Iterative Evidence Retrieval and Abductive Reasoning: A Dual-Agent Framework for Causal Inference

Simone Columbaro, Irene Bartolini, Alex Drapant, Lorenzo Brasolin, Shamim Zare
Politecnico di Torino, Turin, Italy
GitHub Repository: LLM-abductive-event-reasoning

Abstract

Abductive event reasoning requires identifying the most plausible causes for a given event, a task that demands deep causal understanding and precise evidence verification. Standard Retrieval-Augmented Generation (RAG) approaches often fail in this domain because a single retrieval step may miss granular details—such as specific timestamps or actor sequences—necessary to establish a valid causal link. In this paper, we propose a specialized agentic framework designed for multiple-choice causal inference. Our system employs a dual-agent architecture: a Search Agent that performs iterative gap analysis to gather missing evidence, and a Causal Agent that applies abductive logic based on temporal precedence and counterfactual reasoning. To maximize recall, we implement a hybrid retrieval engine using multi-view semantic search and heuristic keyword boosting. Additionally, we explore an extension using Causal Knowledge Graphs to capture multi-hop relational dependencies. Experimental results on the SemEval 2026 Task 12 dataset show that our agentic approach effectively mitigates information gaps. While graph-based methods offer theoretical advantages, we find that our agentic framework currently provides more robust performance across diverse topics due to lower noise in the retrieved context.

1 Introduction

Abductive reasoning, the process of identifying the most plausible explanation for observed events, remains a significant challenge for Large Language Models (LLMs). While general performance has improved, complex causal inference requires verifying precise temporal details and actor sequences that standard architectures often overlook. In this work, we evaluate these reasoning capabilities using the SemEval 2026 Task 12 dataset.

Current approaches largely rely on static Retrieval-Augmented Generation (RAG). However, our preliminary analysis reveals that even high-performing models suffer from hallucinations when restricted to a single retrieval pass, as initial queries often fail to capture the nuanced evidence required to distinguish correlation from causation. To address these limitations, we propose and evaluate two distinct architectural paradigms for causal inference. Our contributions are:

- **Decoupled Agentic Architecture:** We introduce a modular framework separating evidence gathering from inference. Powered by a **Hybrid Retrieval Strategy**, it employs a specialized Search Agent for targeted gap analysis to ensure precise context for causal deduction.

- **CausalRAG Analysis:** We implement a graph-based extension utilizing **Causal Knowledge Graphs (CKG)** to capture multi-hop dependencies. We benchmark this against our text-based agent, providing a critical analysis of the trade-offs between structural reasoning and the noise sensitivity of automated graph construction.

2 Related Works

The evolution of Large Language Models (LLMs) has significantly advanced the field of natural language understanding, yet complex reasoning remains a persistent challenge. Recent research has focused on enhancing these capabilities through structured frameworks and external knowledge integration.

2.1 Retrieval-Augmented Generation (RAG)

Standard RAG (Lewis et al., 2021) paradigms typically rely on a static "retrieve-then-generate" heuristic. This approach presupposes that a single retrieval phase can capture all necessary context. However, in causal inference tasks, this often becomes a bottleneck, leading to "hallucinations" when models attempt to bridge information gaps without sufficient evidence. Our work addresses this by implementing a hybrid retrieval engine that uses multi-view semantic search and heuristic keyword boosting to improve document recall.

2.2 Agentic Reasoning

Our work builds on the *ReAct* paradigm (Yao et al., 2023), interleaving reasoning traces with actions. Following the principle of functional decoupling (Dua et al., 2022), we decompose the workflow into specialized Evidence Evaluator and Causal Reasoner components. This separation reduces cognitive load and prevents the conflation of evidence verification with final causal inference.

2.3 Graph-Based Methods

Recent frameworks like *CausalRAG* (Wang et al., 2025) utilize Causal Knowledge Graphs (CKG) to model events and causal-temporal edges. By traversing these topologies, systems can recover multi-hop causal chains that standard RAG might miss. We implement this as an extension to evaluate the trade-offs between structural connectivity and the noise introduced during graph synthesis.

3 Methodology

We propose a dual-agent framework that integrates iterative information retrieval with abductive reasoning to solve complex causal inference tasks. Unlike static Retrieval-Augmented Generation (RAG) pipelines, our architecture actively verifies evidence sufficiency before attempting a conclusion.

