms (Santhanam et al., 2022). Recently, the field has shifted toward Large Embedding Models like E5-Mistral (Wang et al., 2024), NV-Embed (Lee et al., 2025) and GritLM (Muennighoff et al., 2025), which repurpose LLMs to achieve superior benchmark performance (Muennighoff et al., 2022). However, these models remain constrained by the static nature of vector similarity. By compressing complex reasoning paths into a single geometrical proximity, they lack explicit multi-hop traversal mechanisms and frequently fail when queries and evidence are connected solely through intermediate bridge entities (Gutiérrez et al., 2024).

2.2 Structure-Aware RAG

To transcend the limitations of flat vector spaces, recent works integrate explicit structural priors. Hierarchical approaches like RAPTOR (Sarthi et al., 2024) organize text into recursive trees, while graph-based frameworks such as GraphRAG (Edge et al., 2025) and LightRAG (Guo et al., 2025) leverage Knowledge Graphs to traverse entity relationships. The state-of-the-art neuro-symbolic approach, HippoRAG (Gutiérrez et al., 2024) and its successor, HippoRAG 2 (Gutiérrez et al., 2025), simulates associative memory via PPR to link disparate facts. However, these methods suffer from the "Static Graph Fallacy": edge weights are fixed during indexing and cannot adapt to query-specific intent. This rigidity causes semantic drift, where high-degree "hub" nodes disproportionately dominate traversal probabilities, leading to the retrieval of structurally connected but contextually irrelevant paths.

2.3 Dynamic & Adaptive Retrieval

To address static retrieval limitations, iterative frameworks like IRCoT (Trivedi et al., 2023) and Self-RAG (Asai et al., 2023), or agentic systems such as PRISM (Nahid and Rafiei, 2025) and FAIR-RAG (Asl et al., 2025), employ multi-step loops to refine search queries. While effective, these methods incur high latency and computational costs by requiring repeated LLM calls for multiple

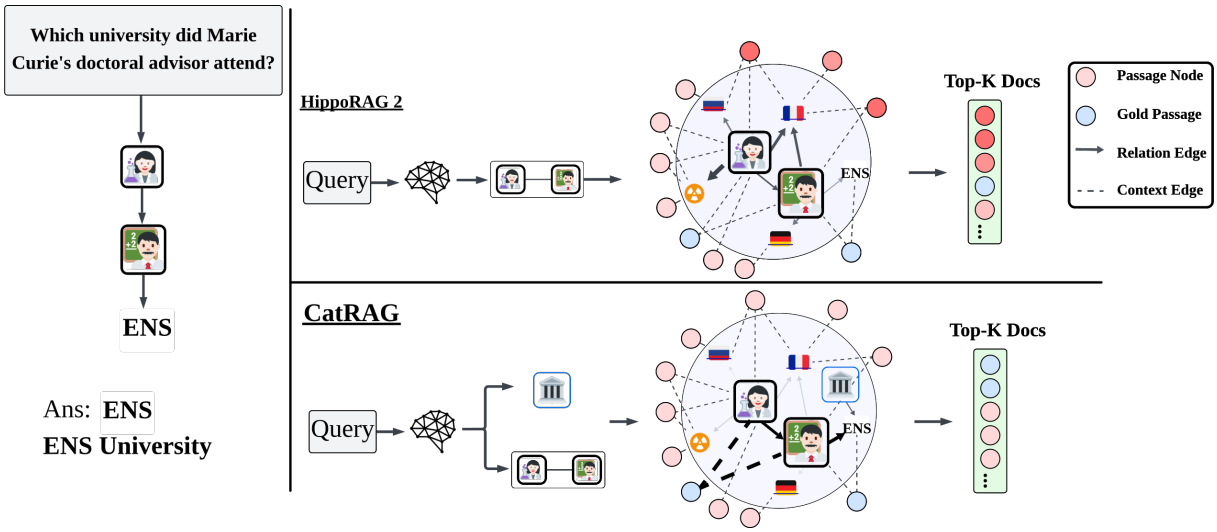

Figure 1: **Comparison of graph traversal between HippoRAG 2 and CatRAG.** We illustrate the retrieval process for the multi-hop query "Which university did Marie Curie's doctoral advisor attend?". In HippoRAG 2 (top), the static graph structure causes semantic drift; probability mass is diverted to high-weight generic edges (e.g., *Marie Curie* → *Radioactivity*), missing the downstream evidence *ENS*. CatRAG (bottom) prevents this by applying (1) Symbolic Anchoring, injecting "University" as a weak seed, (2) Query-Aware Dynamic Edge Weighting amplifying relevant paths (e.g., *Attend in ENS*) while pruning irrelevant ones, and (3) Key-Fact Passage Weight Enhancement to strength, boosting relevant context edge. This steers the random walk to successfully retrieve the complete evidence chain for *ENS*

searches. CatRAG instead introduces a "one-shot" context-aware graph modification that dynamically re-weights edges before traversal. Unlike iterative cycles, our approach maintains the efficiency of a single retrieval pass, effectively combining the reasoning precision of adaptive methods with the speed and structural integrity of graph-based retrieval.

