A Multi-Agent Framework and Challenging Benchmark for Causal Relationship Extraction

Anonymous ACL submission

Abstract

Extracting accurate causal relationships from text is crucial for developing Causal Knowledge Graphs (CKGs), which support advanced reasoning and decision-making. Traditional approaches often struggle with linguistic ambiguity and the complexity of natural language. Existing benchmarks, like SemEval-2007 Task 4, primarily feature short sentences, limiting the evaluation of modern Large Language Models (LLMs) in longer contexts.

In this study, we present two key contributions: (1) a novel Multi-Agent Causal Extraction System that employs a multistage verification process with a Judge agent for relationship extraction and a Critic agent for reasoning verification; and (2) a Categorized Benchmark Dataset containing 10,000 long-context examples across 20 causal and non-causal categories, including "deceptive correlations" to test models' capabilities.

Our experiments reveal that while our system achieves human-level performance (89.66%) on SemEval-2007, accuracy drops to 70.00% on our benchmark, highlighting the need for more rigorous evaluations in causal reasoning.

1 Introduction

Natural Language Processing (NLP) has made significant strides in information extraction; however, accurately identifying *causal relationships* remains a formidable challenge. Precision in causality extraction is critical for building Causal Knowledge Graphs (CKGs) (Jaimini and Sheth, 2022; Fujitsu Limited, 2024), which map the interactions between real-world events to support medical diagnosis, financial risk analysis, and scientific discovery.

Automating CKG construction from unstructured text is difficult because traditional rule-based and statistical approaches cannot capture the diversity of human languages. Although Large Language Models (LLMs) have improved performance, they remain prone to "hallucinations" and often confuse correlation with causation (Joshi et al., 2024; Chi et al., 2025). A significant bottleneck is the saturation of standard benchmarks such as SemEval-2007 Task 4, which primarily consist of short, single-sentence examples. Such benchmarks allow models to rely on surface-level keyword matching rather than deep causal reasoning. In contrast, our dataset introduces 'deceptive correlations' in at least five sentences per context, forcing models to distinguish between mere temporal co-occurrence and true latent causal mechanisms.

To address these limitations, we propose a unified framework comprising the following:

1. **Multi-Agent Extraction System:** A collaborative architecture where an initial agent extracts potential links and a secondary "Critic" agent verifies the logical consistency of the output, mimicking a peer-review process.

2. **Challenging Benchmark Dataset:** A new dataset of 10,000 long-context examples (over five sentences each) categorized into 20 distinct causal and non-causal relationship types, specifically designed to be statistically deceptive.

Our experimental evaluation demonstrates that while modern LLMs achieve near-human performance on standard benchmarks, they struggle significantly with long-context and deceptive scenarios introduced in our new dataset. Specifically, our multi-agent framework demonstrates greater robustness than single-agent baselines, although the 19.66% accuracy drop on the new benchmark highlights substantial room for algorithmic improvement.

The paper is organized as follows: Section 2 reviews **related work** on causal extraction and LLM reasoning. Sections 3 and 4 detail our **Multi-Agent**

architecture and **benchmark dataset construction**. Sections 5 and 6 present **experimental results, ablation studies, and discussions**. Finally, Sections 7–9 address **limitations, ethical considerations, and conclusions**.

2 Related Work

2.1 Causal Relationship Extraction Methods

Causal relationship extraction has evolved from rule-based systems to sophisticated deep-learning models. Early approaches relied heavily on handcrafted patterns and statistical features, which limited their adaptability to the diversity of natural languages. The introduction of deep learning, particularly BERT-based models and Graph Convolutional Networks (GCNs), marked a significant performance leap. Domain-specific models, such as BioBERT and SciBERT, have proven effective in specialized fields (Yang et al., 2021; Akkasi and Moens, 2021). However, these methods still face challenges in handling complex expressions, implicit causality across sentences, and a scarcity of large-scale annotated datasets.

2.2 Causal Knowledge Graph (CKG) Construction

Recent frameworks have focused on automating CKG construction to enable reasoning. *CausalKG* (Jaimini and Sheth, 2022) integrates Causal Bayesian Networks (CBNs) with ontologies to support interventional and counterfactual reasoning. Similarly, Fujitsu's Causal Knowledge Graph framework (Fujitsu Limited, 2024) employs a multiphase approach, using LLMs to extract causal triples from documents. It combines them with statistical causal discovery from numerical data to uncover unknown causal structures. In the medical domain, Lyu et al. (Lyu et al., 2023) proposed a framework for constructing large-scale CKGs linking drugs and diseases by incorporating formal causal definitions and probability distributions derived from cohort data (e.g., the UK Biobank). These efforts highlight the growing need for scalable, automated extraction methods for use in complex reasoning systems.

