Adaptive Semantic-Causal Graph Integration for Robust Knowledge Graph RAG

Takanori Ugai[1,*], Takeshi Morita[2,1]

[1]National Institute of Advanced Industrial Science and Technology, 2-3-26 Aomi, Koto-ku, Tokyo 135-0064, Japan
[2]Aoyama Gakuin University, 5-10-1 Fuchinobe, Chuo-ku, Sagamihara-shi, Kanagawa, 252-5258, Japan

Abstract

Standard Retrieval-Augmented Generation (RAG) often relies on semantic similarity, which can retrieve relevant-looking passages without preserving the causal structure needed for reliable reasoning. For multi-hop questions, this may produce answers that are locally coherent but fail to connect entities or establish the direction of influence. We propose `CausalHippo`, an adaptive framework that combines semantic graph connectivity for broad exploration with causal path discovery for directional verification. Its three-layer controller uses dynamic weighting to balance these complementary signals, confidence-based switching to adjust retrieval emphasis when evidence is unreliable, and two-pass adaptation to revise the initial retrieval decision. We evaluate the framework on MuSiQue, Webis-CausalQA-22, and a novel Causal-Reasoning-QA benchmark, covering disconnected multi-hop reasoning, causal question answering, and distinctions between causal, non-causal, and insufficient evidence. The results show that `CausalHippo` does not uniformly outperform all baselines; its benefits depend on the task and evaluation objective. Compared with graph-only retrieval, it provides stronger evidence grounding and more reliable causal discernment, while its advantage over a fixed hybrid strategy remains selective. Sensitivity analysis identifies node coverage as a more reliable gating signal than raw path count, suggesting that connected supporting evidence is more informative than the number of candidate paths alone. Failure analysis further shows that retrieval-chain misses and imperfect causal graph extraction remain important bottlenecks. Overall, the findings support adaptive semantic-causal integration as a practical approach for choosing between broader semantic exploration and stricter causal verification according to query complexity and evidence density.

Keywords

Causal RAG, HippoRAG, Adaptive Retrieval, Causal Inference, Knowledge Graphs

CEUR-WS.org/Vol-4316/kge-full-1.pdf

1. Introduction

Retrieval-Augmented Generation (RAG) is foundational for knowledge-intensive AI. However, standard architectures often fall into a "correlation trap," prioritizing semantic similarity over the logical or causal grounding required for accurate reasoning. In the Semantic Web, where Knowledge Graphs (KGs) serve as the backbone for structured knowledge [1, 2], distinguishing genuine causal mechanisms from surface-level associations is paramount for reliable AI. Existing RAG systems typically treat semantic connectivity and causal grounding as separate, non-adaptive modalities, leaving a gap in their ability to handle complex queries that require both broad exploration and logical rigor.

Recent research has pursued semantic graph retrieval (e.g., HippoRAG [2]) for multi-hop connectivity and causal path discovery (e.g., CausalRAG [3]) for logical rigor. However, these modalities are often treated as separate, static signals. HippoRAG's PPR-style retrieval can surface semantically adjacent but causally spurious context, whereas CausalRAG's static causal weighting can underperform on purely multi-hop tasks when causal evidence is sparse. We posit that a fixed integration is insufficient; over-prioritizing causal paths can lose essential semantic "bridges" in multi-hop reasoning, while static semantic weights may fail to filter spurious correlations. The optimal balance depends on query complexity and evidence density, requiring a mechanism that can dynamically pivot between broad exploration and rigorous exploitation.

KGE-LLM 2026: International Workshop on Knowledge Graph Engineering in the Era of Large Language Models in conjunction with IJCKG 2026, 19–21 November 2026, Bangkok, Thailand
*Corresponding author.
takanori.ukai@aist.go.jp (T. Ugai); morita@it.aoyama.ac.jp (T. Morita)
0000-0001-5245-9719 (T. Ugai); 0000-0001-8963-2562 (T. Morita)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-06

In this study, we introduce `CausalHippo`, a RAG framework that adaptively integrates graph-based semantic retrieval with causal path discovery. By leveraging HippoRAG's connectivity [2] and the rigor of causal path extraction [3], our system dynamically routes retrieval efforts based on query characteristics and the confidence of the evidence. This maintains high coverage for multi-hop reasoning while defending against spurious correlations, broadening toward semantic exploration when causal signal is sparse and tightening toward causal verification when it is dense.

Our primary contributions are:

1. The `CausalHippo` framework, which unifies dual-indexing with a three-layer adaptive control system
2. A comprehensive analysis of seven RAG configurations across three benchmarks, evaluating both multi-hop performance (MuSiQue [4]) and causal reliability (Webis-CausalQA-22 and our novel Causal-Reasoning-QA dataset)
3. A detailed analysis clarifying the impact of adaptive routing on stability and category-level robustness.

We specifically address the three research questions.

- **RQ1**: Does adaptive semantic-causal routing improve retrieval depth and grounding relative to `CausalRAG` and `CausalHippo-Fixed` on the shared 300-sample subset?
- **RQ2**: When does hybrid integration improve causal abstention over `HippoRAG` and retrieval depth over `CausalRAG`?
- **RQ3**: How stable are these gains across heterogeneous benchmarks and label regimes?

The remainder of this paper is organized as follows. Section 2 reviews the related work. Section 3 details the proposed `CausalHippo` framework. Section 4 describes the experimental design. Section 5 presents the evaluation results. Section 6 provides an ablation and failure analysis, and Section 7 concludes the study.

2. Related Work

Retrieval-Augmented Generation (RAG) has developed from predominantly dense retrieval toward graph-based and adaptive architectures. This shift addresses a recurring limitation of similarity-driven retrieval: passages can be topically relevant yet fail to provide the relational or causal structure required for multi-step reasoning.

Neuro-Symbolic and Graph-based RAG. Graph-based RAG methods introduce an intermediate structure between documents and generation by representing entities, relations, or communities explicitly. HippoRAG [2] uses a hippocampal indexing perspective to support long-range multi-hop retrieval, while GraphRAG [1] organizes corpus-level entity and community structure. *LightRAG* [5] and *PathRAG* [6] further emphasize lightweight graph organization and path-oriented retrieval, respectively. These methods improve connectivity and contextual organization and are compatible with the Semantic Web's emphasis on structured knowledge. However, graph connectivity alone does not establish that a retrieved relation is causal: associative or conventional knowledge-graph edges may connect relevant entities without specifying a mechanism or direction of influence.

Causal Reasoning in Retrieval. Causal RAG methods address this limitation by making cause–effect structure an explicit retrieval signal. CausalRAG [3] identifies and traces causal paths in external knowledge so that retrieved context can support a causal explanation. CF-RAG [7] instead uses counterfactual reasoning to test whether retrieved evidence is robust to spurious correlations. Other recent approaches, including CDF-RAG [8] and Causal Graphs Meet Thoughts [9], combine causal retrieval with feedback or stepwise refinement, while causal-graph RAG has also been used to study causal reasoning in LLMs [10]. Collectively, these studies establish the value of causal structure for reducing unsupported explanations. Their emphasis is primarily on causal evidence and verification, leaving open the question of how a system should retain broad semantic connectivity when causal evidence is sparse or incomplete.

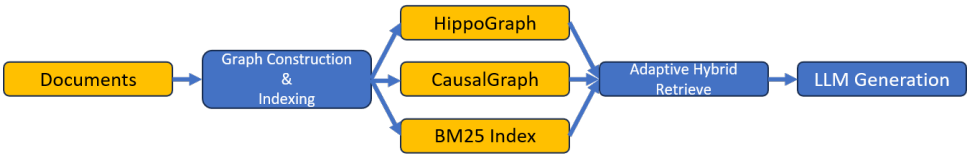


Figure 1: System Architecture and Inference Flow. Orange components represent data and indices, while blue components denote processing steps. The Adaptive Hybrid Retriever operates over the pre-constructed Semantic, Causal, and Lexical indices.