3.1 System Overview

The system is built upon the `Qwen2.5-7B-Instruct` Large Language Model (LLM), quantized to 4-bit (NF4) for computational efficiency. The workflow operates in three distinct phases:

1. **Hybrid Retrieval:** A dense-sparse retrieval module gathers initial context using multi-view embeddings and keyword boosting, followed by cross-encoder reranking to filter irrelevant candidates.

2. **Iterative Evidence Verification:** A dedicated *Search Agent* evaluates context sufficiency and dynamically queries for missing data.

3. **Abductive Inference:** A *Causal Agent* reasons over the finalized context to identify the valid causal set.

3.2 Hybrid Retrieval Engine

To capture high-granularity semantic dependencies, documents are segmented into 800-token chunks with a 256-token overlap. We implement a hybrid scoring function that combines semantic understanding with exact keyword matching.

Multi-View Semantic Search

We employ a multi-view query strategy using the `BAAI/bge-base-en-v1.5` embedding model. This design addresses the information asymmetry inherent in causal inference: searching for the target event alone often yields context that is too broad to distinguish between specific causes. **Analogous to a test-taker who cross-references specific options against their knowledge base rather than answering an open-ended prompt**, our system treats each candidate option as an active retrieval cue. This allows the model to "work backwards," gathering precise evidence to verify or refute each specific hypothesis individually. To operationalize this, for a pair consisting of a question Q and an option O_i, we generate three distinct embedding views:

- *Causal View:* "Evidence that O_i is the cause of Q."
- *Entity View:* The textual representation of O_i.
- *Event View:* The textual representation of the target event Q.

The semantic score S_{sem} for a document chunk d is defined as the maximum cosine similarity across these three views.

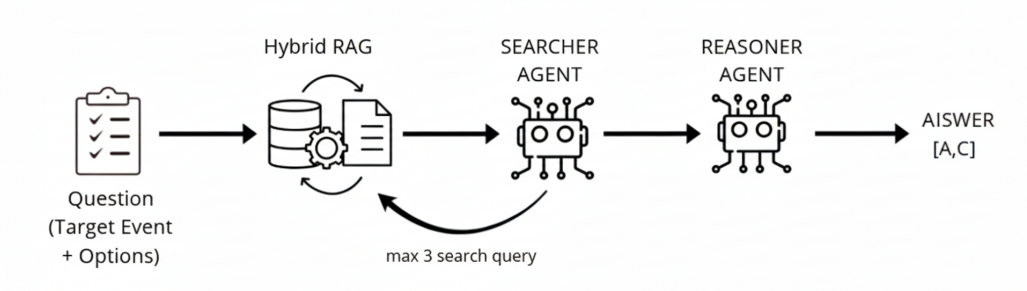

Figure 1: The decoupled Agentic Workflow, illustrating the separation between the Evidence Evaluator and the Causal Reasoner.

Heuristic Keyword Boosting

To ensure the retrieval of specific named entities (e.g., dates, actors), we apply a dynamic boosting algorithm. Let K_{opt} be the set of keywords extracted from option O_i. The final hybrid score $S(d, O_i)$ is calculated as:

$$S(d, O_i) = \max_{v \in \text{views}} \big(\text{sim}(\mathbf{q}_v, d)\big) + \alpha \sum_{w \in K_{opt}} \mathbb{I}(w \in d) \quad (1)$$

where $\mathbb{I}$ is the indicator function and $\alpha = 0.2$ is a hyperparameter balancing semantic relevance and lexical overlap.

Reranking

The top-k candidates ($k = 20$) are reranked using a Cross-Encoder (`ms-marco-MiniLM-L-6-v2`). This step filters out candidates that are semantically related to the topic but lack the specific relational context required for causal verification.

3.3 Agentic Workflow

To mitigate the hallucinations common in monolithic RAG systems, we adopt a decoupled architecture inspired by the ReAct framework (Yao et al., 2023). We separate the process into two specialized roles to prevent the model from conflating evidence retrieval with causal derivation.

The Search Agent (Evidence Evaluator)

This module acts as an iterative investigator. Initialized with the reranked documents, it performs a *Gap Analysis* to determine if the current context supports a definitive conclusion. If critical information is missing, the agent generates structured search actions in the format `Action: Search['query']`. This iterative cycle continues until the agent outputs `SUFFICIENT` or reaches a configurable maximum depth K. In our experiments, we set $K = 1$, relying on the agent's ability to issue the most relevant queries to resolve ambiguity in the first refinement step.