2.3 LLMs and Causal Inference: Capabilities and Limits

The application of LLMs to causal inference has been debated extensively. **Capabilities:** Kiciman et al. (Kiciman et al., 2024) demonstrated that LLMs, such as GPT-4, can go beyond pattern recognition to perform knowledge-based causal reasoning, achieving high accuracy (97%) on pairwise causal discovery tasks (e.g., Tubingen dataset). This suggests that LLMs can use metadata and context to infer relationships that are difficult to obtain with purely statistical methods. **Limitations:** Conversely, recent studies warn that LLMs may act as "causal parrots." Jin et al. (Jin et al., 2024b) showed that LLMs perform poorly on the "Corr2Cause" task, failing to infer causation from correlation when the data contradicts their pre-training priors. Joshi et al. (Joshi et al., 2024) and Chi et al. (Chi et al., 2025) found that LLMs are sensitive to graph encoding and are prone to logical fallacies, often failing to identify hidden confounders or distinguish correlation from causation in novel scenarios. These mixed findings underscore the necessity of a verification mechanism, such as our multi-agent critic, to filter out hallucinations.

2.4 Benchmarks for Causal Reasoning

The evaluation of causal reasoning requires robust benchmarks. Although datasets such as COPA, e-CARE, and CausalNet have been widely used, they often lack the complexity of real-world documents. Newer benchmarks such as *CausalProbe-2024* (Chen et al., 2025) reveal that LLMs suffer from performance drops on fresh corpora that are not seen during training. Additionally, *CausalPitfalls* (Du et al., 2025) benchmarks LLMs against statistical biases, showing their struggle with confounding variables. Our work builds on these insights but addresses a specific gap: the need for a long-context, categorized dataset that explicitly tests the ability to reject "deceptive" non-causal relationships (e.g., spurious correlations), moving beyond the short-sentence limitations of the standard SemEval-2007 Task 4 (Girju et al., 2007).

2.5 Counterfactual Reasoning and Prompting

Evaluating **counterfactual reasoning** is essential to move beyond "causal parrots" that rely on surface-level statistical patterns (Jin et al., 2024a; Joshi et al., 2024). While existing benchmarks like CRASS (Frohberg and Binder, 2022; Kiciman et al., 2024) employ counterfactual prompting to test logical consistency, they primarily focus on short-sentence contexts. Our work extends this scope by integrating a "Counterfactual Reasoning Required" category within deceptive long-context narratives, thereby necessitating more robust causal

inference.

3 Methodology: Multi-Agent Extraction System

Our Multi-Agent system architecture employs four primary components to ensure high-precision extraction.

3.1 System Architecture

The system, as depicted in Figure 1, consists of four primary components: an Entity Extractor that identifies potential causal entities, a Causal Relationship Judge that reasons about their connections, a Validity Critic that acts as a logical adversary to filter errors, and an aggregator that makes the final determination.

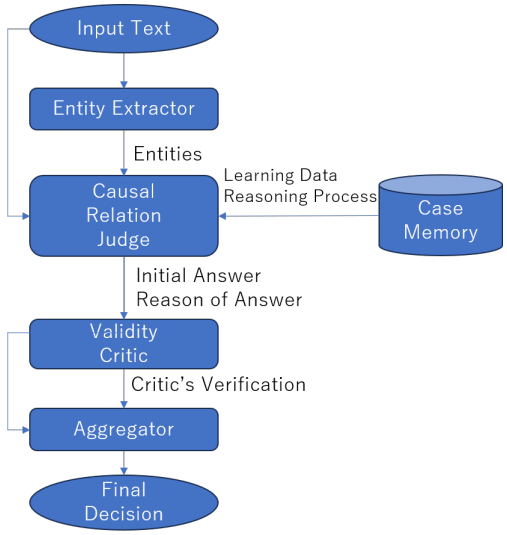

Figure 1: The Multi-Agent System architecture featuring a multistage verification pipeline.

- **Entity Extractor:** Identifies potential causal pairs (e_1, e_2) using a NER-based approach.
- **Causal Relationship Judge:** Analyzes the text and entities using **Case Memory** (few-shot examples). It outputs a *Reasoning Process* and an *Initial Answer*.
- **Validity Critic:** Acts as a logical adversary. It reviews the Judge's logic to identify hallucinations or instances where correlation is mistaken for causation.
- **Aggregator:** Reconciles the outputs of the Judge and Critic to produce the *Final Decision*.

3.2 Prompt Engineering Strategy

We employed a sophisticated prompting strategy to maximize the LLM performance.

- **Chain of Thought (CoT):** We explicitly instruct the model to output a "Reasoning Process" before the final answer. For example, the prompt asks the model to "Outline linguistic cues, logical deductions, or absence of evidence considered."
- **Few-Shot Learning:** The system utilizes a "Case Memory" containing examples of correct and incorrect reasoning. By providing the model with examples of how to distinguish subtle non-causal correlations, we significantly improved its discrimination ability.

An example of the prompt used for the Causal Relationship Judge is provided in Listing 1 in Appendix A.1.

Similarly, the detailed prompt for the Validity Critic, which guides its role as a logical adversary, is provided in Listing 2 in Appendix A.1

If the Judge and Critic disagree, the aggregator is invoked. The detailed prompt for the aggregator, designed to reconcile conflicting inputs, is provided in Listing 3 in Appendix A.1.