Adaptive and Hybrid Retrieval Strategies. A complementary line of work adapts retrieval to the query or to the reliability of intermediate evidence. Dynamic Alpha Tuning (DAT) [11] and Adaptive-RAG [12] adjust the balance between lexical (BM25) and dense semantic retrieval, for example as a function of query complexity. Confidence-based switching [13] and two-pass adaptation [14] provide mechanisms for suppressing unreliable signals or revising retrieval decisions during inference. These approaches motivate adaptive routing, but they do not by themselves resolve the distinct roles of semantic connectivity and causal verification. `CausalHippo` extends this perspective by treating semantic, causal, and lexical retrieval as complementary signals whose relative influence can be adjusted from both query characteristics and evidence agreement.

Evaluation of Complex Reasoning. Evaluation must distinguish retrieval quality from the correctness and causal reliability of the generated answer. MuSiQue [4] tests multi-hop reasoning over disconnected supporting facts, making it useful for measuring whether a retriever preserves bridge evidence. Webis-CausalQA-22 [15] provides a large corpus of more than 1.1 million causal questions spanning consequence, reason, and purpose. GraphRAG-Bench [16] complements these datasets with college-level, domain-specific questions across 16 disciplines and multiple answer formats, including multiple-choice, true/false, multi-select, open-ended, and fill-in-the-blank tasks. It evaluates graph construction, retrieval, answer generation, and reasoning coherence across the full GraphRAG pipeline. Together, these benchmarks expose different failure modes: a system may retrieve a coherent semantic neighborhood while missing a necessary reasoning chain, or it may generate a plausible answer without distinguishing causation from correlation.

Research Gap and Positioning. The literature therefore provides the three ingredients needed for robust causal RAG, but largely treats them as separate design choices: graph structure supplies connectivity, causal paths supply mechanistic constraints, and adaptive retrieval supplies query-dependent control. Our work studies their joint integration in a single retrieval pipeline. Specifically, `CausalHippo` maintains semantic, causal, and lexical indices and uses a staged controller to adjust their relative influence before and after candidate retrieval. We evaluate the resulting system not only with ranking and answer metrics, but also with causal-label metrics that measure both genuine causal recognition and non-causal abstention. This positions `CausalHippo` as an adaptive semantic-causal integration framework for the Semantic Web, rather than as a claim that causal retrieval universally replaces semantic or fixed hybrid retrieval.

Motivation and Failure Modes. The two closest baselines expose a concrete tension. HippoRAG's connectivity can over-retrieve semantically related but causally irrelevant passages, whereas CausalRAG's stronger causal focus can lose bridge documents when explicit causal evidence is sparse. These are not mutually exclusive failure modes: a reliable system must know when to broaden retrieval and when to enforce structural verification. `CausalHippo` is designed around this decision, broadening retrieval under low causal confidence and tightening it when the extracted causal subgraph is dense. The experiments consequently test a selective improvement hypothesis, not universal dominance over every static alternative.

3. The `CausalHippo` Framework

In this section, we introduce the proposed framework, `CausalHippo`, whose architecture is illustrated in Fig. 1. While the figure focuses on the inference-time retrieval and generation flow, the pipeline

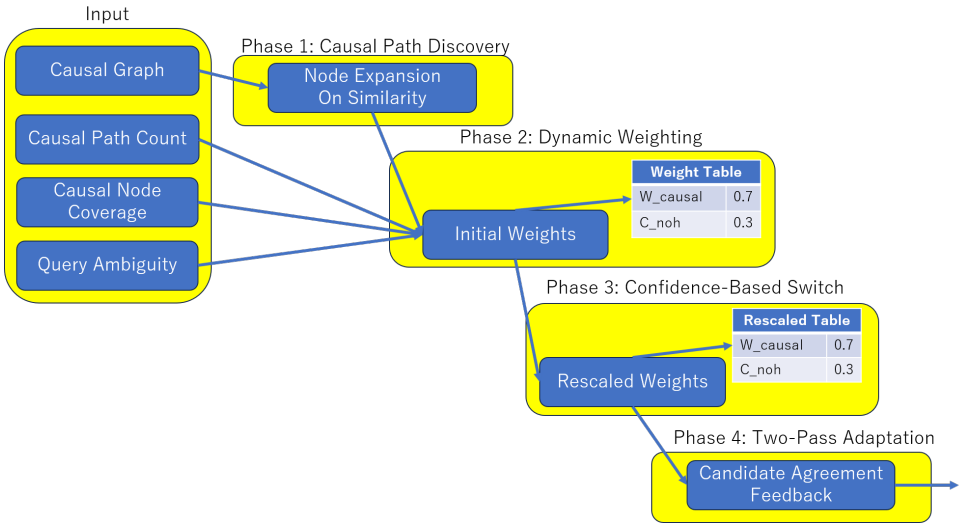


Figure 2: Architecture of Adaptive Hybrid Retriever

fundamentally depends on a pre-processing stage and a dynamic **Adaptive Hybrid Retriever** mechanism. During inference, the system first identifies explicit logical structures through path discovery, which then informs the weighting strategy over the pre-constructed indices to extract the most relevant evidence for the LLM.

3.1. Graph Construction and Indexing

At the outset, documents are indexed across three distinct components to enable multi-modal retrieval. This **dual-indexing** architecture maintains both a semantic HippoGraph and a causal CausalGraph, alongside a standard lexical index, to capture diverse knowledge representations. For semantic graph-based retrieval, we constructed the **HippoGraph** using HippoRAG [2], which structures documents into a knowledge graph to facilitate broad contextual understanding. Concurrently, the **CausalGraph** is constructed via an LLM-driven Open Information Extraction (OpenIE) process to extract relational triples and preserve underlying cause-and-effect structures with CausalRAG. The resulting nodes are stored in a vector database using text embeddings for similarity-based retrieval. Finally, a standard **BM25 Lexical Index** is maintained to capture exact token matches. This strategy ensures that broad semantic themes, specific causal mechanisms, and lexical details are all available for query resolution.

3.2. Adaptive Hybrid Retriever

The core innovation of `CausalHippo` lies in the dynamic fusion of semantic, causal, and lexical signals (Fig. 2). Rather than using a static weighting scheme, the system continuously adjusts the influence of each modality based on query characteristics and retrieval confidence. This recognizes that a universal retrieval posture is suboptimal; multi-hop queries may require expansive semantic exploration, while causal inquiries demand rigorous structural validation. As shown in the architecture diagram, adaptive control is applied in four sequential phases:

- **Phase 1 (Causal Path Discovery):** Performs dynamic path discovery to refine raw signals from the CausalGraph.
- **Phase 2 (Dynamic Weighting):** Sets local relative weights based on path counts, node coverage, and query ambiguity.
- **Phase 3 (Confidence-Based Switch):** Acts as a safety gate measuring absolute evidence density to attenuate unreliable signals.
- **Phase 4 (Two-Pass Adaptation):** Provides a post-retrieval check, refining weights by evaluating candidate alignment with the causal graph.

This three-layer controller (Phases 2–4) is conceptually necessary to safely interpolate semantic exploration with causal exploitation. Relative scaling in Phase 2 cannot judge if the absolute density of causal evidence is sufficient. Phase 3 serves as an essential safety gate that overrides relative weights when absolute evidence is sparse, falling back on semantic retrieval to avoid overreliance on noisy paths. Finally, because Phases 2–3 are pre-retrieval, Phase 4 provides post-retrieval verification, measuring both entity overlap (node match) and the structural preservation of discovered cause-effect links (path match) within candidate text to amplify verified paths and diminish spurious semantic links.