The Causal Agent (Inference Engine)

The Causal Agent receives the finalized context and performs abductive classification. We enforce a strict "Archival Rule" via the system prompt, requiring the agent to interpret prospective statements in documents (e.g., "plans to") as valid antecedents for subsequent events. The agent selects options based on three logical criteria:

- **Temporal Precedence:** The cause must chronologically precede the event.
- **Mechanism:** A direct explanatory link must exist.
- **Counterfactual Validity:** The event would not have occurred without the cause.

The architecture supports multi-label classification, allowing the agent to return a set of answers (e.g., `[A, C]`) if multiple independent causes are verified.

4 Experiments

4.1 Dataset and Evaluation Metrics

We evaluated our framework on the **SemEval 2026 Task 12** dataset, which consists of complex multiple-choice questions requiring abductive reasoning over event-based corpora.

The evaluation metric accounts for partial correctness in multi-answer scenarios:

- **1.0 Point:** The set of predicted answers matches the ground truth exactly.

- **0.5 Points:** The predicted answers are a valid subset of the ground truth (no incorrect options selected).
- **0.0 Points:** The prediction includes any incorrect option.

4.2 Implementation Details

All experiments were conducted on a single NVIDIA Tesla P100 GPU. The system is powered by the `Qwen2.5-7B-Instruct` model, loaded with 4-bit NF4 quantization to optimize memory efficiency.

Retrieval Configuration

We utilized the `bge-base-en-v1.5` embedding model. Documents were partitioned into 800-token chunks with a 256-token overlap. For the retrieval pipeline, we retrieved $k = 20$ candidates per option during the initial search, reranking them to select the top $k_{final} = 2$ chunks for the final context window.

Agent Configuration

The Search Agent was configured with a maximum depth of $K = 1$ iteration, allowing for a single batch of targeted queries to resolve ambiguity while minimizing latency. Heuristic keyword boosting was applied with $\alpha = 0.2$ to ensure high recall of named entities and temporal markers.

4.3 Overall Performance

Table 1 summarizes the global performance of the architecture. Across the full test set of 400 questions, the model achieved a Total Score of 256.00, yielding an average accuracy of **0.64**.

Table 1: Overall Performance of the Agentic Architecture.

Metric	Value
Total Questions	400
Correct Answers (Exact)	212
Partial Answers (Subset)	88
Incorrect Answers	100
Total Score	**256.0**
Average Score	**0.64**

4.4 Behavioral Analysis

A granular analysis of the evaluation logs reveals distinct behavioral characteristics of the decoupled architecture, specifically regarding negative constraints and answer cardinality.

4.4.1 Aversion to the "None of the Other" Option

The model exhibits a significant bias against negative conclusions. In scenarios where the correct answer was "None of the Other" (NOTO), the model achieved a Recall of only 56.5% ($N = 46$). This result illustrates a well-known and persistent limitation in generative AI: the struggle to identify the absence of evidence. As recently analyzed by Kirichenko et al. (Kirichenko et al., 2025), this "forced resolution" bias remains acute even in advanced reasoning models, which often prioritize completing the logical chain over admitting ignorance. Our findings suggest that the agent tends to prioritize the selection of an explicit causal candidate, favoring potential associations over the null hypothesis when evidence is ambiguous.

4.4.2 Decoupling Format Recognition from Content Accuracy

To understand the model's behavior, we decoupled the evaluation into two distinct dimensions: *Format Recognition* (the ability to correctly identify answer cardinality) and *Content Accuracy* (the semantic correctness of the answer).

In terms of **Format Recognition**, the system exhibits a strong bias towards Single-Choice outputs (see Table 2). The model achieved a high Recall of 85.2% for identifying Single-Choice formats, but this came at the cost of significant under-generation in complex scenarios. In 80 instances, the model incorrectly treated a Multi-Choice question as a Single-Choice one, resulting in a low Recall of 57.9% for the Multi-Choice format.