Finally, an illustrative example of the "Reasoning Process" stored in the Case Memory, which guides the models' logical analysis, is provided in Listing 4 in Appendix A.1

4 The Categorized Benchmark Dataset

To rigorously test the system's ability to handle complex and deceptive scenarios, we developed a new benchmark dataset that surpasses existing resources in difficulty and scale.

4.1 Dataset Construction

We utilized GPT-4.1 to generate a large-scale dataset of 10,000 items (5,000 positive/causal and 5,000 negative/non-causal). Specific constraints guided the generation process:

- **Long Context:** Each example consists of at least five sentences, requiring the model to maintain context over a longer span than SemEval-2007.
- **Deceptive Correlations:** Negative examples were generated with prompts explicitly instructing the model to create "plausible but

deceptive correlations." For instance, a "Descriptive Statement, Not Causal" prompt ensures that entities appear related (e.g., temporally) but lack a causal mechanism.

An example of the prompt used to generate positive data (causal relationships) is provided in Listing 5 in Appendix A.2

For negative examples (non-causal relationships), we utilized prompts designed to generate deceptive correlations, such as the one provided in Listing 6 in Appendix A.2

The resulting dataset follows a specific JSON format to facilitate automated parsing. The detailed JSON structure and an example entry are provided in Listing 7 in Appendix A.

4.2 Categories and Difficulty

The dataset is structured into 20 distinct categories (10 Positive, 10 Negative) to allow for fine-grained error analysis. We established a difficulty scale (Easy, Medium, Hard) to evaluate the model performance across different levels of complexity (see Tables 1 and 2).

Category	Difficulty
Simple, Direct Causation	Easy
Common Sense Causation	Easy
Reverse Causation Potential	Medium
Spurious Correlation	Medium
Intermediary/Chain Causation	Medium
Conditional Causation	Medium
Preventative Causation	Medium
Confounding & Multiple Causes	Hard
Counterfactual Reasoning Required	Hard
Subtle, Context-Dependent Causation	Hard

Table 1: Positive Categories and Difficulty Levels

Category (Negative)	Difficulty
Co-occurrence, No Mechanism	Easy
Random Co-occurrence	Easy
Shared Trend, No Direct Link	Medium
Temporal Proximity	Medium
Reverse Correlation	Medium
Weak Correlation, No Plausible Mechanism	Medium
Descriptive Statement, Not Causal	Medium
Subtle Shared Influence	Hard
Complex System, No Direct Link	Hard
Confounding Variable Masking	Hard

Table 2: Negative Categories and Difficulty Levels

4.3 Diversity and Uniformity Analysis via PCA

To ensure that the dataset was not biased toward specific topics or linguistic patterns, we assessed its uniformity using Principal Component Analysis (PCA) applied to text embeddings. We employed `text-embedding-3-small` (OpenAI) to transform the textual data into vector representations.

We defined a new metric of uniformity based on the distribution of variance across the principal components. If a dataset is highly uniform, the variance should be distributed relatively evenly across many components rather than being concentrated in the first few.

Our analysis yielded a cumulative contribution ratio of 0.366 for the top 10 principal components. The contribution ratio of the first principal component was only 0.0741. This indicates a high degree of uniformity, as no single component accounts for most of the variance.

Visual inspection of the PCA projections across the first four principal components (Figure 2) further corroborates our quantitative findings. In all pairwise combinations of the axes (from PC1–PC2 to PC3–PC4), the data points exhibit a dense and continuous elliptical distribution centered around the origin.

Notably, the **absence of distinct clusters or isolated "islands"** suggests that our dataset is free from dominant thematic biases or domain-specific groupings. The stable and broad distribution across these dimensions ensures that the 10,000 examples are semantically diverse and well-balanced. This visual evidence confirms that the benchmark presents a challenging, non-trivial semantic space in which models cannot rely on simple keyword-based clustering or topic-specific heuristics to achieve high performance.

To validate this, we performed a one-sample t-test against the null hypothesis that the top 20 components each contributed 0.05 (perfect uniformity). Although the p-value (1.59×10^{-12}) led to rejection of the strict null hypothesis, a comparison with a dataset of scientific papers (cumulative contribution ratio of 0.496) indicates that our dataset exhibits significantly less bias and greater diversity than the latter. This suggests that the dataset is well-suited as a general-purpose benchmark that is not overly reliant on specific knowledge.

