Team YTY at SemEval 2026 task 12: Option-Aware Retrieval and Cross-Encoder Reasoning Framework for Abductive Event Reasoning

Younghee Jeong, Yixin Zhao, Hoang Bao Trung Le
University of Tübingen
Department of Computational Linguistics
Germany
young-hee.jeong@student.uni-tuebingen.de yixin.zhao@student.uni-tuebingen.de
hoang-bao-trung.le@student.uni-tuebingen.de

Abstract

Abductive reasoning—the capacity to infer the most plausible explanation from incomplete or noisy observations—remains a significant hurdle for language models that often rely on simple associative patterns. In this paper, we present our framework for the SemEval 2026 Task 12. We propose an Option-Aware Retrieval and Cross-Encoder Reasoning Framework designed to bridge the gap between evidence acquisition and causal inference. Our architecture utilizes a dual-track retrieval strategy that gathers both global background and option-specific evidence, ensuring high recall of decisive clues. This is coupled with a 4-pass independent cross-encoder validation using DeBERTa-v3-large, which isolates individual hypotheses to prevent attention dispersion. Experimental results on the official dataset show that our system achieves a robust test score of 0.8358.

1 Introduction

Abductive reasoning, the process of inferring the most plausible explanation for a set of incomplete or noisy observations, is a cornerstone of human-level causal cognition. While language models have demonstrated remarkable performance in deductive and inductive tasks, they often falter in abductive contexts where relevant evidence is sparse or intermixed with irrelevant information. This limitation is particularly critical because real-world event understanding requires shifting from simple associative patterns to robust causal inference.

In this paper, we address this gap by proposing an Option-Aware Retrieval and Cross-Encoder Reasoning Framework. Despite the maturity of individual retrieval and reading comprehension models, their integration for noisy causal inference remains suboptimal. To overcome this, we introduce an architecture specifically configured for the complexities of abductive reasoning. Our approach integrates targeted, option-specific evidence gathering with a strictly isolated, 4-pass cross-encoder validation. By forcing independent hypothesis verification within a concentrated context, and coupling this with a score-based logical post-processing filter, our framework prevents the attention-dispersion and probabilistic contradictions common in multi-class machine reading comprehension (MRC) models. Consequently, this separated architecture enables language models to distinguish true causal links from spurious correlations, thereby significantly enhancing the reliability of abductive reasoning in complex real-world scenarios.

2 Related Work

Early work on abductive reasoning has focused mainly on common sense reasoning and employed statistical NLP methods such as Pointwise Mutual Information (PMI), based on datasets such as COPA (Choice of Plausible Alternatives) and extensions that emphasize causal plausibility (Roemmele et al., 2011; Gordon et al., 2012). With the advent of large language models (LLMs), the focus has increasingly shifted to the use of LLMs for causal inference tasks (Ma, 2025; Liu et al., 2025). The ART dataset, created by (Bhagavatula et al., 2020), introduced Abductive NLI (Natural Language Inference) as a discriminative choice task, pinpointing a persistent gap between model and human performance.

In practice, prompted generative LLMs are often adapted to closed-set reasoning tasks, by mapping each candidate to an independent prompt and selecting those with the highest conditional likelihood (Rahmani et al., 2025). However, prompting-based methods can be relatively unstable due to prompt-sensitivity, as small prompt changes may lead to large performance changes (Pecher et al., 2026). By contrast, recent work reports that structured

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1979–1987

discriminative models can achieve higher accuracy and efficiency than generative models in certain settings like closed-set classification, as they optimize explicit option-level decision boundaries without autoregressive decoding (Pang et al., 2026).

In parallel, to improve factuality and better process domain-specific queries, retrieval augmentation has been explored to ground LLM-based causal inference in external evidence (Lewis et al., 2021; Arslan et al., 2024). Iterative retrieval-based methods further interleave reasoning and evidence acquisition, reporting improved robustness in multi-hop and distractor-rich settings (Yao et al., 2023; Trivedi et al., 2023). Beyond general knowledge grounding, option-aware dense retrieval in multi-choice query answering improves evidence identification, yielding enhanced retrieval quality and better performance (Singh and Shrivastava, 2025).