Causal Path Discovery. As the first stage of adaptive retrieval, the system performs dynamic path discovery to refine raw signals from the CausalGraph. A **causal path** is formally defined as a sequence of directed cause-effect triples (e.g., "X increases $\rightarrow$ Y decreases") connecting a seed entity to a query-relevant entity. It differs fundamentally from a *semantic path* (which connects entities purely by embedding similarity or PPR connectivity in the HippoGraph without causal direction) and a *traditional relational path* (which connects entities via static non-causal relations like `bornIn` or `capitalOf` in standard KGs [6]). Causal relationships are extracted at indexing via OpenIE (cause, relation, effect) and stored as directed edges in the CausalGraph.

Upon receiving a query, seed nodes are retrieved from the CausalGraph by computing cosine similarity between the query embedding and node embeddings in the CausalGraph. These nodes are expanded along directed edges—to both ancestors and descendants—up to a configurable hop limit H, capturing long-range causal relationships. Nodes with semantic similarity above threshold θ_{sim} to seed centroids are included to handle linguistic variation. The system then identifies explicit causal paths, filtering out isolated elements. As a running example, consider the question in `Data-Simple_Direct_Causation_265`: "What is the cause of a reduction in noise complaints?" Phase 1 extracts seed node "reduction in noise complaints" and discovers the directed causal path "implementing quiet hours in the apartment complex $\rightarrow$ reduction in noise complaints."

Dynamic Weighting. Before scoring candidate passages, a logistic model determines a query-specific causal gate based on three factors: the normalized count of discovered causal paths (N_{paths}), the coverage of retrieved causal nodes within candidate passages, and an estimation of query ambiguity derived from the score margin between the top seed nodes. The logit is computed using hyperparameters $(\alpha, \beta, \gamma, \delta, \tau)$ to balance evidence strength and uncertainty:

$$logit = \alpha \cdot \min\left(\frac{N_{paths}}{\tau}, 1.0\right) + \beta \cdot \text{coverage} - \gamma \cdot \text{ambiguity} - \delta \tag{1}$$

where *coverage* is the ratio of unique causal nodes found in any candidate passage to the total retrieved nodes, and *ambiguity* is $1 - (S_1 - S_2)/S_1$ for top-2 seed scores S_1, S_2 (representing the cosine similarity between the query and the closest nodes in the CausalGraph). A narrow margin between S_1 and S_2 indicates that the system is uncertain about the primary causal concept (high ambiguity), whereas a large margin suggests confident entity resolution (low ambiguity).

The system uses a fixed lexical weight, W_{bm25}, to ensure that keyword-level evidence is always considered. The remaining non-lexical retrieval budget $(1 - W_{bm25})$ is dynamically apportioned between the causal weight (W_{causal}) and the semantic weight ($W_{semantic}$) using the sigmoid function $\sigma(x) = (1 + e^{-x})^{-1}$:

$$W_{causal} = (1 - W_{bm25}) \cdot \sigma(logit), \quad W_{semantic} = (1 - W_{bm25}) \cdot (1 - \sigma(logit)) \tag{2}$$

Confidence-Based Switch. Following the initial weighting, a secondary safety mechanism attenuates the causal weight if the absolute evidence of causality is weak. The system computes a confidence score $C_{causal} \in [0, 1]$ by normalizing the raw counts of retrieved causal nodes (N_{nodes}) and paths (N_{paths}) against predefined saturation thresholds τ_{nodes} and τ_{paths}:

$$C_{causal} = \frac{1}{2}\left(\min\left(\frac{N_{nodes}}{\tau_{nodes}}, 1.0\right) + \min\left(\frac{N_{paths}}{\tau_{paths}}, 1.0\right)\right) \tag{3}$$

These thresholds define the minimum amount of evidence required to consider the causal signal fully reliable. If confidence is low, indicating a sparse causal signal, the causal weight is rescaled accordingly, and the remainder of the non-lexical budget is re-apportioned to the semantic weight to ensure the total weight remains normalized ($W_{bm25} + W'_{causal} + W'_{semantic} = 1.0$):

$$W'_{causal} = W_{causal} \cdot C_{causal}, \quad W'_{semantic} = W_{semantic} + (W_{causal} - W'_{causal}) \tag{4}$$

This ensures that when causal evidence is insufficient, the system gracefully falls back to a semantic-heavy retrieval posture without losing total retrieval influence.

Two-Pass Adaptation. The final stage refines the retrieval weights using a first-pass candidate agreement mechanism. Candidates are initially scored using a semantic prior P_{sem}. For each top-ranking passage, a causal alignment score is computed based on node overlap and path structure preservation. We assign importance weights w_{node} and w_{path} to these signals to balance overall relevance with structural rigor:

$$\text{score}_{causal} = w_{node} \cdot \text{node_match} + w_{path} \cdot \text{path_match} \tag{5}$$

where *node_match* is the proportion of causal subgraph nodes present in the passage, and *path_match* is the proportion of discovered cause-effect pairs that are preserved. The system measures the average causal alignment across the top N candidates to form a target causal weight:

$$W_{target} = (1 - W_{bm25}) \cdot \frac{1}{N} \sum_{i=1}^{N} \text{score}^{(i)}_{causal} \tag{6}$$

The final causal influence is interpolated between the current dynamically adjusted weight and this target using a blend factor λ:

$$W''_{causal} = W'_{causal} + \lambda \cdot (W_{target} - W'_{causal}) \tag{7}$$

Consequently, causal influence is amplified when first-pass semantic candidates consistently align with the causal graph, and diminished otherwise. In our running example (`Data-Simple_Direct_Causation_265`), this post-retrieval verification boosts the verified cause-effect link, enabling `CausalHippo` and `CausalRAG` to recover the exact causal answer, whereas semantic-only `HippoRAG` suffers from capitalization and descriptive variation.

4. Experimental Design

4.1. Datasets and Benchmarks

To evaluate the effectiveness of `CausalHippo` in both complex reasoning and causal inference tasks, we utilize three primary datasets:

Although GraphRAG-Bench [16] provides a valuable evaluation of general graph-based reasoning, we do not include it in our experiments because it is not specifically designed to distinguish causal mechanisms from correlations or to evaluate causal abstention. We instead use MuSiQue to assess multi-hop retrieval, and Webis-CausalQA-22 and Causal-Reasoning-QA to assess causal reliability and discrimination.

Webis-CausalQA-22 [15]. This benchmark corpus comprises 1.1 million causal questions and answers derived from ten large-scale question-answering datasets. It features a diverse typology of causal questions, such as *consequence*, *reason*, and *purpose* questions. For our experiments, we used a balanced subset of 300 samples across nine source datasets (e.g., GooAQ [17], MS MARCO [18], and SQuAD [19]), each augmented with three hard-negative distractors to evaluate the system's robustness against lexical overlap and its ability to generalize across various causal query types.

MuSiQue [4]. This dataset contains complex 2–4 hop questions requiring multi-step reasoning. For our evaluation, we construct a balanced experiment subset of 300 answerable samples. This subset is

stratified to maintain equal distribution across hop counts (2, 3, and 4+ hops) and to ensure the system is tested across varying levels of reasoning depth. Furthermore, to increase the difficulty of the retrieval task and minimize surface-level lexical shortcuts, each sample is augmented with three "hard-negative" distractor paragraphs. These distractors are selected from the broader dataset for their high lexical overlap with the query but lack the supporting evidence required to answer the question.