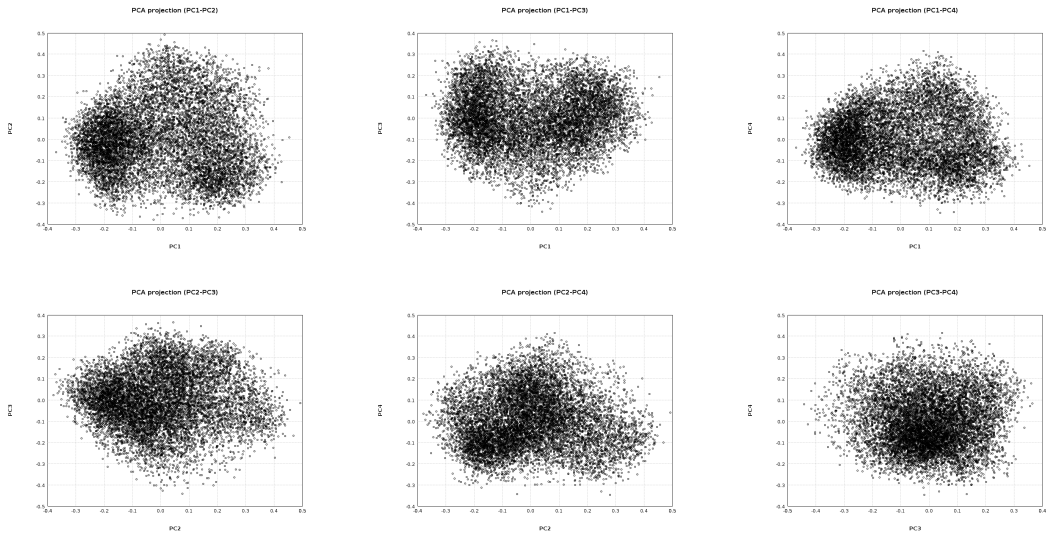

Figure 2: PCA projections of the benchmark embeddings across the first four principal components (PC1–PC4). The uniform, dense distribution without distinct clustering indicates a lack of thematic bias.

5 Experiments

5.1 Dataset Statistics

The proposed benchmark dataset comprises 10,000 items, balanced with 5,000 positive (causal) and 5,000 negative (non-causal) instances each. These were distributed across 20 distinct categories (10 positive and 10 negative), with 500 examples per category to ensure comprehensive coverage. The average text length per problem was approximately 5.2 sentences, which was significantly longer than the single-sentence examples typically found in SemEval-2007.

5.2 Evaluation Metrics

To evaluate the performance of our causal extraction system, we employ several standard classification metrics based on the counts of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Specifically, we report Accuracy, Sensitivity (Recall), Specificity, Precision, F1 Score, and the Matthews Correlation Coefficient (MCC). The mathematical definitions for each of these metrics are provided in **Appendix A.4**.

5.3 Experimental Design

We conducted a series of experiments to evaluate the performance of our proposed Multi-Agent System and the quality of the new benchmark dataset. We utilized the SemEval-2007 Task 4 dataset as a standard baseline and our newly created Categorized Benchmark Dataset for advanced evaluation.

Baseline Experiment (Experiment 0): We utilized `GPT-4.5-preview`(as the primary reasoning agent) and `GPT-4o-mini` (as the efficiency-focused critic) for our Multi-Agent baseline. Note that `GPT-4.5-preview` was exclusively employed during the dataset construction phase to generate the initial 10,000 examples, as described in Section 4.1.

Metric	Score (%)
Accuracy	89.66
Sensitivity (Recall)	91.67
Specificity	88.24
Precision	84.62
F1 Score	88.00
MCC	79.13

Table 3: Baseline Performance on SemEval-2007 Task 4

The accuracy of 89.66% is comparable to the reported human inter-annotator agreement (approximately 89%) for this task, indicating that our system has achieved human-level performance on standard, short-sentence causal extraction.

5.4 Ablation Study 1: Impact of Reasoning Process

To verify the contribution of the explicit "Reasoning Process" in the prompt (as shown in Listing 1), we compared the accuracy of the Causal Relationship Judge with and without this requirement. The results are summarized in Table 4.

To visualize these results, Figure 3 shows the

impact of reasoning on accuracy.

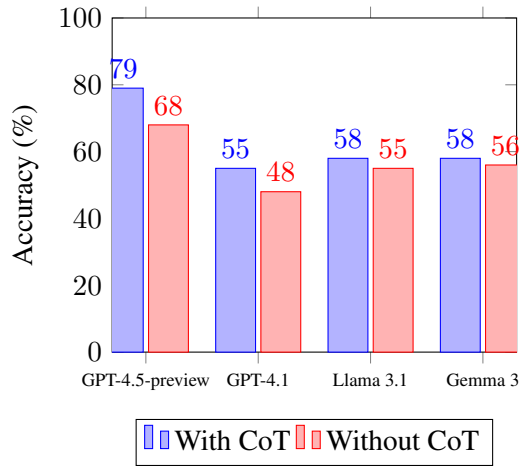

Figure 3: Impact of Reasoning Process on Model Accuracy.

Model	With CoT (%)	W/O CoT (%)
GPT-4.5-preview	**79**	68
GPT-4.1	55	48
Llama 3.1:70b	58	55
Gemma 3:27b	58	56

Table 4: Impact of CoT Reasoning on Accuracy.

As shown in Table 4, adding the reasoning process consistently improved the accuracy across all the tested models. Notably, even for `Gemma 3`, a smaller model, the performance gain indicates that CoT prompting is effective for causal tasks regardless of the model scale.

5.5 Ablation Study 2: Multi-Agent Configuration

We investigated the robustness of the multi-agent architecture by varying the models used for the Judge (front end) and the Critic (back end).

Experiment 2 (Small LLM as Judge): We fixed the Critic and Aggregator to `GPT-4o-mini` and tested smaller or different models (`GPT-4.1`, `Llama 3.1:70b`, `Gemma 3:27b`) as the Judge. Remarkably, the final system accuracy remained comparable to the baseline using `GPT-4.5-preview`. This suggests that a strong critic can effectively correct errors made by a weaker judge, thereby maintaining the system's overall robustness.