While substantial work has contributed to the effectiveness of retrieval-augmented generation (RAG) pipelines (Uhm et al., 2025; Brown et al., 2025), where an external retriever fetches evidence to condition generation, few studies systematically integrate option-aware retrieval with a structured discriminative reasoner tailored for abductive reasoning. Our work bridges this gap by combining bi-encoder retrieval with a finetuned cross-encoder reasoner under isolated option-level decision objectives.

3 Methodology

Constructing causal reasoning system for unconstrained textual data requires navigating a fundamental trade-off between evidence recall and reasoning precision. To address this, we propose a two-stage *Option-Aware Retrieval and Cross-Encoder Reasoning Framework*. Our architecture is explicitly designed to decouple the search phase from the reasoning phase, allowing each component to specialize in its respective role.

As visually shown in Figure 1, our framework is composed of three interconnected modules. The left-most block illustrates the Input Phase, where the original query is decomposed into a global event and four option-specific branches. These branches independently feed into the Retriever Module using all-MiniLM-L12-v2, which utilizes deep self-attention distillation (Wang et al., 2020) and is implemented via the Sentence Transformers framework (Reimers and Gurevych, 2019). The module performs targeted searches to construct a

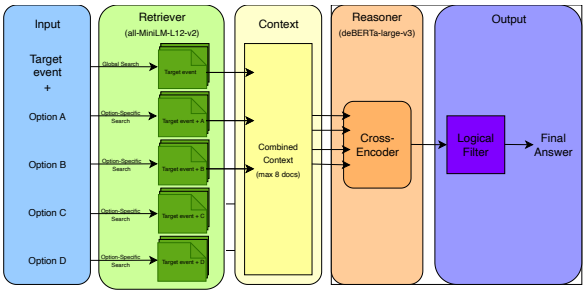

Figure 1: Overview of the proposed Option-Aware Retrieval and Cross-Encoder Reasoning Framework. The pipeline explicitly divides into three stages: (1) extracting dense evidence for the target event and each specific option (Retriever), (2) integrating the retrieved documents into a consolidated context to perform independent 4-pass hypothesis verification (Reasoner), and (3) applying a score-based heuristic to enforce mutually exclusive deterministic outputs (Logical Filter).

highly concentrated, de-duplicated Combined Context (capped at 8 documents). Moving to the right, the Reasoner Module, powered by DeBERTA-v3-large, introduced in (He et al., 2023), receives this context block. Crucially, the figure shows four distinct pathway entering the Cross-Encoder. This indicates our *4-Pass Inference* mechanism, where the model evaluates each candidate causal link sequentially rather than simultaneously. Finally, the four resulting probability scores converge into the Logical Filter block, which resolves any probabilistic contradictions and outputs a single, definitive causality prediction.

3.1 Option-Aware Hybrid Retrieval

If a retrieval system searches documents based solely on the Target Event, it may acquire broad background information but often fails to capture the decisive clues unique to each candidate answer. Based on the insight that a single perspective is insufficient for noisy abductive reasoning, we concluded that relevant evidence is frequently embedded within the specific interaction between the event and its potential causes.

Therefore, we configured our retriever to perform a dual-track search: first it retrieves k_{global} documents using only the Target Event; second, it retrieves an additional k_{option} documents for each Target Event + Option X pair. For the retrieval architecture, we adopted a Bi-Encoder to ensure efficient token-embedding-based search. Finally, the framework filters out redundant documents and concatenates the unique set into a single, comprehensive context.

3.2 Cross-Encoder Reasoning

To maximize precision and validate the candidate causal links identified during the retrieval stage, we employ a high-capacity Cross-Encoder architecture. While Bi-Encoders are efficient for large-scale semantic search, they compute representations for the query and document independently, failing to capture the deep, word-level interactions necessary for complex causal inference. To address this, we utilize DeBERTa-v3-large as our backbone reasoner, allowing the model to perform full-attention across both the hypothesis and the retrieved context simultaneously.

3.2.1 4-Pass Independent Inference

A common approach in MRC is to evaluate all multiple-choice options concurrently using a multi-class softmax layer. However, in noisy abductive reasoning tasks, exposing the model to all candidate options at once often leads to attention dispersion and vulnerability to distractor documents.