Causal-Reasoning-QA. This dataset is a novel benchmark developed for this study to test the model's ability to distinguish genuine causal mechanisms from deceptive correlations. Recognizing that large-scale causal reasoning datasets with long-context ground truths are currently a major gap in the field, we constructed this benchmark using `GPT-4.5-preview` to generate 10,000 long-context examples across 20 distinct categories, explicitly pairing genuine causal queries with correlation-only distractors to target causal discernment. To mitigate potential generation bias and ensure evaluation integrity, we implemented a multi-agent peer-review pipeline involving a *Causal Relationship Judge* to provide initial reasoning, a *Validity Critic* explicitly prompted to identify logical fallacies or hallucinations, and an *Aggregator* to reconcile disagreements. We evaluated this synthetic dataset alongside the non-synthetic, human-annotated Webis-CausalQA-22 benchmark to ensure that our framework's performance generalizes beyond synthetic contexts. For our RAG evaluation, we derived a balanced experimental subset of 300 samples, each embedded within five noise paragraphs to simulate a realistic retrieval environment.

4.2. Baseline Systems and Variations

To rigorously evaluate the proposed framework, we compared seven distinct RAG configurations:

1. `CausalHippo` **(Primary):** This is our core proposed system. It integrates the graph-based semantic retrieval of HippoRAG with a causal path discovery mechanism. All three adaptive control layers—Dynamic Weighting (Phase 2), Confidence-Based Switch (Phase 3), and Two-Pass Adaptation (Phase 4)—are enabled to dynamically route retrieval effort between semantic and causal signals.
2. `CausalHippo-Fixed` **(Variation):** A non-adaptive version of our primary system. It shares the same hybrid architecture but employs a static weighting scheme ($W_{semantic} = 0.4, W_{causal} = 0.5, W_{bm25} = 0.1$). All three layers of the adaptive control system are disabled in this configuration.
3. `CausalHippo (no Ph3)` **(Variation):** An ablation version of `CausalHippo` where the **Confidence-Based Switch (Phase 3)** is disabled. It retains Dynamic Weighting (Phase 2) and Two-Pass Adaptation (Phase 4).
4. `CausalHippo (no Ph4)` **(Variation):** An ablation version where **Two-Pass Adaptation (Phase 4)** is disabled. It retains Dynamic Weighting (Phase 2) and the Confidence-Based Switch (Phase 3).
5. `CausalHippo (no Ph3&4)` **(Variation):** An ablation version where both Phase 3 and Phase 4 are disabled, leaving only **Dynamic Weighting (Phase 2)** to manage the retrieval posture.
6. `CausalRAG` [3]: The original, baseline implementation of CausalRAG. It uses a static weighting scheme to combine causal-path evidence with standard vector-based semantic retrieval.
7. `HippoRAG` [2]: A baseline implementation of HippoRAG operating in its graph-based retrieval mode. It utilizes a knowledge graph indexed from the document corpus and employs Personalized PageRank (PPR) to identify relevant context.

4.3. Experimental Configuration

We conducted the experiments using a unified retrieval-and-generation pipeline implemented in Kotlin. All the RAG systems used in this study were faithful re-implementations in Kotlin. While the original baselines were developed in Python, we ported them to the same JVM environment as `CausalHippo` to ensure a fair comparison and to leverage the architectural benefits of the Kotlin/JVM ecosystem.

This choice was motivated by three primary factors: (1) **Type Safety**: Kotlin's static typing and null safety provide superior rigor for managing complex graph structures and multi-stage RAG pipelines; (2) **Interoperability**: JVM-based implementation allows for seamless, high-performance integration with established Semantic Web tools; and (3) **Production Efficiency**: The JVM's robust concurrency model and JIT optimization ensure the system is suitable for high-throughput, real-world reasoning tasks. To ensure functional equivalence, we performed a line-by-line logic migration and validated each component through parity testing, confirming that the Kotlin implementations produce identical retrieval and scoring outputs to their original Python counterparts when provided with standardized reference inputs.

We utilized `gpt-5.4-mini` as the primary large language model for both answer synthesis and causal path extraction. For vector-based semantic retrieval and node embedding, we employed the `text-embedding-3-small` model. Both models were accessed via the OpenAI API with default temperature settings ($T = 0.0$) to ensure the reproducibility of the generated reasoning paths. For systems other than `CausalHippo`, we used the parameters and prompts from the original implementation for all components except the LLM model. We used the same LLM model as `CausalHippo` across all evaluations.

To ensure technical comparability, all systems are evaluated on the same benchmark subsets, with the same top-K retrieval cutoff, model family, and aggregation procedure. The intended controlled variables are the retrieval architecture and routing strategy, rather than the generator identity.

Across all evaluated systems, the retrieval component was configured with a cutoff of $K = 5$ to identify the most relevant supporting context. We selected this value as a conservative compromise: it is large enough to preserve bridge evidence for multi-hop reasoning, while remaining small enough to limit context noise and keep the generator focused. For the `CausalHippo` framework, the four phases of the adaptive control system were initialized with hyperparameters derived from preliminary pilot experiments (conducted with 20 samples for each dataset) to optimize the balance between coverage and precision:

- **Phase 1 (Causal Path Discovery):** The system utilizes a hop limit $H = 2$ for graph expansion and a semantic similarity threshold $\theta_{sim} = 0.8$ for node inclusion. We selected $H = 2$ because a single hop was often insufficient to surface bridge evidence in multi-step reasoning, whereas larger hop limits introduced substantial noise and redundant expansions in pilot runs. Likewise, $\theta_{sim} = 0.8$ offered the best precision-recall trade-off for node inclusion, preserving near-paraphrastic matches while filtering out loosely related nodes that degraded causal grounding.
- **Phase 2 (Dynamic Weighting):** The causal gate logit utilizes coefficients $\alpha = 2.0$ and $\beta = 2.0$ for path count and node coverage, $\gamma = 1.5$ for ambiguity, a saturation threshold $\tau = 3.0$ for path evidence, and a bias $\delta = 0.5$.
- **Phase 3 (Confidence-Based Switch):** Causal confidence is normalized against saturation thresholds $\tau_{nodes} = 6.0$ and $\tau_{paths} = 3.0$.
- **Phase 4 (Two-Pass Adaptation):** The system employs a semantic prior $P_{sem} = 0.85$ during the first-pass agreement phase to ensure high-quality thematic filtering. Alignment scores are computed with $w_{node} = 0.7$ and $w_{path} = 0.3$, with a blend factor $\lambda = 0.7$ for the final weight interpolation.

For the non-adaptive `CausalHippo-Fixed` variation, the static weights were set to $W_{semantic} = 0.4$, $W_{causal} = 0.5$, and $W_{bm25} = 0.1$. We selected these weights based on preliminary pilot experiments to establish a "causal-heavy" but balanced posture, where structural signals (W_{causal}) are prioritized for grounding, while semantic signals ($W_{semantic}$) maintain broad connectivity, anchored by a minimal lexical weight (W_{bm25}) for exact token matching.

All indexing operations, including the construction of the knowledge and causal graphs, were performed using the same embedding model to maintain representational consistency across the semantic and causal modalities.

4.4. Evaluation Metrics

To evaluate retrieval and generation, we employ a multifaceted suite of metrics. For ranking quality, we report **Recall@5**, **Mean Reciprocal Rank (MRR@5)**, and **nDCG@5**. For generation quality, we use word-level **F1-score**, **Exact Match (EM)**, and **BERTScore (Precision, Recall, F1)** [20] using the `paraphrasemultilingualMiniLML12v2` model for semantic similarity. Trustworthiness is assessed via **Faithfulness** and **Groundedness** from the Ragas framework [21]. For causal discrimination, we report class-specific **F1 (True)**, **F1 (False)**, and **Unknown Recall** (ratio of correctly identified non-causal cases). Finally, we calculate the arithmetic mean across samples and measure **Stability** as the standard deviation of F1-scores across benchmarks. Samples that resulted in execution errors were excluded from the mean calculation to ensure that the metrics reflected successful logical execution.