Experiment 3 (Small LLM as Critic): Conversely, we fixed the Judge to `Llama 3.1:70b` and used weaker models (`Gemma 3:27b` and `Phi-4:17b`) as the Critic The results shown in Table 5 reveal a significant drop in performance compared to using `GPT-4o-mini` as the critic.

Metric	Gemma 3 (27b)	Phi-4 (17b)
Accuracy	65.52%	79.31%
Sensitivity	83.33%	58.33%
Specificity	52.94%	94.12%
Precision	55.56%	87.50%
F1 Score	66.67%	70.00%
MCC	36.82%	57.80%

Table 5: Performance when using smaller models as the Critic (with Llama 3.1:70b as Judge). Comparing these to be effective.

This result (Experiment 3) underscores that the "Critic" role is computationally demanding and requires strong reasoning capabilities to evaluate the Judge's logic. A weak critic fails to filter out hallucinations effectively, degrading the final output.

5.6 Uniformity Evaluation via PCA

Table 6: Comparison of Cumulative Variance Proportions

Components	Scientific Paper	Our Benchmark Data
5	0.3299	0.1878
10	0.4910	0.2794
15	0.5899	0.3484
20	0.6629	0.3996
25	0.7199	0.4403
30	0.7668	0.4738
35	0.8063	0.5031
40	0.8392	0.5287
45	0.8667	0.5512
50	0.8909	0.5709

To quantitatively assess the diversity and domain-independence of our dataset, we performed Principal Component Analysis (PCA) on the text's semantic embeddings. Each paragraph was mapped into a high-dimensional vector space using the `text-embedding-3-small` model. This analysis allows us to measure **"semantic concentration"**—a metric of how much information is "packed" into a few recurring patterns.

Mathematical Formulation of Uniformity: To quantify this diversity, we analyze the contribution ratio of each principal component. Let λ_i be the i-th eigenvalue obtained from the covariance matrix of the text embeddings. The contribution ratio C_k of the k-th principal component is defined as:

$$C_k = \frac{\lambda_k}{\sum_{i=1}^{d} \lambda_i} \quad (1)$$

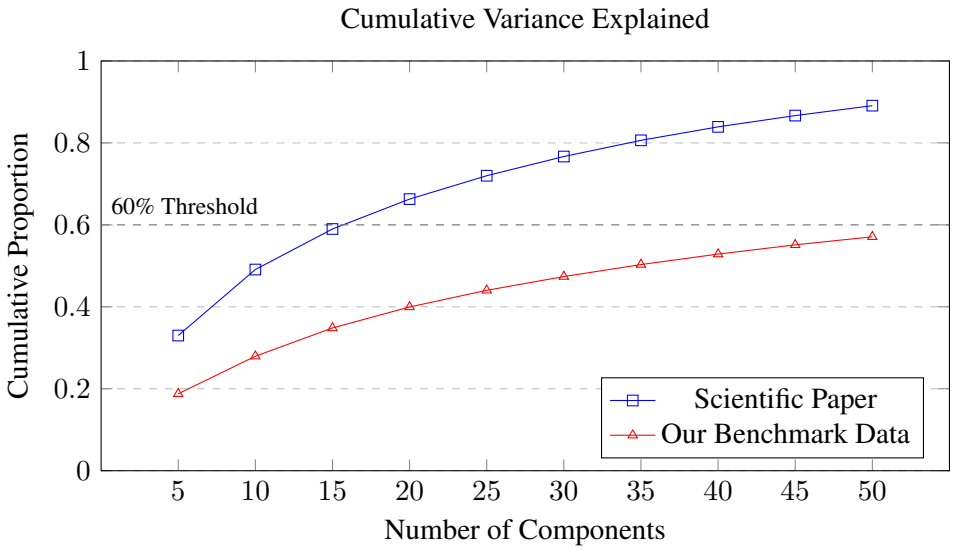

Figure 4: Visual representation of information retention across components.

where d is the total number of dimensions in the embedding space. The **cumulative contribution ratio** (which we refer to as the **"rate of capture"**) for the top K components is then expressed as:

$$R_K = \sum_{k=1}^{K} C_k \tag{2}$$

In our evaluation, a lower R_K for a fixed K indicates that the dataset's information is more widely distributed across the semantic space, signifying higher uniformity and less thematic bias.

As shown in Table 6 and Figure 4, our benchmark exhibits a significantly lower capture rate than the scientific paper baseline. This quantitative evidence, combined with the dense distribution in Figure 2, confirms that our dataset is **semantically diverse** and free from thematic shortcuts that models often exploit. Our results show that the first principal component (C_1) accounts for only 7.41% of the total variance, and the top ten components together (R_{10}) capture only 36.6%.

In contrast, while the scientific corpus achieves the 60% information retention threshold with approximately 16 components, our dataset requires significantly more dimensions to reach the same level. **This lower rate of capture confirms that our data is not dominated by a few obvious topics or sentence structures, but is instead composed of a vast array of diverse linguistic features**.