To mitigate this, we decompose the multiple-choice problem into four independent binary classification tasks. For each candidate option ($O_i \in \{A, B, C, D\}$), we construct a unified input sequence by concatenating the target event (T), the specific option (O_i), and the consolidated context (C) retrieved from Stage 1:

$$\text{Input}_\text{i} = \text{[CLS] T} \mid O_i \text{ [SEP] } C \text{ [SEP]}$$

As visually summarized in Figure 2, by isolating each hypothesis, the Cross-Encoder concentrates its entire self-attention mechanism on verifying a single causal narrative at a time. The model outputs a raw logit score (S_i) representing the causal validity of O_i given C. This raw logit score is then transformed into a probability(P_i) using a sigmoid activation function at output layer: $P_i = sigmoid(S_i)$. This probability P_i is subsequently used for all thresholding operations. This binary classification format is significantly more robust against the interference of spurious correlations found in distractor options.

3.2.2 Logical Consistency Filter

A fundamental challenge in applying independent binary classifiers to a mutually exclusive task is the emergence of probabilistic contradictions. For instance, the model may confidently assign high probabilities to both a specific causal option (e.g., Option A) and the "None of the others" option(e.g., Option D), resulting in a logical paradox impossible in the ground truth.

To enforce logical consistency without retraining the neural architecture, we implement a *Highest-Score-Wins* heuristic as a definitive post-processing layer. When a logical conflict is detected among the predicted options (i.e., multiple options including "None" exceed the classification threshold), we bypass the binary thresholding and compare the post-sigmoid probabilities directly:

$$\text{Prediction} = \arg\max_{\text{O}_\text{i} \in \text{set}} (\text{P}_\text{i})$$

This structural logic gate forces the system to commit exclusively to its highest-confidence prediction. By explicitly resolving these neural contradictions, this filter acts as a robust fail-safe, significantly improving the exact-match accuracy of the pipeline and ensuring that the final output strictly adheres to the mutually exclusive constraints of the task.

4 Experimental Setup

4.1 Dataset & Evaluation Metric

We follow the official data split provided by the shared task without using external data. Baseline methods are evaluated on the sample data, as the gold labels for the test set were unavailable before the closure of submission. Models are trained on the training set and finetuned on the development set. Final results are reported on the test set (612 instances) via CodaBench. Each instance consists of a target event and four options as candidate causal hypotheses. Some instances contain a candidate option of the form “None of the others are correct causes” option. Additionally, semantically duplicate options may occur with different option keys. Multiple choices may be correct.

The evaluation logic follows the official metric: an exact match between predicted and gold answer sets receives a score of 1.0; a non-empty subset receives 0.5; all other predictions receive 0. The final score is the mean over all instances.

Furthermore, we assess the system’s performance as a ranking task utilizing post-sigmoid probability score to order potential candidates. To quantify effectiveness, we employ standard information retrieval metrics, specifically Mean Reciprocal Rank (MRR) and Precision@1 (P@1). These results are then benchmarked against other methods, such as cosine similarity, to analyze the performance-to-complexity trade-offs.