For transparency, we release per-question metric records and logs in the supplemental artifact, enabling recomputation with alternative aggregation choices (including error-as-zero analyses).

5. Evaluation Results

In this section, we present a comparative evaluation of the seven RAG configurations across the Webis-CausalQA-22, MuSiQue, and Causal-Reasoning-QA datasets. We analyzed the results using standard accuracy metrics, retrieval quality, and robustness to causal reasoning.

5.1. Results on Webis-CausalQA-22

The evaluation results for the Webis-CausalQA-22 [15] balanced dataset are presented in Table 1.

Table 1
Webis-CausalQA-22 [15] Evaluation Results

System	EM	F1	Recall@5	MRR@5	nDCG@5	BS-P	BS-R	BS-F1	Faith.	Ground.
`CausalHippo`	0.333	0.517	0.853	0.687	0.729	0.715	0.671	0.683	**0.940**	**0.910**
`CausalHippo-Fixed`	0.333	0.512	0.847	0.690	0.729	0.714	0.669	0.681	0.922	0.896
`CausalHippo (no Ph3)`	0.355	0.531	**0.860**	**0.693**	**0.735**	0.727	0.683	0.695	0.931	0.904
`CausalHippo (no Ph4)`	0.357	0.521	0.850	0.683	0.725	0.721	0.674	0.687	0.920	0.897
`CausalHippo (no Ph3&4)`	0.317	0.498	0.847	0.693	0.732	0.714	0.670	0.681	0.923	0.890
`CausalRAG`	**0.360**	0.533	0.817	0.665	0.703	**0.737**	0.693	0.704	0.936	0.902
`HippoRAG`	0.283	**0.549**	0.800	0.589	0.642	0.707	**0.740**	**0.714**	0.892	0.775

On Webis-CausalQA-22, `HippoRAG` achieved the highest lexical F1 (0.549), while `CausalRAG` achieved the best EM (0.360). Among the `CausalHippo` variants, the full adaptive system achieved the best Faithfulness (0.940) and Groundedness (0.910), while disabling the confidence-based switch (`no Ph3`) yielded the best overall retrieval Recall@5 (0.860) and MRR@5 (0.693). This pattern suggests a practical trade-off: while adaptive routing improves evidence quality and grounding behavior, individual components, such as dynamic weighting, can provide substantial coverage gains in high-noise causal contexts.

5.2. Results on MuSiQue

The evaluation results for the MuSiQue multi-hop reasoning task are presented in Table 2.

Table 2
MuSiQue [4] Evaluation Results

System	EM	F1	Recall@5	MRR@5	nDCG@5	BS-P	BS-R	BS-F1	Faith.	Ground.
`CausalHippo`	0.243	0.361	0.586	**0.822**	**0.600**	0.647	0.647	0.644	0.900	0.866
`CausalHippo-Fixed`	0.267	0.372	0.562	0.782	0.567	**0.648**	0.648	0.646	**0.903**	**0.879**
`CausalHippo (no Ph3)`	0.233	0.351	0.583	0.800	0.590	0.640	0.635	0.634	0.893	0.846
`CausalHippo (no Ph4)`	0.233	0.349	0.586	0.808	0.593	0.639	0.639	0.636	0.893	0.867
`CausalHippo (no Ph3&4)`	0.237	0.359	0.577	0.805	0.587	0.644	0.639	0.639	0.890	0.868
`CausalRAG`	0.223	0.350	0.545	0.747	0.540	0.616	0.618	0.615	0.885	0.868
`HippoRAG`	**0.293**	**0.403**	**0.622**	0.626	0.538	0.644	**0.663**	**0.649**	0.762	0.697

On MuSiQue, the graph-based baseline HippoRAG achieved the highest F1 (0.403) and Recall@5 (0.622), suggesting that semantic graph connectivity is effective for purely multi-hop tasks. However, the CausalHippo variants outperform both HippoRAG and CausalRAG in retrieval precision, with the full system achieving the best MRR@5 (0.822) and nDCG@5 (0.600). This demonstrates that integrating HippoRAG's connectivity with adaptive causal filtering provides a more robust foundation for multi-step reasoning than static retrieval schemes. Furthermore, the CausalHippo family consistently maintains superior trustworthiness, with CausalHippo-Fixed achieving the best Faithfulness (0.903) and Groundedness (0.879).

5.3. Results on Causal-Reasoning-QA

The results for the Causal-Reasoning-QA task, which focuses on distinguishing causation from correlation, are presented in Table 3.

Table 3
Causal-Reasoning-QA Evaluation Results

System	EM	F1	Recall@5	MRR@5	nDCG@5	BS-P	BS-R	BS-F1	Faith.	Ground.
CausalHippo	**0.367**	**0.459**	0.500	0.497	0.498	**0.616**	**0.632**	**0.621**	0.407	0.543
CausalHippo-Fixed	0.357	0.450	0.500	0.497	0.498	0.605	0.624	0.612	0.395	0.534
CausalHippo (no Ph3)	0.320	0.406	0.500	0.498	0.499	0.568	0.588	0.575	0.404	0.542
CausalHippo (no Ph4)	0.318	0.406	**0.501**	**0.499**	**0.500**	0.571	0.591	0.578	0.395	0.534
CausalHippo (no Ph3&4)	0.326	0.417	**0.501**	0.498	0.499	0.575	0.594	0.582	0.400	0.543
CausalRAG	0.343	0.437	0.493	0.490	0.491	0.595	0.613	0.602	0.386	0.529
HippoRAG	0.157	0.291	0.500	0.439	0.455	0.461	0.565	0.500	**0.847**	**0.810**

On Causal-Reasoning-QA, the primary CausalHippo system obtained the highest aggregate F1 (0.459) and EM (0.367). In terms of trustworthiness, CausalHippo achieved a Faithfulness of 0.407 and shared the highest Groundedness (0.543) with the no Ph3&4 variant, outperforming the baseline CausalRAG (0.386 and 0.529).

Importantly, HippoRAG has no general faithfulness advantage across benchmarks: mean Faithfulness on Webis-CausalQA-22 is CausalHippo 0.940, CausalHippo-Fixed 0.922, CausalRAG 0.936, and HippoRAG 0.892; on MuSiQue, it is CausalHippo 0.900 vs. HippoRAG 0.762. Its higher apparent Faithfulness on Causal-Reasoning-QA (0.847) is a sample-level metric artifact of Ragas Faithfulness. In sample naturalquestions_839, for example, CausalHippo answers the exact ground-truth causal label ("human activities", EM/F1 = 1.0) but receives a Faithfulness score of 0.0 because the concise target string does not echo redundant background context. In contrast, HippoRAG outputs a lengthy, ungrounded explanation (EM/F1 = 0.0) but receives Faithfulness 1.0 because Ragas rewards broad, topically compatible narrative overlap over exact causal targets. As shown below, HippoRAG's significantly lower F1 (0.291) and near-zero Unknown Recall (0.033) demonstrate that semantic connectivity alone is insufficient for robust causation-correlation discrimination.

5.4. Fine-grained Causal Analysis

Table 4 provides a breakdown of the performance across causal (True) and non-causal (False) classes, including the semantic similarity profile using BERTScore.