To validate this statistically, we performed a one-sample t-test against the null hypothesis H_0 that the top 20 components each contributed equally ($C_i = 0.05$), representing perfect uniformity. Although the p-value (1.59×10^{-12}) led to the rejection of the strict null hypothesis, the comparison with the scientific paper baseline (cumulative ratio of 0.496 at $K = 10$) demonstrates that our dataset exhibits significantly greater diversity. This ensures that the benchmark is robust and not overly reliant on specific domain knowledge.

5.7 Performance on the New Benchmark

Finally, we evaluated our best-performing multi-agent system (Experiment 0) on the proposed **Categorized Benchmark Dataset**. As summarized in Table 7, the system's accuracy significantly declined to **70.00%**, representing a **19.66 percentage point drop** compared to its performance on the SemEval-2007 baseline.

This substantial decrease confirms that the new dataset presents a significantly more challenging task. The **"deceptive" negative categories**—specifically designed to mimic plausible causal links—successfully misled the models, which indicates that current LLMs still struggle to distinguish between **actual causation and spurious correlations** when presented with long-form, linguistically complex texts.

While our system achieves state-of-the-art results on short, standard texts, these results reveal critical limitations in handling **logical fallacies** and **hidden confounders** in novel scenarios. This gap underscores the need for a more rigorous evaluation framework to advance models that extend beyond surface-level pattern recognition.

Metric / Feature	SemEval-2007	Proposed Benchmark
Accuracy	89.66%	**70.00%**
Context Length	Short sentences	**Long-context ($\geq$5 sentences)**
Key Challenge	Saturated/Simple	**Deceptive Correlations**
Difficulty Level	Human-level	**Substantial Reasoning Gap**

Table 7: Performance comparison between the standard SemEval-2007 Task 4 and our challenging Categorized Benchmark Dataset.

6 Discussion

The significant performance gap between the SemEval-2007 dataset (89.66%) and our proposed benchmark (70.00%) underscores the complexities of modern causal extraction tasks. While the high score on SemEval confirms that our Multi-Agent System is state-of-the-art for short, standard texts, the drop in performance on our dataset reveals critical limitations in the current LLMs' ability to handle "deceptive correlations" and long-context dependencies.

Specifically, the Multi-Agent architecture proved effective in mitigating simple errors. The "Critic" agent effectively filtered hallucinations in standard contexts. However, the system's struggle with "Hard" categories like *Confounding & Multiple Causes* underscores that current LLMs lack the **deep world knowledge** required to reject plausible but false causal chains in complex, long-form narratives.

Furthermore, the "Reasoning Process" ablation study highlights that simply asking for an answer is insufficient; forcing the model to articulate its logic significantly improves its accuracy. This suggests that future improvements may come not only from larger models but also from more structured "thinking" processes or integration with external knowledge bases to verify factual consistency.

7 Limitations

Despite its robustness, our framework has several limitations. First, system effectiveness **depends heavily on the Critic agent's reasoning capabilities**; using weaker models significantly reduces accuracy and fails to suppress hallucinations, necessitating **substantial computational resources** for high-precision tasks. Second, the system still **struggles with "Hard" categories** (e.g., *Confounding & Multiple Causes*) that require deep world knowledge and temporal analysis beyond linguistic patterns. Third, LLMs remain prone to acting as **"causal parrots,"** where internal biases and logical fallacies can override pure inference. Finally, as our approach relies on **prompt engineering**, future work should integrate **external knowledge bases** or structured thinking processes to improve factual verification.

8 Ethical Considerations

Our study involves several ethical considerations. First, the 19.66% performance drop on deceptive correlations highlights the risk of relying on LLMs for critical decisions in medical or financial domains. Second, the tendency of models to act as "causal parrots" poses a risk of reinforcing systemic biases present in pre-training data. Third, the high computational demands of the "Critic" agent raise concerns about energy consumption. Finally, our benchmark utilizes entirely synthetic data, ensuring that no personally identifiable information (PII) was compromised.

9 Conclusion

This study presents a comprehensive framework for advancing causal relationship extraction. We introduced a Multi-Agent System that achieves state-of-the-art results on standard benchmarks by mimicking the peer-review process. Furthermore, we addressed benchmark saturation by releasing a large-scale, categorized dataset of complex, long-context examples. The experimental results show a clear gap between the current model capabilities and the requirements of complex causal reasoning. Our system excels at standard tasks but faces difficulties with deceptive correlations in our new benchmark. This dataset will serve as a valuable resource for the community, driving the development of models that go beyond surface-level pattern recognition to achieve accurate causal understanding.

References

Abbas Akkasi and Mari-Francine Moens. 2021. Causal relationship extraction from biomedical text using deep neural models: A comprehensive survey. *Journal of Biomedical Informatics*, 119:103820.

Meilin Chen, Jian Tian, Liang Ma, Di Xie, Weijie Chen, and Jiang Zhu. 2025. Unbiased evaluation of large language models from a causal perspective. *Preprint*, arXiv:2502.06655.