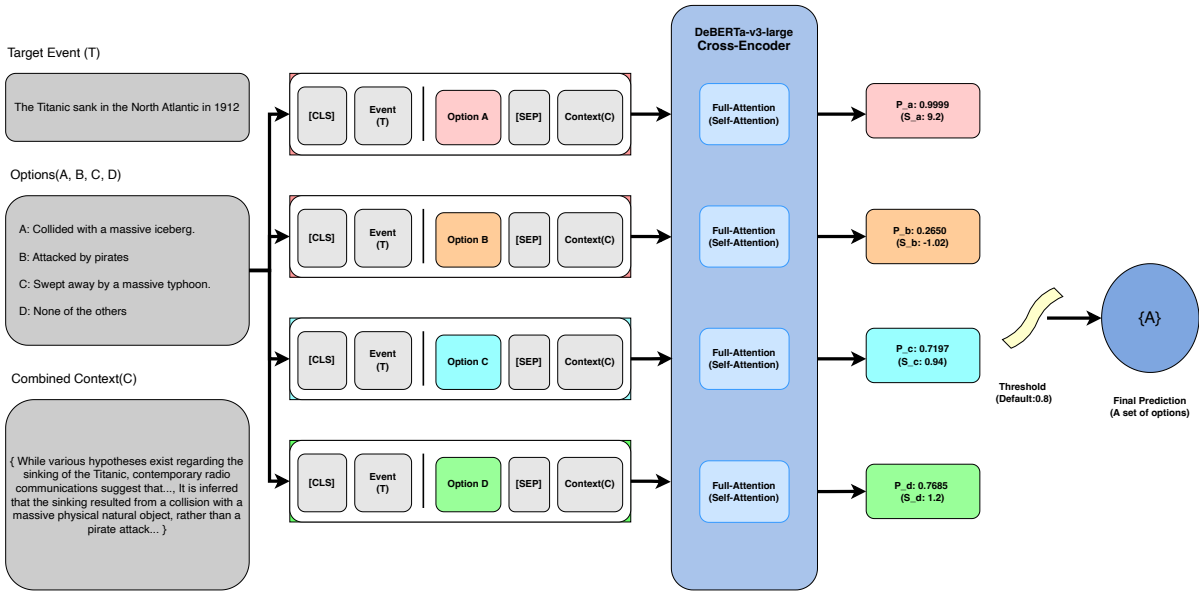

Figure 2: Illustration of the 4-Pass Independent Inference mechanism. As depicted, the single Target Event is paired individually with Option A, B, C, and D. Each pair passes through the DeBERTa Cross-Encoder alongside the Combined Context to produce four distinct probability scores, culminating in a final truth value via the logical filter threshold

4.2 Main System

4.2.1 Retrieval

We retrieve evidence using an option-aware, dual-track retrieval strategy. Queries are constructed in two forms: (1) the Target Event alone (global query), and (2) the concatenation of the Target Event with each Option X pair (option-aware query). Document and query embeddings are computed using the pre-trained bi-encoder all-MiniLM-L12-v2, implemented in SentenceTransformers (Reimers and Gurevych, 2019). Each document is truncated and combined with its title to fit within the bi-encoder's input limit. Cosine similarity is used for ranking. We retrieve the top two documents for the global query and the top document for each option-aware query. Duplicate documents are removed. The retrieved documents are then merged and truncated to no more than six documents per instance in typical cases, with a hard upper bound of eight to satisfy input-length constraints.

4.2.2 Training

The reasoner is a cross-encoder initialized from microsoft/deberta-v3-large. The model is trained as a binary classifier over the 4-pass inference mechanism. A candidate receives label 1 if it appears in the gold causal answer set and 0 otherwise. Each training example consists of the option-aware query and the retrieved evidence context. Training is performed for 3 epochs with batch size 4 and learning rate 5×10^{-6}, using AdamW optimization with 300 warm-up steps. Mixed-precision training is enabled. Model selection is based on the development set using option-level binary accuracy with a fixed threshold of 0.5. The best-performing checkpoint is retained.

4.2.3 Inference

At inference time, each option is evaluated independently. The raw logit scores for each option are transformed into probabilities via a sigmoid function. Options with probabilities ≥ 0.8 are initially selected as plausible causes. If no option's probability reaches this threshold, the option with the highest probability is selected. If all options' probabilities exceed the threshold, the option with the lowest probability is removed [1] to refine the selection.

Furthermore, unlike the substantive causal options, the "None of the others" option does not correspond to a concrete causal hypothesis and is therefore less naturally supported by retrieval evidence. In our current system, we nevertheless score it within the same pipeline for architectural consistency. However, the logical consistency filter, de-

[1] This design choice is empirically motivated by the label distribution of the training and development sets, in which no instance has all four options as correct causes.

scribed in Section 3.2.2, allows it to function partly as a fallback when the evidence for specific causal candidates is weak or contradictory: the "None of the others" option is selected only if it receives the highest probability among all four options, thereby resolving probabilistic contradictions. Additionally, the "None of the others" option can also serve as a warning flag in cases where the model assigns a high score to a non-substantive candidate.

5 Supplementary Experiments

Beyond the primary proposed system, we evaluate several baseline architectures to benchmark performance and analyze the trade-off between computational cost and retrieval accuracy. Moreover, to boost the performance of the model, we explored an alternative strategy and an ensemble strategy with LLM-as-judge for adjudication.