Table 4
Causal Label-Aware Metrics

System	Unk. P	Unk. R	Unk. F1	F1(T)	F1(F)	BS-P(T)	BS-R(T)	BS-F1(T)	BS-P(F)	BS-R(F)	BS-F1(F)
CausalHippo	**0.626**	0.447	0.521	0.472	**0.447**	**0.637**	0.648	0.639	**0.595**	**0.616**	**0.603**
CausalHippo-Fixed	0.600	0.460	0.521	0.436	**0.464**	0.602	0.622	0.608	0.608	**0.627**	**0.616**
CausalHippo (no Ph3)	0.542	0.400	0.460	0.411	0.400	0.586	0.600	0.590	0.551	0.576	0.561
CausalHippo (no Ph4)	0.557	0.414	0.475	0.404	0.408	0.586	0.601	0.590	0.557	0.581	0.567
CausalHippo (no Ph3&4)	0.555	0.413	0.474	0.421	0.413	0.586	0.603	0.592	0.563	0.585	0.572
CausalRAG	0.595	0.440	0.506	0.440	0.433	0.604	0.621	0.609	0.587	0.606	0.594
HippoRAG	0.625	0.033	0.063	**0.509**	0.073	0.633	**0.736**	**0.672**	0.288	0.394	0.328

Label-aware analysis of the Causal-Reasoning-QA experiment revealed that while HippoRAG shows high F1 (True), its F1 (False) and **Unknown Recall** ($Recall_{unknown} \approx 0.03$) are near zero, indicating a severe bias toward predicting causality regardless of evidence. In contrast, the CausalHippo family maintained more balanced profiles. The primary system achieved the best **Unknown Precision** (0.626) and F1 (True) (0.472) among high-accuracy models, while CausalHippo-Fixed achieved the best F1 (False) (0.464). This confirms that integrating causal paths and adaptive control enables the system to reliably avoid false causal claims while preserving accuracy for genuine causal relationships.

These results characterize CausalHippo as a practically useful, precision-oriented hybrid rather than a universally dominant system. Its value is especially clear when a system must preserve broad semantic coverage while making causal claims selectively: the adaptive controller can use causal paths when evidence is dense and retain semantic retrieval when causal evidence is weak. This operating point is useful for applications in which unsupported causal assertions are more costly than an abstention, even though a fixed hybrid may remain preferable for some aggregate metrics or task distributions.

6. Ablation Analysis and Discussion

In this section, we provide a deeper analysis of the CausalHippo framework, examining its unique performance profile and positioning within the broader RAG design space.

Table 5
Aggregate Performance Across All Benchmarks

System	EM	F1	Recall@5	MRR@5	nDCG@5	BS-P	BS-R	BS-F1	Faith.	Ground.
CausalHippo	0.314	**0.446**	**0.646**	**0.669**	**0.609**	0.659	0.650	**0.649**	0.749	**0.773**
CausalHippo-Fixed	**0.319**	0.445	0.636	0.656	0.598	0.656	0.647	0.646	0.740	0.770
CausalHippo (no Ph3)	0.303	0.429	**0.648**	0.664	0.608	0.645	0.635	0.635	0.743	0.764
CausalHippo (no Ph4)	0.303	0.425	0.646	0.663	0.606	0.644	0.635	0.634	0.736	0.766
CausalHippo (no Ph3&4)	0.293	0.425	0.642	0.665	0.606	0.644	0.634	0.634	0.738	0.767
CausalRAG	0.309	0.440	0.618	0.634	0.578	0.649	0.641	0.640	0.736	0.766
HippoRAG	0.244	0.414	0.641	0.551	0.545	0.604	**0.656**	0.621	**0.834**	0.761

6.1. Architectural Trade-offs

Our results highlight CausalHippo's ability to balance multi-hop connectivity with causal discrimination. As shown in Table 5, the full CausalHippo achieves the highest aggregate F1 score (0.446) and MRR@5 (0.669). Interestingly, disabling the confidence-based switch (no Ph3) yielded the highest aggregate Recall@5 (0.648). This suggests a performance trade-off: while Phase 3 improves precision and grounding, Phase 2 in isolation provides more aggressive coverage. This indicates that integrating causal reasoning significantly improves both structural reliability and category-level robustness.

Synergistic Integration and Statistical Significance Paired t-tests on the 300-sample subset (with 10,000 bootstrap replicates and Holm correction) show that CausalHippo does not significantly outperform CausalHippo-Fixed across all metrics; hence, we do not claim that adaptive routing is a universal win over a well-tuned static hybrid. Against CausalRAG, the only statistically robust gain is Recall@5 (+0.040, 95% CI [0.0167, 0.0667], adjusted $p = 0.0088$), while other metric differences are not statistically significant. Conversely, CausalHippo strongly outperforms HippoRAG on EM, Recall@5, MRR@5, nDCG@5, Faithfulness, and Groundedness (as well as non-causal abstention, Unknown Recall: 0.447 vs. 0.033, Table 4), providing robust support for hybrid semantic-causal integration over graph-only baselines. The practical value of the proposed technique therefore lies not in universally replacing a well-tuned static hybrid, but in offering query- and evidence-dependent routing that preserves semantic coverage while selectively strengthening causal verification and abstention.

Compared with CausalHippo-Fixed, the adaptive CausalHippo does not achieve uniformly better performance across all metrics. It improves aggregate Recall@5 (0.646 vs. 0.636), MRR@5 (0.669 vs. 0.656), and Groundedness (0.773 vs. 0.770), while the fixed variant performs better on aggregate EM (0.319 vs. 0.314) and achieves higher Faithfulness and Groundedness on MuSiQue (0.903 and 0.879 vs.

Table 6
Stability and robustness across evaluated systems. Lower is better for σ_{F1}, and CV_{F1}; higher is better for Min_{F1}. The symbol "–" denotes a metric not reported in the current analysis.

System	$\sigma_{F1} \downarrow$	$Min_{F1} \uparrow$	$CV_{F1} \downarrow$
CausalHippo	0.064	0.361	0.434
CausalHippo-Fixed	0.057	0.372	–
CausalHippo (no Ph3)	0.075	0.351	–
CausalHippo (no Ph4)	0.072	0.349	–
CausalHippo (no Ph3&4)	0.057	0.359	0.401
CausalRAG	0.075	0.350	0.488
HippoRAG	0.106	0.291	0.867

0.900 and 0.866). These results indicate that the benefit of adaptive routing is selective and depends on the benchmark and evaluation metric, rather than representing a universal improvement over static hybrid weighting.

Quantifying Stability and Robustness To analyze the consistency across benchmarks, we defined three key metrics:

- **Stability (σ_{F1}):** Standard deviation of F1 scores across benchmarks (lower indicates higher consistency).
- **Floor (Min_{F1}):** The minimum performance bound observed.
- **Robustness (CV_{F1}):** The coefficient of variation across the 20 logical categories in Causal-Reasoning-QA.

Table 6 summarizes these metrics for all evaluated configurations. The stability and floor values are recomputed from the three benchmark-level F1 scores reported in Tables 1, 2, and 3; the CV is included where available for the corresponding analyses.

CausalHippo demonstrates higher stability ($\sigma = 0.064$ vs. 0.075 for CausalRAG) and a higher performance floor (0.361 vs. 0.350) than the causal-only baseline. Its inter-category robustness was significantly better ($CV_{F1} = 0.434$ vs. 0.488 for CausalRAG, and 0.867 for HippoRAG). Notably, the no Ph3&4 variant achieved the best overall Robustness ($CV_{F1} = 0.401$), suggesting that basic dynamic weighting provides a robust defense against category-level failure modes.

Ablation of Adaptive Gating Analysis of the non-adaptive variant, CausalHippo-Fixed, revealed a higher performance floor (0.372) and the best average rank (1.33). This suggests that while adaptive gating excels in specialized causal tasks, static weighting provides a consistent "causal filter" that prevents semantic noise from dominating the retrieved context. However, the adaptive mechanism in CausalHippo remains superior in terms of aggregate retrieval coverage and category-level robustness.

Summary of Baselines Compared to the pure semantic graph RAG (HippoRAG), CausalHippo improves the average F1 and drastically enhances non-causal abstention, which is critical for avoiding false causal claims. Compared to the pure causal graph RAG (CausalRAG), our framework improves aggregate retrieval quality (Recall@5: 0.646 vs. 0.618) and downstream answer accuracy.