Haoang Chi, He Li, Wenjing Yang, Feng Liu, Long Lan, Xiaoguang Ren, Tongliang Liu, and Bo Han. 2025. Unveiling causal reasoning in large language models: Reality or mirage? *Preprint*, arXiv:2506.21215.

Jin Du, Li Chen, Xun Xian, An Luo, Fangqiao Tian, Ganghua Wang, Charles Doss, Xiaotong Shen, and Jie Ding. 2025. Ice cream doesn't cause drowning: Benchmarking llms against statistical pitfalls in causal inference. *Preprint*, arXiv:2505.13770.

Juri Frohberg and Frank Binder. 2022. CRASS: A novel data set and benchmark to test counterfactual reasoning of large language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2126–2140. European Language Resources Association.

Fujitsu Limited. 2024. Fujitsu causal knowledge graph: Transform to data-driven decision-making based on logical reasoning. White paper, Fujitsu Limited.

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2007. SemEval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of SemEval-2007*, pages 13–18. ACL.

Utkarshani Jaimini and Amit Sheth. 2022. Causalkg: Causal knowledge graph explainability using interventional and counterfactual reasoning. *IEEE Internet Computing*, 26(1):43–50.

Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. 2024a. Can large language models infer causation from correlation? *arXiv preprint arXiv:2306.05836*.

Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. 2024b. Can large language models infer causation from correlation? *Preprint*, arXiv:2306.05836.

Nitish Joshi, Abulhair Saparov, Yixin Wang, and He He. 2024. LLMs are prone to fallacies in causal inference. In *Proceedings of EMNLP 2024*, pages 10553–10569.

Emre Kiciman, Robert Osazuwa Ness, Amit Sharma, and Chenhao Tan. 2024. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research (TMLR)*.

Kewei Lyu, Yu Tian, Yong Shang, Tianshu Zhou, Ziyue Yang, Qianghua Liu, Xi Yao, Ping Zhang, Jianghua Chen, and Jingsong Li. 2023. Causal knowledge graph construction and evaluation for clinical decision support of diabetic nephropathy. *Journal of Biomedical Informatics*, 139:104298.

Jie Yang, Soyeon Caren Han, and Josiah Poon. 2021. A survey on extraction of causal relations from natural language text. *Preprint*, arXiv:2101.06426.

A Appendix

A.1 Prompt Details

To ensure the reproducibility of our Multi-Agent system, we provide the full text of the prompts used in our experiments.

```
You are an expert in linguistic analysis specializing in causal inference. Your task is to determine whether a causal-effect relationship is implied between two specific entities within a given text.

### Task
$task

### Input
**Entity 1 ('{{E1}}'): The first entity to consider as a potential cause.
**Entity 2 ('{{E2}}'): The second entity to consider as a potential effect.
***Text ('{{TEXT}}'): The passage you need to analyze.

### Output Format
Please provide your answer in the following strict format:

Answer: ["Yes" or "No"]
Confidence: [a numerical score between 0.0 and 1.0]
Reasoning Process: [A concise, step-by-step breakdown of how you arrived at your conclusion.] This should outline the linguistic cues, logical deductions, or absence of evidence considered.]
Explanation of the reason: [A clear and detailed explanation summarizing why a causal relationship is (or isn't) implied, referencing specific parts of the text or general causal principles.]

### Examples for Clarification (if applicable)
* '$example1'
* '$example2'
```

Listing 1: Prompt for Causal Relationship Judge

```
You are an expert in linguistic analysis. Your task is to critically evaluate the proposed answer to a causal-effect relationship identification task.

Task Description:
$task

Text to Analyze:
'{{TEXT}}'

Proposed Answer:
'{{ANSWER}}'

Evaluation Question:
Given the 'Task Description' and the 'Text to Analyze', is the 'Proposed Answer' (including its stated causal determination and accompanying explanation) fully consistent, accurate, and logically sound?

Please provide your evaluation in the following format:

Answer: [Your 'Agree' to the proposed answer or 'Not agree' to the proposed answer response indicating if the proposed answer is correct]

Explanation: [Your detailed reasoning supporting your 'Agree' or 'Not agree' answer, referencing the text and the definition of a causal-effect relationship. Explain why the proposed answer is correct or incorrect.]
```

Listing 2: Prompt for Validity Critic

```
You are a critical evaluator tasked with reviewing and comparing a response and the review for the response to a given task. Your objective is to determine the correct answer to the original task, based on the provided text and the quality of each response.

**Task to Evaluate:** "$task"
**Text for Analysis:** "{{TEXT}}"
**Response to Evaluate:** "{{ANS1}}"
**Review of Response:** "{{ANS2}}"

Evaluation Question:
Based on the 'Task', 'Text for Analysis', 'Response to Evaluate', and 'Review of Response', what is the correct answer to the original 'Task'?

Output Format (in plain text):
Answer: ["Yes" or "No" in plain text without any Markdown syntax]
The reason: [Provide a comprehensive explanation detailing your examination of both responses. Clearly state which response (the 'Response to Evaluate') is correct or more accurate, or if it's incorrect. Justify your assessment by explaining why the response is accurate or inaccurate, referencing the 'Task' and 'Text for Analysis'. If the 'Review of Response' highlights valid points, incorporate them into your justification. If the 'Review of Response' is incorrect or misleading, explain why. Your explanation should lead directly to your final "Answer" for the 'Task' itself.]
```