However, this format confusion did not drastically impact the **Content Accuracy**, which remained remarkably stable (0.65 for Single vs. 0.63 for Multi). This stability can be attributed to an emergent **"Safety Bias."** Since the evaluation metric awards partial credit (0.5 points) for subset matches, the model's tendency to halt after finding a single `SUFFICIENT` proof acts as a risk-minimization strategy. Much like a student avoiding penalties for incorrect guesses, the model prefers to secure points with one high-confidence answer rather than risking

the selection of uncertain additional options. This effectively buffers the overall score against the lower format recall.

Table 2: Comparison of Format Detection Capability vs. Semantic Score.

Metric	Single-Choice	Multi-Choice
Format Recognition (Cardinality)		
Format Precision	0.691	0.780
Format Recall	0.852	0.579
Format Confusion	31 (Over-gen)	80 (Under-gen)
Content Accuracy (Score)		
Average Score	**0.652**	**0.626**

5 CausalRAG Extension

In addition to our primary strategy, we implemented a specialized CausalRAG (Wang et al., 2025) framework. The foundational premise of this approach lies in the Graph-based Preprocessing phase. Unlike traditional RAG systems that treat document corpora as collections of isolated text chunks, CausalRAG organizes information into a Causal Knowledge Graph (CKG). Within this architecture, nodes represent discrete events or states, while edges explicitly encode the causal-temporal relationships extracted from the source material. We hypothesized that this technique would be particularly effective for our task because this preprocessing step is fundamental for Context Augmentation. By transforming unstructured text into a structured topology, the system can traverse relational edges to recover multi-hop causal chains. This capability allows the model to bridge the gap between a trigger event and a distal consequence, even when they are separated by multiple intermediate steps in the narrative. Consequently, this approach effectively mitigates the "context window fragmentation" typical of standard RAG, where the lack of explicit links between retrieved chunks often obscures long-range dependencies. To retrieve the optimal context, we employ the following strategy: for a given target event and its associated options, we select the top k nodes based on a combined score from **FAISS** (dense semantic similarity) and **BM25** (sparse lexical matching), the nodes with the higher total score are retrieved. Once these "seed nodes" are identified, the system extracts a relevant subgraph through a bidirectional traversal logic:

- **Backward Expansion**: Starting from the target event nodes, the system traverses outgoing edges to identify potential causes
- **Forward Expansion:** Starting from the options nodes, the system traverses incoming edges to trace back to potential consequences.

This traversal is constrained to a maximum of s hops to maintain causal relevance and prevent context dilution.

From the resulting subgraph, all valid causal chains are extracted. These raw paths are then processed by a specialized agent, the Synthesizer. This agent transforms the structured graph data into a coherent, synthetic textual summary. By distilling complex multi-hop relationships into a natural language narrative, the Synthesizer provides the reasoning model with a streamlined and causally-dense context, specifically optimized for decision-making.

5.1 Experiments

The evaluation of the CausalRAG approach was restricted to the first twenty Topic IDs within the dataset. This constraint was necessitated by the high computational overhead of the preprocessing phase, specifically the generation of a dedicated Causal Knowledge Graph for each topic, which proved prohibitively resource-intensive given our available hardware. Counter-intuitively, our preliminary results indicate that this graph-based technique underperforms relative to the baseline RAG method. This performance gap suggests that while the structural topology of a graph offers theoretical advantages for multi-hop reasoning, the complexity of the graph construction or the potential noise introduced during the extraction of causal edges may currently outweigh the benefits of structured context. As illustrated in Table 3, while the CausalRAG strategy demonstrates high effectiveness in specific topics, the overall trend is inversely proportional to our initial expectations. Notably, in Topics 14 and 19, the CausalRAG model exhibits significantly poor performance. This discrepancy is likely attributable to graph density and noise within the preprocessed knowledge structures. In these instances, the extraction of spurious causal links or the inclusion of irrelevant nodes likely