3 Methodology

In this section, we propose three mechanisms to optimize HippoRAG 2's retrieval on a knowledge graph: *Symbolic Anchoring*, *Query-Aware Dynamic Edge Weighting* and *Key-Fact Passage Weight Enhancement*, also present in Figure1.

3.1 Preliminaries

We build our approach upon the graph structure defined in HippoRAG 2. The knowledge base is modeled as a directed graph $G = (V, E)$. The node set $V = V_E \cup V_P$ consists of entity phrases V_E and passage nodes V_P.

The edge set E is composed of three distinct types of semantic connections:

- **Relation Edges (E_{rel}):** Edges between entity nodes ($u, v \in V_E$) derived from OpenIE triples.
- **Synonym Edges (E_{syn}):** Edges connecting entity nodes with high vector similarity, capturing linguistic variations of the same concept.
- **Context Edges (E_{ctx}):** Edges linking a passage node $p \in V_P$ to the entity nodes $e \in V_E$ contained within it.

We adopt the Personalized PageRank (PPR) algorithm to model the retrieval process. The probability distribution over nodes at step k is updated as:

$$\mathbf{v}^{(k+1)} = (1 - d) \cdot \mathbf{e}_s + d \cdot \mathbf{v}^{(k)}\mathbf{T} \qquad (1)$$

where $\mathbf{e}_s$ is the personalized probability distribution over seed nodes, and $\mathbf{T}$ is the row-normalized transition matrix. In the standard framework, $\mathbf{T}$ is static. Our work focuses on dynamically refining $\mathbf{T}$ into a query-specific transition matrix $\hat{\mathbf{T}}_q$ to better capture the reasoning requirements of the user query.

3.2 Symbolic Anchoring

While the "Query to Triple" retrieval in HippoRAG 2 effectively captures implicit semantic cues, we argue that relying solely on dense vector alignment leaves the graph traversal susceptible to semantic drift. Without explicit constraints, the PPR propagation can easily be siphoned into high-degree

"hub" nodes that have high similarity but lack precise relevance to the query. To mitigate this, we introduce **Symbolic Anchoring**, a regularization strategy that grounds the stochastic walk using explicit query constraints.

Rather than treating NER as an alternative retrieval path, we utilize extracted entities as strictly auxiliary *topological anchors*. We extract a set of entities and inject them as weak seed, assigning reset probabilities for retrieval. We assign these symbolic anchors with small reset probabilities ϵ, to ensure that their influence is subordinate to the initial entity from contextual triples.

This weak seeding serves a specific regulatory function: it aligns the PPR propagation with the query's intent. By placing a non-zero probability on the exact named entities mentioned in the query, we create a gravitational pull that resists the diffusion of probability mass into generic graph hubs. Even as the random walk explores the neighborhood defined by the static graph, these weak anchors ensure the traversal recurrently grounded to the specific entities in the query, effectively suppressing semantic drift. As a secondary benefit, this mechanism naturally balance the system's capability: it retains the triplet-based strength in interpreting implicit clues while ensuring robust coverage for containing explicit entity mentions.

3.3 Query-Aware Dynamic Edge Weighting

Current graph-based RAG models rely on a static transition matrix T, where transition probabilities are fixed during indexing. We argue that this rigidity induces stochastic drift: without query-specific guidance, the random walk indiscriminately diffuses probability mass into high-degree "hub" nodes that are structurally prominent but semantically irrelevant. To mitigate this, we approximate a query-conditional transition matrix $\hat{T}_q$, concentrating the random walk on edges that maximize information gain. We implement a **two-stage coarse-to-fine strategy** to dynamically modulate the weights of **relation edges** (E_{rel}).

3.3.1 Adaptive Entity Contextualization

To assist the LLM in evaluating the relevance of a transition from seed u to neighbor v, we augment the prompt with a semantic summary of v. Since providing all connected facts for dense nodes is computationally intractable, we employ a conditional summarization strategy. Let $\mathcal{F}(v)$ be the set of fact triples connected to entity node v. We define the context content $C(v)$ as:

$$C(v) = \begin{cases} \text{Summary}(\mathcal{F}(v)) & \text{if } |\mathcal{F}(v)| > \tau \\ \text{Concat}(\mathcal{F}(v)) & \text{otherwise} \end{cases} \quad (2)$$

where τ is a density threshold. For information-dense nodes ($|\mathcal{F}(v)| > \tau$), we generate a concise summary; for sparse nodes, we use raw triples. This hybrid approach balances context completeness with token efficiency.