Listing 3: Prompt for Aggregator

```
Sentence: "This, as well as kinetic data, supports the hypothesis of inhibition through altered membrane properties."

Answer: Yes. The sentence explicitly states that the hypothesis of inhibition is supported by evidence of altered membrane properties. The phrasing "inhibition through altered membrane properties" strongly implies that the altered membrane properties are the mechanism or cause of the inhibition. A clear directional relationship is presented, indicating a causal link.

Reasoning Process:
1. Analyze the Sentence Structure: The sentence is constructed to link "inhibition" and "altered membrane properties" through the phrase " support the hypothesis of inhibition through altered membrane properties." This structure suggests a directional relationship between the two, with inhibition leading to altered membrane properties.
2. Identify Keywords and Phrases: "Support": This word implies evidence backing the idea of a connection between the two entities. It suggests that the data confirms or strengthens the hypothesis of a causal link. "Kinetic data ": This phrase indicates that experimental measurements were taken... "Hypothesis of inhibition": This phrase explicitly proposes that inhibition is the causative factor.
3. Consider Alternative Interpretations: While the sentence strongly suggests a causal relationship, it's crucial to consider other possibilities: Correlation, not causation... Reverse causation...
4. Evaluate the Strength of the Evidence: The sentence states that both "this" and the kinetic data support the causal hypothesis. This convergence of multiple lines of evidence supports the conclusion that inhibition is responsible for the altered membrane properties.
5. Formulate the Conclusion: Based on the sentence structure, keywords, and the strength of the evidence presented, it is highly likely that the sentence implies a causal-effect relationship between inhibition and altered membrane properties.
```

Listing 4: Example of Reasoning Process in Case Memory

A.2 Dataset Generation Prompts

Listing 5 presents the specific instructions provided to the model to generate the positive causal examples for our benchmark.

```
Make a list of 25 paragraphs for benchmarking causal analysis.
Annotated words with e1 and e2 must be entities.
The entities are causally related but could ultimately be misinterpreted as unrelated.
This paragraph should subtly suggest a potential causal relationship without explicitly stating it, allowing the reader to draw their own (potentially erroneous) conclusions.
Avoid explicitly stating a clear timeline or causal relationship.
Instead, focus on creating a plausible yet deceptive correlation.
* Each paragraph should consist on more than five sentences.
* An entity is one word or a few words.
* The causal relationship should be "Preventative Causation".
```

Listing 5: Prompt for Generating Positive Examples (e.g., Preventative Causation)

Listing 6 shows the prompt used to create negative instances. These instructions explicitly guide the model to produce nuanced and "deceptive" scenarios that lack a true causal mechanism.

```
Make a list of 25 paragraphs for benchmarking causal analysis.
Annotated words with e1 and e2 must be entities.
The entities appear to be causally linked to a casual observer, but in reality, are independent. The paragraph should subtly suggest a potential causal relationship without explicitly stating it, allowing the reader to infer the false connection.
Avoid obvious or simplistic examples. Aim for a nuanced scenario that could plausibly mislead someone.
* Each paragraph should consist on more than five sentences.
* An entity is one word or a few words.
* The relationship between the entities should be "Descriptive Statement, Not Causal".
```

Listing 6: Prompt for Generating Negative Examples (e.g., Descriptive Statement, Not Causal)

A.3 Dataset Structure Details

Listing 7 illustrates the JSON structure of an example from our categorized benchmark dataset.

```
{
"e1": "heavy rainfall",
"e2": "power outage",
"text": "Residents noticed that after several days of heavy rainfall, there were widespread reports of a power outage across many neighborhoods...",
"answer": "true",
"category": "Intermediary/Chain Causation"
}
```

Listing 7: JSON Structure of the Generated Dataset

A.4 Mathematical Definitions of Evaluation Metrics

The metrics used to evaluate the causal extraction system are defined as follows based on the confusion matrix components (TP, TN, FP, FN):

- **Accuracy**: The proportion of total correct predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

- **Sensitivity (Recall)**: The ability of the model to correctly identify positive causal relationships.

$$Sensitivity = \frac{TP}{TP + FN} \quad (4)$$

- **Specificity**: The ability of the model to correctly identify negative instances.

$$Specificity = \frac{TN}{TN + FP} \quad (5)$$

- **Precision**: The proportion of identified causal relationships that are truly causal.

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

- **F1 Score**: The harmonic mean of Precision and Sensitivity.

$$F1Score = 2 \cdot \frac{Precision \cdot Sensitivity}{Precision + Sensitivity} \quad (7)$$

- **Matthews Correlation Coefficient (MCC)**: A robust measure for binary classification that accounts for all four quadrants of the confusion matrix.
  Let $S_1 = TP + FP$, $S_2 = TP + FN$, $S_3 = TN + FP$, and $S_4 = TN + FN$. The MCC is then defined as:

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{S_1 S_2 S_3 S_4}} \quad (8)$$