sutta at SemEval-2026 Task 12: A Multi-Perspective Retrieve-Verify-Aggregate Framework for Abductive Event Reasoning

Junliu Zou, Liang Yang*, Jingjie Zeng
School of Computer Science and Technology
Dalian University of Technology, China
1539606967@mail.dlut.edu.cn, liang@dlut.edu.cn

Abstract

We present our system for SemEval-2026 Task 12: Abductive Event Reasoning (AER). The task asks models to identify the direct causes of real-world events from multiple-choice options using retrieved documents. Rather than fine-tuning on the training data, we built a zero-shot "Retrieve-Verify-Aggregate" pipeline around Qwen3-8B. We first isolate relevant evidence using BM25 and cross-encoder reranking. To evaluate causal links, we prompt the model with several distinct "personas" and aggregate their independent decisions through majority voting. Our system scored 0.7614 on the official test set. This performance suggests that strict retrieval combined with diverse reasoning prompts can help compact open-source models ignore irrelevant context and perform complex causal inference, entirely without task-specific training.

1 Introduction

Abductive Event Reasoning (AER) requires models to identify the underlying causes of real-world events from document collections(Cao et al., 2026). Unlike standard information extraction, AER involves reasoning under uncertainty to bridge observation and causation, which is critical for applications like misinformation detection. Although Large Language Models (LLMs) perform well on general reasoning tasks(DeepSeek-AI et al., 2025; Yang et al., 2025; Team et al., 2026), they struggle in abductive scenarios. Standard approaches often suffer from context distraction(Liu et al., 2023). When processing retrieved documents, models frequently attend to plausible but non-causal distractors. The inability to separate actual causes from mere correlations leads to causal hallucinations and incorrect logical deductions(Wang et al., 2024; Li et al., 2025).

To address this limitation, we propose a training-free "Retrieve-Verify-Aggregate" framework that replaces single-pass generation with multi-perspective verification. Rather than relying on large proprietary models or task-specific fine-tuning, both of which can overfit to narrow causal patterns, we introduce a Multi-Perspective Reasoning Ensemble mechanism using Qwen3-8B. By combining BM25 retrieval and cross-encoder reranking with this ensemble, we decompose the abductive process into independent hypothesis-testing streams. This mechanism acts as a logical regularizer. By simulating distinct skeptical personas, the model cross-verifies its own outputs and discards false causal links induced by noisy context before final aggregation via majority voting.

We evaluate our approach on the SemEval-2026 Task 12 dataset. Comparisons with standard retrieval-augmented generation baselines show the effectiveness of our framework. Our analysis yields two key findings. First, increasing retrieved context degrades model focus, which supports the need for precise passage-level retrieval. Second, while complex standard prompts struggle to suppress causal hallucinations, our multi-persona mechanism effectively regularizes the deduction process. As a result, our lightweight ensemble achieves highly competitive performance and demonstrates the viability of zero-shot causal reasoning in compact models.

In summary, our main contributions are threefold. We propose a parameter-efficient and training-free "Retrieve-Verify-Aggregate" framework that addresses the context-distraction issue in abductive event reasoning without task-specific fine-tuning. We also introduce a Multi-Perspective Reasoning Ensemble mechanism based on Qwen3-8B, showing that modeling diverse skeptical personas serves as a crucial regularizer against context-induced hallucinations. Finally, we validate our approach on the SemEval-2026 Task 12 dataset, achieving a competitive average score of 0.7614 and establishing a robust open-source baseline for lightweight causal inference.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2564–2574

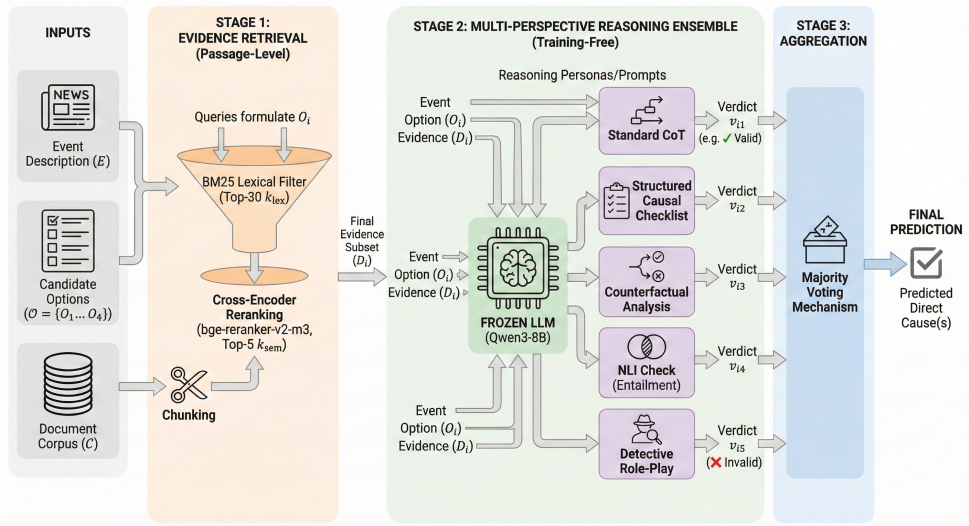

Figure 1: Overall architecture of our Retrieve-Verify-Aggregate framework. The system first retrieves relevant evidence using BM25 and a cross-encoder, then verifies causal links using a multi-perspective reasoning ensemble, and finally aggregates the results via majority voting.