3.3.2 Stage I: Coarse-Grained Candidate Pruning

Evaluating the semantic relevance of every edge using an LLM is computationally prohibitive. Therefore, we first apply a topological filter to constrain the search space to the most plausible local neighborhoods. We define two hyperparameters: the maximum number of seed entities N_{seed} and the maximum number of edges per seed K_{edge} for fine-grained alignment. First, we select the top-N_{seed} entity nodes based on their initial reset probabilities (derived from the dense retrieval alignment). Let u be such a selected seed. For the seed phrase u within top-N_{seed}, if the number of outgoing relation edges exceeds a threshold K_{edge}, we prune its outgoing edges by prioritizing the top-K_{edge} neighbors based on the vector similarity between the query embedding and fact embeddings of relation edges. Neighbors $v \notin \mathcal{N}_{top}(u)$ are bypassed by the scoring module and assigned a minimal *Weak* weight. This step acts as a low-pass structural filter, discarding statistically improbable paths before the intensive semantic scoring.

3.3.3 Stage II: Fine-Grained Semantic Probability Alignment

In the second stage, we refine the weights of the surviving edges in $\mathcal{N}_{top}(u)$ to minimize semantic drift. While vector similarity (Stage I) captures general relatedness, it often fails to distinguish between generic associations and precise evidentiary links. We employ a Large Language Model (LLM) as a discrete approximation of the conditional transition probability $P(v|u, q)$. The LLM evaluates the necessity of the transition $u \to v$ given the query q and the neighbor's summary $C(v)$. We prompt the model to classify the relationship into discrete tiers $\mathcal{L} \in \{\text{Irrelevant, Weak, High, Direct}\}$. We define a mapping function $\phi : \mathcal{L} \to \mathbb{R}^+$ to project these judgments into scalar weights. The updated dynamic weight $\hat{w}_{uv}$ is computed as:

$$\hat{w}_{uv} = \phi\big(\text{LLM}(q, u, v, C(v))\big) \cdot w_{uv}^{(static)}$$

This modulation is asymmetric, applied only to forward edges originating from the seed set. By suppressing irrelevant edges and amplifying critical ones, we actively steer the PPR propagation, ensuring the traversal tunnels through the graph along the query's intent rather than diffusing into topological sinks.

3.4 Key-Fact Passage Weight Enhancement

In the directed graph setting, a seed entity node $u \in V_E$ may connect to multiple passage nodes V_P via context edges. We aim to bias the walk towards passages containing "Key Facts"—fact triplets that were explicitly identified and filtered during the filtering Recognition memory filtering proposed in HippoRAG 2.

Let $\mathcal{T}_{seed}$ be the set of verified seed triples. We identify a "Key Fact" connection if the edge E_{ctx} from seed entity u to passage p is supported by a triple in $\mathcal{T}_{seed}$. We enhance the weight of such edges:

$$\hat{w}_{up} = w_{up} \cdot (1 + \beta \cdot \mathbb{I}(u, p \in \mathcal{T}_{seed})) \quad (3)$$

where β is a boost factor and $\mathbb{I}(\cdot)$ is an indicator function.

This enhancement prioritizes passages providing evidentiary support. Unlike the previous module which requires LLM inference, the Key-Fact Enhancement is a purely algorithmic adjustment based on triple-matching. It incurs zero additional token cost and negligible latency, making it a highly efficient approach to guide the random walk.

3.5 Unified Retrieval Process

We integrate Symbolic Anchoring, Dynamic Edge Weighting, and Passage Enhancement to construct a query-adapted graph. Standard PPR (Eq. 1) is executed on this refined structure. The resulting stationary distribution of PPR provides the final passage ranking, prioritizing nodes reachable via semantically relevant reasoning paths.

4 Experimental Setup

4.1 Baselines

We evaluate CatRAG against a comprehensive suite of baselines spanning two paradigms: standard RAG with retrieval methods, and structure-aware RAG.

For standard retrieval comparisons, we employ several strong and widely used retrieval model,including **BM25** (Robertson and Walker, 1994), **Contriever** (Izacard et al., 2022b), **GTR** (Ni et al., 2022), **text-embedding-3-small** [1] model, to represent standard embedding-based approaches.