6.2. Component Impact Analysis

Adaptive Controls. Comparing CausalHippo with its static counterpart, CausalHippo-Fixed, and individual ablations shows that three-layer adaptive control successfully mitigates the "noisy connectivity" of pure semantic graphs. By dynamically routing efforts, CausalHippo balances multi-hop exploration with causal verification. Turning off the confidence-based switch (no Ph3) often increases retrieval coverage (Recall@5) but can decrease the overall answer accuracy (F1). In contrast,

the two-pass adaptation (no Ph4) significantly contributes to category-level robustness. In Webis-CausalQA-22, the full system achieved superior Groundedness (0.910), confirming that the integrated gating mechanism is more robust than static schemes in high-ambiguity cases.

Synergistic Effect of Graph Integration. Integrating semantic retrieval with causal path discovery provides a stronger foundation for multi-step reasoning. Semantic graphs offer connectivity but lack grounding; causal paths provide rigor but are often sparse. This synergy allows CausalHippo to resolve complex dependencies that static or vector-only models cannot. On MuSiQue, CausalHippo outperformed CausalRAG (F1 0.361 vs. 0.350), demonstrating that the semantic context provides essential structural support for multi-step causal inference.

6.3. Hyperparameter Tuning: Sensitivity Analysis of Adaptive Gating

To verify the robustness of the Adaptive Causal Gating mechanism, we conducted a systematic sensitivity analysis using a 3×3 grid search over the hyperparameters α (Path Influence) and β (Coverage Influence) on the Causal-Reasoning-QA dataset. Table 7 summarizes the nine configurations evaluated with $\alpha, \beta \in \{0.0, 2.0, 4.0\}$.

Table 7
Sensitivity Analysis of Adaptive Causal Gating on Causal-Reasoning-QA

α	β	F1 Score	Exact Match	Recall@5	Avg. Causal Gate	Var. Causal Gate
0.0	0.0	0.3014	0.1767	**0.5233**	0.1730	0.0022
0.0	2.0	0.2974	0.1767	**0.5233**	0.2930	0.0432
0.0	**4.0**	**0.3114**	**0.1906**	0.5217	0.3896	0.1126
2.0	0.0	0.2802	0.1605	0.5217	0.2626	0.0234
2.0	2.0	0.2958	0.1837	0.5170	0.3727	0.0909
2.0	4.0	0.2837	0.1739	0.5217	0.4109	0.1305
4.0	0.0	0.2941	0.1833	**0.5233**	0.3382	0.0644
4.0	2.0	0.3063	0.1867	**0.5233**	0.3879	0.1157
4.0	4.0	0.2941	0.1873	0.5217	**0.4484**	**0.1508**

The results show that node coverage (β) is a more reliable indicator of causal relevance than the raw path count (α). The system achieved peak performance at $\alpha = 0.0, \beta = 4.0$, representing a significant improvement over the unweighted baseline ($\alpha = 0, \beta = 0$). As α and β increase, the average causal gate shifts from 0.1730 to 0.4484, with a corresponding increase in variance (0.0022 to 0.1508). This behavior matches our success criteria for "smooth gating": the system becomes more discriminative without saturating.

6.4. Research Question Synthesis

RQ1 (Adaptive routing and reliability-coverage trade-off). *Supported with caveats.* CausalHippo improves retrieval depth over CausalHippo-Fixed on the aggregate benchmark view (Recall@5 0.646 vs. 0.636; MRR@5 0.669 vs. 0.656) and attains the strongest grounding on Webis-CausalQA-22. Paired analysis confirms that CausalHippo achieves a statistically robust gain over CausalRAG in Recall@5 (+0.040, adjusted $p = 0.0088$). However, it does not statistically outperform CausalHippo-Fixed across all metrics, confirming a selective trade-off rather than universal dominance.

RQ2 (Hybrid integration vs. semantic-only/causal-only retrieval). *Supported.* Compared with HippoRAG, CausalHippo dramatically improves causal abstention behavior (Unknown Recall: 0.447 vs. 0.033, Table 4), while remaining competitive in terms of retrieval quality. Compared with CausalRAG, CausalHippo improves retrieval depth on the aggregate benchmark view and on MuSiQue, with paired analysis identifying Recall@5 as the most robust gain. This supports semantic-causal hybridization as a selective, query-dependent gain rather than a blanket improvement.

RQ3 (Stability across benchmarks and failure modes). *Supported with caveats.* `CausalHippo` achieves better cross-benchmark stability than the causal-only baseline (σ_{F1}: 0.064 vs. 0.075) and improved inter-category robustness (CV_{F1}: 0.434 vs. 0.488). Failure analysis (Section 6.5) still shows sensitivity to retrieval-chain misses and OpenIE extraction fidelity; therefore, the stability gains are meaningful but not universal.

6.5. Failure Analysis

Our systematic failure analysis revealed two primary performance boundaries. First, the most significant bottleneck in multi-hop reasoning is the **retrieval-chain miss** (affecting 85.4% of MuSiQue failures), where missing a single bridge document breaks the entire reasoning path. Our ablation experiments (Table 2) show that while the full adaptive system achieves the highest precision (MRR), the `no Ph3` variant slightly improves aggregate recall. This indicates that the **Confidence-Based Switch**, while effective at filtering noise and improving grounding, can occasionally over-filter essential semantic bridge documents when the initial causal signal is sparse. Second, the system faces a **robustness challenge** in high-noise environments where distractors share entity-level overlap with the query; this can cause the ambiguity estimator to revert to semantically heavy weights, increasing the risk of spurious correlations. Interestingly, the `no Ph3&4` configuration achieved the best category-level robustness ($CV_{F1} = 0.401$), suggesting that complex adaptive layers can sometimes introduce their own sensitivities to specific logic distributions.

At a finer granularity, the MuSiQue evaluation contained 227 incorrect answers because the system obtained EM $= 0.243$ (73/300 correct). Of these failures, 194 (85.5%) were associated with incomplete retrieval chains (*SupportRecall@5* < 1.0), whereas the remaining 33 (14.5%) occurred despite retrieving all supporting passages. The latter cases were primarily reasoning or answer-normalization errors, including entity-type substitutions, such as "American League" for "New York Yankees", and near-matches, such as "Exeter College, Oxford" for "Exeter College". This distinction is important for optimization: improving graph coverage addresses the dominant failure source, while the residual errors require better answer normalization and generation control.

On Causal-Reasoning-QA, the 190 failures associated with EM $= 0.367$ also break down into distinct error sources. Log-level inspection identified 86 false positives in which non-causal contexts were treated as causal, 39 false-negative abstentions in which genuine causal relationships were labeled "unknown", and 65 phrasing or extraction errors in which a causal link was recognized but the exact gold token was not produced. False positives were especially common in *Confounding Variable Masking* (11/12) and *Subtle Shared Influence* (10/11), while false-negative abstentions were concentrated in *Spurious Correlation* (10/15) and *Reverse Causation Potential* (7/15). These patterns suggest that future calibration should distinguish between detecting causal evidence and recognizing when evidence is insufficient, rather than applying a single global gate threshold.

6.6. Limitations

The current results should be interpreted in light of several limitations. First, the framework depends on the **fidelity of causal graph construction**: OpenIE extraction failures, especially in complex literary contexts, can force the system to rely on fallible semantic retrieval. Second, the adaptive controller introduces an explicit coverage–precision trade-off. The Confidence-Based Switch can over-filter semantic bridge documents when causal evidence is sparse, while high-noise distractors can still cause the system to over-assert causality. Third, our conclusions are limited to the three selected benchmarks and their 300-sample evaluation subsets; GraphRAG-Bench was not included because it evaluates general graph-based reasoning rather than causal discrimination and abstention. Finally, although adaptive routing was not slower than the fixed variant in our run, the shared retrieval stack and sequential LLM calls remain substantial sources of latency. Future work should therefore improve causal graph extraction, calibrate positive and negative evidence separately, parallelize independent retrieval branches, and evaluate the framework on broader domain-specific benchmarks.