5.1 Retrieval Methods

5.1.1 Semantic Vector Search

As a baseline for topical relevance, we implemented a dense retrieval approach using a pre-trained text embedding model. This method calculates the Cosine similarity between document and query embeddings. The underlying hypothesis is that events linked by a causal relationship often share semantic overlap or exist within similar latent topical spaces.

$$\text{score}(q, d) = \frac{\mathbf{e}_q \cdot \mathbf{e}_d}{\|\mathbf{e}_q\| \|\mathbf{e}_d\|} \quad (1)$$

5.1.2 LLM-Augmented Vector Search

Recent advancements in LLMs have demonstrated significant proficiency in abductive reasoning tasks. Following the approach suggested by Zandie et al. (2023), we leverage the LLM's internal world knowledge to bridge the gap between an effect and its potential cause during the retrieval phase.

In this configuration, we prepend a generative step to the retrieval pipeline using the following prompt:

> *"Given the event: '[target event]', what is a plausible cause? Answer concisely in one sentence."*

The resulting generated hypothesis is then embedded and used as the query vector for Cosine similarity search. This "Query Expansion via Reasoning" aims to align the search vector more closely with the semantic space of the target causal documents.

5.1.3 Chunk-level Search

To explore whether an alternative strategy could improve system performance, we replaced the bi-encoder retriever with a chunk-level retriever. By applying semantic chunking with dynamic splits based on cosine similarity, we segmented the context documents into variable-length chunks. We then employed a combination of BM25 for sparse retrieval and BGE-M3 for dense retrieval, whose results were merged via Reciprocal Rank Fusion (RRF) to produce a top-30 candidate pool. A cross-encoder reranker (bge-reranker-v2-m3) then selected top-10 chunks, yielding approximately 300 tokens per option, also a potential causal hypothesis, as input to the DeBERTa reasoner.

5.2 LLM-as-judge Ensemble

After implementing the chunk-level search with the reasoner (deBERTa-v3-large), complementary errors were observed from a comparison with our main system. Following that, we designed a disagreement-driven ensemble strategy. For instances received identical predictions from the main model and the chunk-level-search-based model, we retained the results directly. For samples where the two models disagreed, we submitted both predictions along with the retrieved evidence to GPT-o3 as an LLM-as-judge for adjudication. The judge was instructed to select all plausible causes based on the evidence, with explicit constraints on the consistent treatment of duplicate options and the conditional use of the "none of the others are causes" option (full prompt provided in Appendix A), aligning with the logical filter design.

6 Results and Analysis

6.1 Overview

We mainly evaluate our system performance using the official metric. Table 1 shows this result for our training set, the validation (dev) set, as well as the released test set.

Table 1: System Performance: Average Scores

Dataset Split	Average Score
Train	0.9164
Dev	0.9750
Test	0.8358

The system achieves a high level of accuracy across all splits, peaking at an average score of

0.9750 on the development set. While the test performance remains robust (0.8358), there is a noticeable delta between the development and test phases that suggests the model may be highly sensitive to the specific distribution of the development set. Which is understandable considering the topic distribution shift. Specifically, the development set shares the same topics as the training set, whereas the test set introduces entirely new topics. The performance drop exhibited on out-of-distribution data indicates limitations in cross-topic generalization.

6.2 Comparison

Table 2: Ranking Performance in comparison with our supplementary experiments. Results shown for test set only

Method	MRR	Precision@1
Semantic Vector Search	0.6844	0.4534
LLM-Augmented Vector Search	0.7327	0.5285
Chunk-level Search + Finetuned deBERTa	0.8764	0.7958
Doc-level Search + Fine-tuned deBERTa	**0.9039**	**0.8420**
LLM-as-judge Ensemble	0.9623	0.9412

As described in table 2, the transition from traditional vector-based methods to a transformer-based architecture results in a substantial performance gain. Specifically, deBERTa outperforms Augmented Cosine Similarity by approximately 17% in MRR and 31% in Precision@1. This indicates that while cosine similarity provides a respectable baseline, it lacks the semantic depth required to accurately identify the top-ranked candidate in more complex scenarios.

It should be noted that the gap is much wider for Precision@1 (0.5285 vs. 0.8420) than for MRR (0.7327 vs. 0.9039). This implies that while cosine similarity can often "get close" (ranking the correct answer in the top 2 to 4), it struggles to consistently place the correct answer at the very top. In contrast, deBERTa demonstrates exceptional precision, making it far more suitable for systems where only the top result is shown to the user.