Our primary comparison targets structure-aware RAG frameworks. We compare against **RAPTOR** (Sarthi et al., 2024), which constructs a recursive tree structure for hierarchical summarization. Crucially, our main baseline is **HippoRAG 2** (Gutiérrez et al., 2025),the state-of-the-art in graph-based neuro-symbolic retrieval. We omit the original HippoRAG (Gutiérrez et al., 2024) from our evaluation, as HippoRAG 2 has demonstrated that it consistently outperforms its predecessor; thus, HippoRAG 2 serves as the most rigorous and relevant control. Since our method is instantiated upon the HippoRAG 2 architecture, this comparison isolates the specific gains attributable to our proposed dynamic edge weighting and passage enhancement mechanisms.

Table 1: Dataset statistics.

Dataset	MuSiQue	2Wiki	HotpotQA	HoVer
# of Queries	1,000	1,000	1,000	1,000
# of Passages	11,656	6,119	9,811	9,440

4.2 Datasets

To evaluate the ability of *CatRAG* to maintain precise retrieval in multi-hop scenarios, we conduct experiments on four benchmarks across two challenge types: **Multi-hop QA** and **Multi-hop Fact Verification**. We summarize the key statistics of these datasets in Table 1.

Multi-hop QA. We conduct experiments on **MuSiQue** (Trivedi et al., 2022), **2WikiMultiHopQA** (Ho et al., 2020), and **HotpotQA** (Yang et al., 2018). These datasets require the system to reason over multiple passages to derive an answer. To ensure a fair comparison and reproducibility, we utilize the subsets defined in prior work (Gutiérrez et al., 2024), which sampled 1,000 queries randomly and collected all candidate passages (including supporting and distractor passages) to form a corpus for each dataset. Crucially, **HotpotQA** and **2WikiMultiHopQA** are composed of 2-hop queries, while **MuSiQue** presents more challenging questions requiring 2 to 4 hops. This gradation is essential for our analysis.

[1] https://platform.openai.com/docs/models/text-embedding-3-small

Multi-hop Fact Verification. We extend our evaluation to the **HoVer** dataset (Jiang et al., 2020) to test the robustness of our model in a claim verification setting. HoVer is adapted from HotpotQA but increases reasoning complexity by substituting named entities in the original claims with details from linked Wikipedia articles, thereby extending the reasoning chain to 3 and 4 hops. This substitution process creates deep, fragile reasoning chains where a single missed retrieval step results in failure. Following the protocol in HippoRAG, we randomly sample 1,000 claims from the dataset (specifically 3 and 4 hops) and form the retrieval corpus by collecting all candidate passages (supporting evidence and distractors) associated with the original HotpotQA lineage questions of our selected claims.

4.3 Metrics

We report Recall@5 for standard retrieval evaluation and F1 for downstream QA. However, these aggregate metrics often mask incomplete reasoning, as models may retrieve partial evidence or guess correct answers without grounding. To rigorously assess reasoning integrity, we introduce **Full Chain Retrieval (FCR)**, defined as the percentage of queries where the retrieved context contains the *entire* set of gold supporting documents. Furthermore, we report the **Joint Success Rate (JSR)**, which counts a query as successful only if the system achieves FCR *and* the generated response contains the correct answer. This metric conceptually aligned with the strict evaluation established in the FEVER Shared Task (Thorne et al., 2018) and HoVer (Jiang et al., 2020), ensuring that accurate answer stem from complete evidentiary support rather than hallucinated or accidental correctness.

4.4 Implementation Details

We implement CatRAG upon the HippoRAG 2 architecture, using GPT-4o-mini[2] as the backbone for all LLM components and text-embedding-3-small as the retriever. While newer open-weight models like NV-Embed-v2 (Lee et al., 2025) show strong performance, our primary objective is to isolate the topological gains provided by the CatRAG mechanism from the raw semantic capacity of the underlying encoder. For fair comparison, all structure-augmented baselines are reproduced using the same extractor and retriever. Downstream responses are generated by Llama-3.3-70B-Instruct using the top-5 retrieved passages. Key hyperparameters include: symbolic anchor reset probability $\epsilon = 0.2$ (weighted by inverse passage count $|P_i|^{-1}$), boost factor $\beta = 2.5$, dynamic weighting limits $N_{seed} = 5$ and $K_{edge} = 15$. More implementation details and hyperparameters are provided in Appendix A.1.

[2]https://platform.openai.com/docs/models/gpt-4o-mini

5 Results

Standard Retrieval and QA. Table 2 and Table 3 demonstrate that CatRAG consistently outperforms all baselines across standard metrics. On the complex MuSiQue dataset (2–4 hops), CatRAG achieves a Recall@5 of 64.9%, surpassing the dense retriever text-embedding-3-small by a substantial 8.1% margin and confirming the necessity of structure-aware methods.