7. Conclusion

This study introduced `CausalHippo`, a framework that adaptively integrates graph-based semantic retrieval with causal path discovery. Across various RAG configurations, including comprehensive ablations and three benchmarks, the results show that hybrid semantic-causal retrieval can improve the balance between retrieval coverage and causal reliability, but the gains are selective rather than uniform. In particular, the primary `CausalHippo` system achieved strong grounding behavior on Webis-CausalQA-22 (Faithfulness 0.940, Groundedness 0.910) and competitive causal discrimination on Causal-Reasoning-QA (`F1(True)` 0.472, Unknown Recall 0.447). At the same time, the analyses indicate that the adaptive controller is not a blanket improvement over the fixed hybrid variant, and that the strongest gain over `CausalRAG` is retrieval depth. Nevertheless, this limitation does not negate the utility of the proposed approach: its value lies in adapting retrieval behavior to query complexity and evidence density, thereby improving causal abstention and grounding relative to graph-only retrieval while retaining broad semantic coverage. Future work will focus on stronger causal graph extraction, latency reduction, and integration with formal Semantic Web reasoning components.

The source code and experimental materials are publicly available at https://github.com/takanori-ugai/CausalHippo.

Acknowledgements

This study is based on results obtained from a project under the Programs for Bridging the Gap between R&D and the IDeal Society (Society 5.0) and Generating Economic and Social Value (BRIDGE), "R7-H05: Demonstration, Research, and Development for Safety in AI and Robotics Fields / Research and Development for Strengthening AI Safety in Industrial Domains", implemented by the Cabinet Office, Government of Japan.

Declaration of use of Generative AI

During this study, the authors used generative AI to translate and proofread the manuscript and to translate Python source code to Kotlin. The authors reviewed and edited all content and take full responsibility for the manuscript's integrity.

References

[1] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitansky, R. O. Ness, J. Larson, From local to global: A graph rag approach to query-focused summarization, arXiv preprint arXiv:2404.16130 (2024). URL: https://arxiv.org/abs/2404.16130.

[2] B. J. Gutiérrez, Y. Shu, G. Yu, Y. Chen, Y. Su, Hipporag: Neurobiologically inspired long-term memory for large language models, in: Advances in Neural Information Processing Systems (NeurIPS), 2024. URL: https://arxiv.org/abs/2405.14831.

[3] N. Wang, X. Han, J. Singh, J. Ma, V. Chaudhary, CausalRAG: Integrating causal graphs into retrieval-augmented generation, in: Findings of the Association for Computational Linguistics: ACL 2025, Association for Computational Linguistics, Vienna, Austria, 2025, pp. 22680–22693. URL: https://aclanthology.org/2025.findings-acl.1165. doi:10.18653/v1/2025.findings-acl.1165.

[4] H. Trivedi, N. Balasubramanian, T. Khot, A. Sabharwal, Musique: Multihop questions via single-hop question composition, Transactions of the Association for Computational Linguistics 10 (2022) 539–554. URL: https://aclanthology.org/2022.tacl-1.31. doi:10.1162/tacl_a_00475.

[5] Z. Guo, Y. Xia, Lianghao rival, T. Ao, C. Huang, Lightrag: Simple and fast retrieval-augmented generation, in: Findings of the Association for Computational Linguistics: EMNLP 2024, 2024, pp. 10746–10761.

[6] B. Chen, Z. Guo, Z. Yang, Y. Chen, J. Chen, Z. Liu, C. Shi, C. Yang, Pathrag: Pruning graph-based retrieval augmented generation with relational paths, arXiv preprint arXiv:2502.14902 (2025).

[7] H. Qin, C. Wei, Y. Chen, Y. Wang, Counterfactual reasoning for retrieval-augmented generation,

in: The 14th International Conference on Learning Representations (ICLR), 2026. URL: https://openreview.net/forum?id=pDbytHwHVJ.

[8] E. Khatibi, Z. Wang, A. M. Rahmani, Cdf-rag: Causal dynamic feedback for adaptive retrieval-augmented generation, 2025. URL: https://arxiv.org/abs/2504.12560. arXiv:2504.12560.

[9] H. Luo, J. Zhang, C. Li, Causal graphs meet thoughts: Enhancing complex reasoning in graph-augmented llms, 2025. URL: https://arxiv.org/abs/2501.14892. arXiv:2501.14892.

[10] C. Samarajeewa, D. De Silva, E. Osipov, D. Alahakoon, M. Manic, Causal reasoning in large language models using causal graph retrieval augmented generation, in: 2024 16th International Conference on Human System Interaction (HSI), IEEE, 2024, pp. 1–6. doi:10.1109/HSI61632.2024.10613566.

[11] H.-L. Hsu, et al., Dat: Dynamic alpha tuning for hybrid retrieval in retrieval-augmented generation, 2025. URL: https://arxiv.org/abs/2503.23013. arXiv:2503.23013.

[12] S. Jeong, J. Baek, S. Cho, S. J. Hwang, J. C. Park, Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 7036–7050. URL: https://aclanthology.org/2024.naacl-long.391.

[13] S. Ozaki, Y. Kato, S. Feng, M. Tomita, K. Hayashi, W. Hashimoto, R. Obara, M. Oyamada, K. Hayashi, H. Kamigaito, T. Watanabe, Understanding the impact of confidence in retrieval augmented generation: A case study in the medical domain, in: Proceedings of the 24th Workshop on Biomedical Language Processing, Association for Computational Linguistics, Vienna, Austria, 2025, pp. 1–17. URL: https://aclanthology.org/2025.bionlp-1.1. doi:10.18653/v1/2025.bionlp-1.1.

[14] J. Wu, L. Yang, Z. Wang, M. Okumura, Y. Zhang, Cofca: A step-wise counterfactual multi-hop qa benchmark, in: The Thirteenth International Conference on Learning Representations (ICLR), 2025.

[15] A. Bondarenko, M. Wolska, S. Heindorf, L. Blübaum, A.-C. Ngonga Ngomo, B. Stein, P. Braslavski, M. Hagen, M. Potthast, CausalQA: A benchmark for causal question answering, in: N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, S.-H. Na (Eds.), Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 3296–3308. URL: https://aclanthology.org/2022.coling-1.291/.

[16] Y. Xiao, J. Dong, C. Zhou, S. Dong, Q.-w. Zhang, D. Yin, X. Sun, X. Huang, Graphrag-bench: Challenging domain-specific reasoning for evaluating graph retrieval-augmented generation, arXiv preprint arXiv:2506.02404 (2025). URL: https://arxiv.org/abs/2506.02404.

[17] D. Khashabi, A. Ng, T. Khot, A. Sabharwal, H. Hajishirzi, C. Callison-Burch, Gooaq: Open question answering with diverse answer types, arXiv preprint arXiv:2104.08678 (2021).

[18] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, M. Rosenberg, X. Song, A. Stoica, S. Tiwary, T. Wang, Ms marco: A human generated machine reading comprehension dataset, 2018. URL: https://arxiv.org/abs/1611.09268. arXiv:1611.09268.

[19] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, SQuAD: 100,000+ questions for machine comprehension of text, in: J. Su, K. Duh, X. Carreras (Eds.), Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 2383–2392. URL: https://aclanthology.org/D16-1264/. doi:10.18653/v1/D16-1264.

[20] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, in: International Conference on Learning Representations, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

[21] VibrantLabs, Ragas: Supercharge your llm application evaluations, https://github.com/vibrantlabsai/ragas, 2024.