When replacing the doc-level, option-aware hybrid retrieval strategy (our main architecture) with chunk-level search, the performance exhibits a relatively uniform decrease across both Precision@1 (0.8764 vs. 0.9039) and MRR (0.7958 vs. 0.8420), receiving a score of 0.7132. This could be attributed to a loss of semantic coherence caused by chunking (Merola and Singh, 2025). or DeBERTa's disentangled attention mechanism is particularly sensitive to input coherence (He et al., 2020). Its fine-grained relational reasoning could be undermined when the input is fragmented.

The result of LLM-as-judge ensemble system, by capturing the complementary aspects of the two strategies, receives higher performance across the two metrics (0.9623 vs. 0.9039 for MRR and 0.9412 vs. 0.8429 for Precision@1) with a final score of 0.9020. However, this approach incurs significant workload of implementation overhead and relies on the availability of two independent systems, which makes it less suitable as a standalone pipeline.

6.3 Error Analysis

To understand the limitations of the proposed model, we conducted a qualitative analysis of the instances where the model assigned high probability scores (> 0.90) to non-golden options, shown in table 3. We identified the following primary failure modes:

- **Lexical Overlap Bias:** The model frequently over-relies on keyword matching. For instance, in question `q-2494`, the presence of the verb "launched" in both the event and the option triggered a high score, despite the options describing different stages of the mission (lunar vs. terrestrial launch).
- **Narrative Conflation:** The model struggles with temporal and causal boundaries within the same news cycle. In the context of the Syrian conflict (e.g., `q-2610`), it conflated the retreat of the army with specific rebel offensives, failing to isolate the unique factual relationship required for the "golden" label.
- **Entity Distraction:** High-profile entities (e.g., *Hezbollah* or *Chang'e 5*) act as anchors. The model often assigns high scores to any option that contains these entities, regardless of the specific action or outcome described in the premise.

7 Conclusion

In this work, we introduced a two-stage retrieval and reasoning framework tailored for the nuances of abductive event reasoning. We demonstrated that option-aware dense retrieval effectively captures the specific interactions between events and their potential causes. While our system achieved high precision and strong benchmarking results on

the SemEval 2026 test set, our error analysis highlights a persistent "Lexical Overlap Bias" and a tendency for the model to be distracted by high-profile entities. These findings suggest that while structured discriminative models offer both accuracy over simple vector comparisons and flexibility over generative approaches, future research should focus on enhancing cross-topic generalization and developing mechanisms to better handle temporal and causal boundaries in complex news cycles.

References

Muhammad Arslan, Hussam Ghanem, Saba Munawar, and Christophe Cruz. 2024. A survey on rag with llms. *Procedia Computer Science*, 246:3781–3790. 28th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2024).

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. *Preprint*, arXiv:1908.05739.

Andrew Brown, Muhammad Roman, and Barry Devereux. 2025. A systematic literature review of retrieval-augmented generation: Techniques, metrics, and challenges. *Big Data and Cognitive Computing*, 9:320.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention.

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.

Xiaoyu Liu, Paiheng Xu, Junda Wu, Jiaxin Yuan, Yifan Yang, Yuhang Zhou, Fuxiao Liu, Tianrui Guan, Haoliang Wang, Tong Yu, Julian McAuley, Wei Ai, and Furong Huang. 2025. Large language models and causal inference in collaboration: A survey. *Preprint*, arXiv:2403.09606.

Jing Ma. 2025. Causal inference with large language model: A survey. *Preprint*, arXiv:2409.09822.

Carlo Merola and Jaspinder Singh. 2025. Reconstructing context: Evaluating advanced chunking strategies for retrieval-augmented generation.

Zhanzhong Pang, Dibyadip Chatterjee, Fadime Sener, and Angela Yao. 2026. On discriminative vs. generative classifiers: Rethinking MLLMs for action understanding. In *The Fourteenth International Conference on Learning Representations*.

Branislav Pecher, Michal Spiegel, Robert Belanec, and Jan Cegin. 2026. Revisiting prompt sensitivity in large language models for text classification: The role of prompt underspecification. *Preprint*, arXiv:2602.04297.

Hossein A. Rahmani, Satyapriya Krishna, Xi Wang, Mohammadmehdi Naghiaei, and Emine Yilmaz. 2025. Self-correcting large language models: Generation vs. multiple choice. *Preprint*, arXiv:2511.09381.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Bejan, and Andrew Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium - Technical Report*.

Manish Singh and Manish Shrivastava. 2025. Options-aware dense retrieval for multiple-choice query answering. *Preprint*, arXiv:2501.16111.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037. Association for Computational Linguistics.