Compared to the state-of-the-art static baseline, HippoRAG 2 across all benchmarks, CatRAG raises Recall@5 to 89.5% on HotpotQA and 76.8% on HoVer. This retrieval quality directly translates to downstream performance, where CatRAG yields the highest F1 scores across all datasets (e.g., 45.0% on MuSiQue), validating that query-conditional edge weighting surfaces relevant evidence without disrupting structural integrity.

Strict Reasoning Completeness Evaluation. While standard metrics indicate general relevance, they often mask a critical failure mode in multi-hop retrieval: the loss of intermediate "bridge" documents that connect disjoint facts. To assess the recovery of the full evidence paths, we evaluate FCR and JSR in Table 4. CatRAG effectively mitigates probability dilution, achieving an FCR of 34.6% compared to 30.5% for HippoRAG 2. The gain is most pronounced on HoVer, where precise 3–4 hop claim verification is required. CatRAG improves JSR to 31.1%, a relative gain of 18.7% over the HippoRAG2. These results confirm that our dynamic steering successfully anchors the traversal to the specific bridge documents required for grounded reasoning.

5.1 Ablation Study

We conduct an ablation experiment, to isolate the contributions of Symbolic Anchoring, Query-Aware Dynamic Edge Weighting (E_{rel}), and Key-Fact Passage Weight Enhancement, with results summarized in Table 5. First, the removal of Symbolic Anchoring precipitates a consistent perfor-

Table 2: **Retrieval Performance (Recall@5).** Retrieval performance on multi-hop QA and fact verification datasets.

Method	MuSiQue	2Wiki	HotpotQA	HoVer
Standard Retrieval				
BM25	31.6	52.1	64.8	50.8
Contriever	46.6	57.5	75.3	62.3
GTR (T5-base)	49.1	67.9	73.9	55.6
text-embedding-3-small	55.4	70.8	81.3	65.7
Structure-Aware RAG				
RAPTOR	53.3	69.8	79.5	62.4
HippoRAG 2	61.4	85.9	87.1	71.2
CatRAG	**64.9**	**87.0**	**89.5**	**76.8**

Table 3: **Downstream QA Performance.** QA performance on multi-hop QA and fact verification datasets using Llama-3.3-70B-Instruct as the QA reader. We report F1 for QA datasets, accuracy for the HoVer dataset. * denotes the results from (Gutiérrez et al., 2025).

Method	MuSiQue	2Wiki	HotpotQA	HoVer
Standard Retrieval				
None	26.1*	42.8*	47.3*	–
BM25	22.9	39.9	54.1	61.4
Contriever	31.3	41.9	62.3	66.0
GTR (T5-base)	34.6	52.8	62.8	62.7
text-embedding-3-small	36.1	56.9	64.6	64.2
Structure-Aware RAG				
RAPTOR	36.0	56.7	64.4	65.3
HippoRAG 2	43.2	68.1	69.4	67.2
CatRAG	**45.0**	**69.7**	**71.4**	**69.0**

Table 4: **Reasoning Completeness Evaluation.** We evaluation the **Full Chain Retrieval** (**FCR**) and **Joint Success Rate** (**JSR**) on multi-hop QA and fact verification datasets.

Method	MuSiQue	2Wiki	HotpotQA	HoVer
Standard Retrieval				
BM25	6.4/4.5	20.5/19.1	38.3/26.1	8.2/6.3
Contriever	11.3/8.3	27.2/23.9	54.1/37.6	18.1/13.8
GTR (T5-base)	15.0/11.5	35.8/31.3	53.0/37.9	10.3/6.8
text-embedding-3-small	21.1/13.8	41.6/34.9	64.9/46.3	22.1/16.0
Structure-Aware RAG				
RAPTOR	19.6/13.2	40.1/34.2	61.4/44.2	18.6/13.7
HippoRAG 2	30.5/21.5	66.1/53.0	75.5/53.4	34.8/26.2
CatRAG	**34.6/24.3**	**67.6/55.0**	**80.4/56.8**	**42.5/31.1**

mance degradation, most notably a 3.2% drop on HoVer. This confirms that injecting extracted entities as weak topological anchors is critical for mitigating semantic drift. Second, excluding E_{rel} weighting results in significant losses across all benchmarks, confirming that dynamically pruning irrelevant semantic branches is foundational to mitigating drift. Finally, we observe that Key-Fact Enhancement provides consistent gains across unstructured datasets (HotpotQA, MuSiQue, HoVer) where evidence is buried in dense text. On the highly structured dataset 2WikiMultiHopQA, this heuristic introduces slight noise, leading to a minor performance regression . However, given that real-world RAG scenarios involve messy, unstructured corpora, we prioritize the gains on the unstructured datasets.