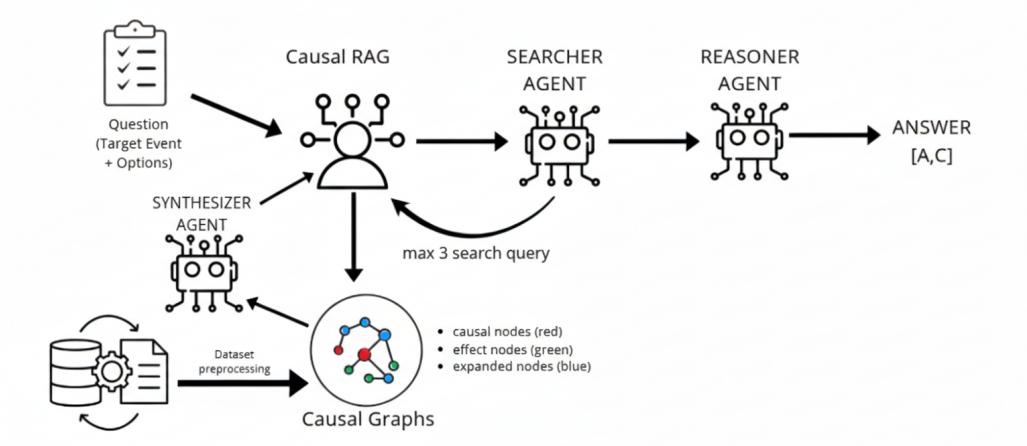

Figure 2: Agent Workflow with Causal RAG

Table 3: Comparative scores per topic (Hybrid vs. Causal RAG).

ID	Hybrid RAG	Causal RAG
1	7.00	**8.00**
2	**7.50**	5.50
3	**11.00**	10.50
4	5.50	**7.00**
5	3.00	**4.00**
6	**1.00**	0.00
7	**5.00**	4.50
8	5.00	**7.50**
9	1.00	**3.00**
10	**6.50**	5.50
11	**3.00**	2.00
12	**8.50**	5.50
13	**8.50**	6.50
14	**3.00**	0.00
15	5.50	**7.00**
16	**6.00**	5.50
17	2.00	2.00
18	**6.50**	5.00
19	**16.00**	10.50
20	7.50	**9.00**
TOT	**119.00**	**108.50**

led to contextual drift, where the synthesized summary distracted the model from the core evidence rather than reinforcing it. These results suggest that the quality of the Causal Knowledge Graph is highly topic-dependent, requiring more robust filtering mechanisms to ensure edge reliability.

6 Conclusion and Discussion

In this work, we introduced an agentic framework for abductive event reasoning, demonstrating that separating evidence verification from causal inference yields a stable, conservative, reasoning engine. Our results on the SemEval 2026 Task 12 dataset highlight that reasoning in LLMs is an emergent property heavily dependent on architectural context.

A primary finding of this research is the emergent "Safety Bias" of the decoupled architecture. While the system frequently undergenerated answers in multi-cause scenarios, often defaulting to single-choice formats, it maintained high semantic accuracy on the options it did select. This behavior suggests that our model act as risk-minimizers, preferring high-confidence partial proofs over the noise of exhaustive retrieval. However, this cautious approach backfired when the answer was 'None of the Other.' In these cases, the model struggled to admit that no cause existed. Instead of accepting that none of the options worked, it often forced a connection where there wasn't one, effectively guessing rather than staying silent.

Finally, our comparative analysis with the graph-based baseline (CausalRAG) challenges the assumption that structural complexity intrinsically yields better performance. Despite its theoretical benefits, CausalRAG failed to outperform the text-based baseline in our experiments. We attribute this to "contextual drift," where spurious connections within the automatically generated graphs distracted the model rather than reinforcing the core evidence. This finding suggests that navigating dense, noisy graph structures may require agents with significantly higher capacity than the 7B-parameter model employed here. While a larger model might effectively filter this structural noise, our results indicate that for efficient, mid-sized agents, the **Hybrid Retrieval Strategy coupled with the Multi-Agent architecture** remains the superior choice, of-

fering robustness without the excessive cognitive overhead required to parse imperfect graphs.

References

Dheeru Dua, Shivanshu Gupta, Sameer Singh, and Matt Gardner. 2022. Successive prompting for decomposing complex questions. *Preprint*, arXiv:2212.04092.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J. Bell. 2025. Abstentionbench: Reasoning llms fail on unanswerable questions. *Preprint*, arXiv:2506.09038.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Nengbo Wang, Xiaotian Han, Jagdip Singh, Jing Ma, and Vipin Chaudhary. 2025. Causalrag: Integrating causal graphs into retrieval-augmented generation. In *Findings of the Association for Computational Linguistics (ACL)*. ArXiv preprint arXiv:2503.19878.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. *Preprint*, arXiv:2210.03629.