Miyoung Uhm, Jaehee Kim, Seungjun Ahn, Hoyoung Jeong, and Hongjo Kim. 2025. Effectiveness of retrieval augmented generation-based large language models for generating construction safety information. *Automation in Construction*, 170:105926.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Preprint*, arXiv:2002.10957.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

Rohola Zandie, Danny Brahman, and Mohammad Mahoor. 2023. COGEN: Abductive commonsense language generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 295–302, Toronto, Canada. Association for Computational Linguistics.

A LLM-as-Judge Prompt

The following prompt was used for GPT-o3 in the LLM-as-Judge ensemble experiment described in 5.2.

LLM-as-Judge Prompt

```
You are an expert in causal event reasoning.
Your task is to determine which of the
following options are plausible causes
of the target event, given the retrieved
evidence and two model predictions.

Target Event: {target_event}

Retrieved Evidence:
{evidence}

Options:
A. {option_a}
B. {option_b}
C. {option_c}
D. {option_d}

Model 1 Prediction: {model_1_prediction}
Model 2 Prediction: {model_2_prediction}

Instructions:

  1. Based on the retrieved evidence and
     your own reasoning, determine which
     options are plausible causes of the
     target event. Multiple options may be
     selected.

  2. If two or more options have identical
     content, treat them consistently —
     either all selected or all rejected.

  3. Select "None of the others" only if you
     are confident that none of options A,
     B, and C are plausible causes. Do not
     select it as a default or fallback.

  4. Consider both model predictions as
     references, but make your final
     judgment based on the evidence and
     causal plausibility, not by simply
     deferring to either model.

Respond strictly in the following format:
[option] A, C
[reasoning] {your reasoning here}
```

B Error Analysis

Table 3: Examples of fail predictions

q-2836	On Dec 5, the **Chang'e 5** ascender docked with the return spacecraft.	
Status	**Option Text**	**Probability Score**
Non-Golden	Chang'e 5 separated into the orbiter-reentry capsule combination and the lander-ascender combination.	**0.9998**
Non-Golden	**Chang'e 5** launched on November 23 atop a Long March 5 rocket.	**0.9993**
Golden	The **Chang'e 5** lander gathered rocks and soil from the Moon.	0.0001
Non-Golden	Chang'e-5 raised a Chinese flag on the Moon.	**0.9993**

q-2494	**Chang'e 5** launched from the lunar surface Dec. 3.	
Status	**Option Text**	**Probability Score**
Non-Golden	**Chang'e 5** launched on November 23 atop a Long March 5 rocket.	**0.9997**
Non-Golden	**Chang'e 5** mission entered lunar orbit on November 28.	0.0002
Golden	The **Chang'e 5** lander gathered rocks and soil from the Moon.	0.0001
Non-Golden	Chang'e-5 raised a Chinese flag on the Moon.	0.0001

q-2610	Syrian army notified officers that Assad's rule had ended.	
Status	**Option Text**	**Probability Score**
Non-Golden	Rebels take control of Aleppo city on November 30 after an offensive that kills dozens of government soldiers, causing the Syrian army to retreat.	**0.9998**
Golden	Former Syrian President Bashar al-Assad flees to Russia with his family.	0.0001
Non-Golden	Insurgents entered Saydnaya military prison and freed prisoners.	0.0001
Golden	Former Syrian President Bashar al-Assad flees to Russia with his family.	0.0001

q-2490	Lebanon, Iraq, and Syria declared three days of mourning after the airstrike.	
Status	**Option Text**	**Probability Score**
Non-Golden	Israeli airstrikes on Saturday targeted areas near Beirut and **Hezbollah** weapons caches, resulting in at least 33 killed and 195 wounded, according to Lebanon's health ministry.	**0.9998**
Golden	The airstrike killed **Hezbollah** Secretary-General Hassan Nasrallah, commander Ali Karaki, additional **Hezbollah** commanders, and IRGC Brigadier General Abbas Nilforooshan.	**0.9998**
Non-Golden	Residents fled Dahiyeh after intense Israeli strikes, seeking shelter elsewhere in Beirut.	0.0001
Non-Golden	Prime Minister Benjamin Netanyahu cut short his U.S. visit and ended a UN briefing after being informed of the attack.	0.0001