6 Discussion

6.1 Impact on Multi-Hop Dependency: Mitigating Hub Bias

A fundamental limitation of static graph retrieval is Hub Bias (or degree centrality bias). In standard formulations like HippoRAG 2, transition probabilities are determined by static structural proper-

Table 5: **Ablations.** We report passage recall@5 on multi-hop benchmarks using several alternatives to our final design in dynamic update.

	MuSiQue	2Wiki	HotpotQA	HoVer
CatRAG	**64.9**	87.0	**89.5**	**76.8**
w/o Symbolic anchor	63.0	86.1	88.6	73.6
w/o E_{rel} weighting	63.2	85.6	88.1	75.0
w/o Passage Enhance	64.7	**88.4**	89.0	76.6

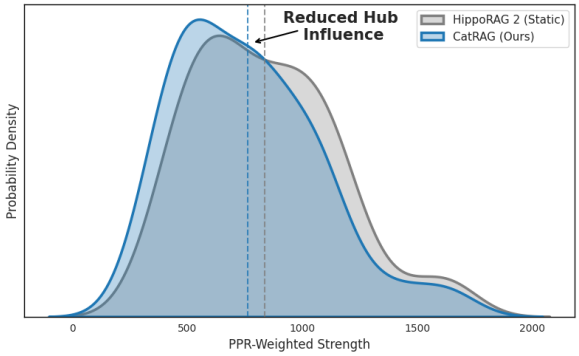

Figure 2: **Distribution of PPR-Weighted Node Strength ($\mathcal{S}_{ppr}$).** Comparison of the HippoRAG 2 versus CatRAG. The distribution for CatRAG is shifted to the left, indicating a reduction in the retrieval of high-degree "Hub" nodes. The dashed lines represent the mean $\mathcal{S}_{ppr}$ for each method.

ties. Consequently, random walks disproportionately converge on high-degree nodes (e.g., generic entities like 'United States' or 'Song'), which act as "topological sinks". We hypothesize that this structural noise disrupts multi-hop dependency by diverting the retrieval path away from the specific "bridge" entities required to connect disjoint facts.

Quantifying Semantic Drift. To assess whether our proposed framework mitigates this drift, we analyzed the topological properties of the top-10 retrieved entity nodes after PPR across 100 randomly sampled queries from MuSiQue. We introduce *PPR-Weighted Strength* ($\mathcal{S}_{ppr}$) to measure the effective structural prominence of the retrieved context:

$$\mathcal{S}_{ppr}(q) = \sum_{v \in \mathcal{V}_{top}} \hat{p}(v) \cdot \text{Strength}(v) \qquad (4)$$

where $\hat{p}(v)$ is the PPR probability mass of node v re-normalized over the retrieved set $\mathcal{V}_{top}$ (i.e., $\sum \hat{p}(v) = 1$), and $\text{Strength}(v)$ is the weighted degree of the node. A higher $\mathcal{S}_{ppr}$ indicates that the PPR result is more reliant on generic, high-connectivity nodes.

Mitigation of Hub Bias. As illustrated in Figure 2, CatRAG exhibits a systematic structural shift toward specificity. The distribution of PPR-Weighted Strength for CatRAG is distinctively shifted to the left compared to the static baseline HippoRAG. CatRAG reduces the Mean PPR-Weighted Strength from 837.0 to 761.7. Furthermore, we quantified the probability mass allocated to "Super Hubs" (nodes in the top 1% of weighted degree). While the baseline allocates 45.7% of its probability mass to these generic hubs, our method significantly reduces this to 42.5%.

Correlation with Reasoning Completeness. This structural correction directly explains the improvements in reasoning integrity observed in Table 4. While the relative reduction in hub mass (7%) may appear moderate, it represents a critical redistribution of probability mass away from topological distractors and toward specific bridge entities. This aligns with our results on the HoVer dataset, where avoiding generic associations is crucial for verification; specifically, this structural enhancement enables the 11% relative improvement in JSR. By structurally decoupling prominence from relevance, CatRAG ensures that the retrieved context preserves the complete dependency chain, bridging the gap between partial recall and grounded reasoning.

7 Conclusion

We identify and address the "Static Graph Fallacy" inherent in current structure-aware RAG systems, where fixed transition probabilities predispose retrieval to semantic drift and prevent the recovery of complete evidence chains. We propose CatRAG, a framework that transforms the Knowledge Graph Traversal into a context-aware navigation structure. Experiment across multi-hop benchmarks demonstrate that CatRAG consistently outperforms baselines, including HippoRAG 2, while significantly reducing the bias of high-degree hub nodes. Our analysis reveals that these topological adjustments yield substantial improvements in reasoning completeness, effectively bridging the gap between retrieving partial context and enabling fully grounded, multi-hop reasoning.

Limitations

While CatRAG significantly enhances reasoning completeness, it introduces certain trade-offs regarding efficiency. First, the mechanism for query-

aware dynamic edge weighting requires run-time LLM inference to assess semantic relevance, which incurs additional computational overhead and latency compared to purely static graph traversals. Although we mitigate this via coarse-grained pruning, the approach remains more computationally intensive than standard dense retrieval. Furthermore, our experimental evaluation intentionally utilized standard embedding models (text-embedding-3-small) rather than larger, state-of-the-art embedding models to strictly isolate the topological gains provided by our framework from the raw semantic capacity of the encoder. Consequently, while our results demonstrate the superiority of dynamic traversal, the absolute performance ceiling of CatRAG could potentially be further elevated by integrating these larger foundational models in future work. Due to proprietary data protection policies, the source code cannot be publicly released. To mitigate this, we have provided full hyperparameter tables in to facilitate reimplementation.

8 Ethical considerations

This study utilizes four publicly available benchmark datasets, MuSiQue, 2WikiMultiHopQA, HotpotQA, and HoVer, which are standard in the field. These datasets are derived from Wikipedia/Wikidata sources and may therefore contain publicly available information about real people and may incidentally include sensitive topics; however, we did not collect new personal data or interact with human participants. Regarding computational resources and model access, we utilized GPT-4o mini and text-embedding-3-small via the Microsoft Azure API, and accessed Llama-3.3-70B-Instruct through the OpenRouter API.

According with AI Assistance policies, we acknowledge that we used generative AI tools to assist with code implementation and language polishing. All scientific content and results were verified by the authors.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *Preprint*, arXiv:2310.11511.

Mohammad Aghajani Asl, Majid Asgari-Bidhendi, and Behrooz Minaei-Bidgoli. 2025. Fair-rag: Faithful adaptive iterative refinement for retrieval-augmented generation. *Preprint*, arXiv:2510.22344.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2025. From local to global: A graph rag approach to query-focused summarization. *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*, KDD '24, page 6491–6501, New York, NY, USA. Association for Computing Machinery.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2025. Lightrag: Simple and fast retrieval-augmented generation. *Preprint*, arXiv:2410.05779.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. Hipporag: Neurobiologically inspired long-term memory for large language models. *Preprint*, arXiv:2405.14831.

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. *Preprint*, arXiv:2502.14802.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *Preprint*, arXiv:2011.01060.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022a. Unsupervised dense information retrieval with contrastive learning. *Preprint*, arXiv:2112.09118.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022b. Unsupervised dense information retrieval with contrastive learning. *Preprint*, arXiv:2112.09118.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. Hover: A dataset for many-hop fact extraction and claim verification. *Preprint*, arXiv:2011.03088.

Sathvik Joel, Jie Wu, and Fatemeh Fard. 2025. A survey on llm-based code generation for low-resource and domain-specific programming languages. *ACM Trans. Softw. Eng. Methodol.* Just Accepted.

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. *Preprint*, arXiv:2405.17428.

Jiawei Li, Yizhe Yang, Yu Bai, Xiaofeng Zhou, Yinghao Li, Huashan Sun, Yuhang Liu, Xingpeng Si, Yuhao Ye, Yixiao Wu, Yiguan Lin, Bin Xu, Bowen Ren, Chong Feng, Yang Gao, and Heyan Huang. 2024. Fundamental capabilities of large language models and their applications in domain scenarios: A survey. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11116–11141, Bangkok, Thailand. Association for Computational Linguistics.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024. A survey on hallucination in large vision-language models. *Preprint*, arXiv:2402.00253.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2025. Generative representational instruction tuning. *Preprint*, arXiv:2402.09906.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Md Mahadi Hasan Nahid and Davood Rafiei. 2025. Prism: Agentic retrieval with llms for multi-hop question answering. *Preprint*, arXiv:2510.14278.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xubin Ren, Jiabin Tang, Dawei Yin, Nitesh Chawla, and Chao Huang. 2024. A survey of large language models for graphs. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 6616–6626. ACM.

S. E. Robertson and S. Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR '94*, pages 232–241, London. Springer London.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. Colbertv2: Effective and efficient retrieval via lightweight late interaction. *Preprint*, arXiv:2112.01488.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. 2024. RAPTOR: Recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations*.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Preprint*, arXiv:2108.00573.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *Preprint*, arXiv:2212.10509.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving text embeddings with large language models. *Preprint*, arXiv:2401.00368.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2025. Hallucination is inevitable: An innate limitation of large language models. *Preprint*, arXiv:2401.11817.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *Preprint*, arXiv:1809.09600.

A Appendix

A.1 Implementation Details and Hyperparameters

We summarize the core hyperparameters for CatRAG in Table 6. To ensure fair comparison, we maintain the QA prompts established in the HippoRAG 2 benchmark (Gutiérrez et al., 2025).

Dynamic Edge Scoring Schedule. To translate the LLM's semantic assessment into topological structure, we employ a tiered projection strategy. We define four distinct semantic tiers—Irrelevant, Weak, High, and Direct—and map the discrete

Table 6: **Hyperparameters for CatRAG.** Note that Synonym Edge weights are dynamic, scaled by their vector similarity, whereas the standard HippoRAG 2 framework uses raw vector similarity.

Parameter	**Value**
Synonym Similarity Threshold	0.8
Synonym Edge Weight	$2.0 \times$ Similarity
PPR Damping Factor (d)	0.5
LLM Temperature	0.0
Symbolic Anchor (ϵ)	0.2
Max Seed Nodes for scoring (N_{seed})	5
Max Pruning Edges for scoring (K_{edge})	15
Passage Boost Factor (β)	2.5

LLM scores $s \in \{0, \ldots, 10\}$ to specific weight intervals (Table 7). This non-linear mapping acts as a high-pass filter, strictly pruning noise (scores ≤ 3) while exponentially amplifying high-confidence evidence paths.

Table 7: **LLM Score to Edge Weight Projection.** Discrete relevance scores are mapped to weight intervals. Within the *Weak* and *High* tiers, weights are linearly interpolated.

Semantic Tier	**LLM Score** (s)	**Output Weight** $\phi(s)$
Irrelevant	$0-3$	0
Weak	$4-6$	$0.2-0.3$
High	$7-9$	$2.0-3.0$
Direct	10	5.0

B Prompts

Entity Summarization Prompt (Adaptive Entity Context)

```
— Task —
Generate a concise, entity-focused summary that captures the core identity and key relationships of a
given entity based on its associated fact triplets.

— Instructions —
1. **Input Format**: You will receive:
  - A ‘target_entity‘ (the entity being summarized)
  - A ‘fact_triplets‘ list in JSON format containing relationships where this entity appears

2. **Output Requirements**:
  - Focus on the **target entity** as the summary’s subject
  - Integrate ALL key relationships from the provided triplets
  - Explain **what the entity is** and **what it connects to** through its relationships
  - Maintain strict coherence and factual accuracy
  - Maximum length: 150 tokens
  - Language: English (preserve proper nouns in original form when needed)

3. **Content Guidelines**:
  - Start with the entity’s core identity/type
  - Group related relationships logically (e.g., all professional roles together)
  - Highlight notable connections to other significant entities
  - Avoid listing facts mechanically - synthesize into narrative form

— Example Structure —
[Entity Name] is a [core type/description] known for [key attributes]. It [main
relationships/activities] with entities such as [notable connections]...

[... One in-context learning examples ...]

— Input —
Target node: ${entity}
Fact Triplets: ${fact_triplets}
```

Table 8: Prompt for generating entity summaries.

Knowledge Graph Neighbor Scoring Prompt (Fine-Grained Semantic Probability Alignment)

```
You are a knowledge graph reasoning expert.  Score neighbor entities (0-10) on their utility for
answering a QUERY.

### Input Data
1. A user QUERY.
2. The CURRENT ENTITY node we are exploring.
3. A set of RETRIEVED FACTS (trusted evidence).
4. A list of NEIGHBORS, each with:
  - The specific LINKING TRIPLET(s) connecting the current entity to this neighbor.
  - A short summary of the neighbor information.

### Scoring Criteria
- **10 (Solution):** The neighbor IS the answer or contains it.
- **7-9 (Bridge):** Critical step in the reasoning chain (e.g., Subject -> Attribute).
- **4-6 (Weak):** Valid semantic link, but tangential to query intent.
- **0-3 (Noise):** Irrelevant, generic, or contradicts facts.

### Rules
1. **Trust Facts:** If a neighbor contradicts RETRIEVED FACTS, score 0.
2. **Output Format:** - ‘ID (Entity Name): Score‘ (if Score < 4)
  - ‘ID (Entity Name): Score | Concise reasoning‘ (if Score >= 4)
3. **Constraint:** You must copy the Entity Name exactly as it appears in the input.

[... Two in-context learning examples ...]

Output ONE line per neighbor: ‘ID (Entity Name): Score | (Reasoning if Score >= 4)‘
```

Table 9: The prompt for scoring neighbor nodes.