2 Background

2.1 Abductive Event Inference

Computational abduction identifies plausible antecedents for observed consequents, a process formalized by the Abductive Natural Language Inference (aNLI) framework(Bhagavatula et al., 2019). While early benchmarks treated this as narrative classification, the contemporary AER task grounds causal deduction in unstructured, real-world corpora(Bhagavatula et al., 2019; Sap; Bosselut et al., 2019). This paradigm enforces "strong abduction", requiring systems to isolate true causal signals from deliberate semantic distractors(Nie et al., 2020).

Autoregressive language models inherently struggle under these constraints(Bubeck et al., 2023). Because they optimize for statistical sequence likelihood rather than logical entailment(Dziri et al., 2023), they frequently generate causal hallucinations to artificially bridge non-causal observations(Ji et al., 2023). Additionally, processing extended retrieved contexts triggers the "Lost in the Middle" phenomenon(Liu et al., 2023), where attention mechanisms degrade and prioritize plausible but irrelevant noise(Tay et al., 2022). These architectural pathologies demonstrate that single-pass generative pipelines are inadequate for AER, necessitating a structural decoupling of evidence retrieval and multi-perspective causal verification.

2.2 Retrieval-Augmented Generation

To ground language models in factual reality, Retrieval-Augmented Generation (RAG) paradigms are widely adopted(Lewis et al., 2020; Borgeaud et al., 2022). Standard RAG frameworks typically employ bi-encoder architectures for retrieval efficiency(Hofstätter et al., 2021). However, bi-encoders compress documents into single vectors without early query interaction, missing the granular semantic nuances required for strict causal verification(Yates et al., 2021).

Furthermore, expanding the context window to ingest more retrieved documents exacerbates contextual degradation by triggering the "Lost in the Middle" phenomenon, where attention mechanisms are derailed by peripheral semantic noise. To optimize the signal-to-noise ratio, recent architectures enforce strict passage-level filtration. Two-stage pipelines integrate probabilistic lexical filtering, such as BM25(Robertson and Zaragoza, 2009; Robertson and Walker, 1994), with joint-attention cross-encoder reranking. This structural constraint aggressively distills the context and isolates the reasoning engine from corpus-induced distraction.

2.3 Multi-Perspective Ensemble

To mitigate the fragility of single-pass generative reasoning, recent literature explores inference-time compute scaling through methods like Self-Consistency and multi-agent debate(Wang et al., 2023a; Du et al., 2024; Shinn et al., 2023). However, in strong abductive scenarios, homogeneous

sampling often amplifies the model's inherent confirmation bias toward plausible semantic distractors rather than correcting it(Yang et al., 2024).

Framing language model execution through explicit role-play alters initial token probabilities and activates structurally diverse reasoning pathways(Park et al., 2023). By simulating an artificial society of deliberative personas, systems achieve epistemic regularization(Wang et al., 2023b; Huang et al., 2023). Specialized personas evaluate identical evidence from distinct vantage points—ranging from standard narrative deduction to strict counterfactual skepticism. Structurally forcing the model to evaluate hypotheses through these skeptical personas acts as a stringent regularizer against causal hallucinations. Aggregating these divergent trajectories via discrete consensus mechanisms, such as majority voting, balances precision and recall while mathematically limiting individual prompt errors, entirely without task-specific fine-tuning.

3 System Overview

We formulate the AER task as a retrieve-then-verify pipeline, as illustrated in Figure 1. Formally, given an event E, a set of candidate causal options $\mathcal{O} = \{O_1, \ldots, O_4\}$, and a corpus of retrieved documents $\mathcal{C}$, our system first extracts the most relevant evidence subset $D_i \subset \mathcal{C}$ for each option O_i. Subsequently, it employs an ensemble of reasoning prompts to independently verify whether O_i constitutes a valid cause of E conditioned on D_i.

3.1 Evidence Retrieval

The provided document corpus $\mathcal{C}$ often exhibits significant length variations and contains a mixture of relevant signals and irrelevant background noise. To effectively filter semantic distractors and accommodate the context window constraints of the LLM, we implement a two-stage retrieval-then-rerank pipeline operating at the passage level.

Document Chunking. Before indexing, we segment each document into smaller, semantically coherent passages to enhance retrieval granularity. We split text by sentence delimiters and group sentences until the chunk hits 1,000 characters ($L_{\max}$). This heuristic ensures the preservation of semantic context within sentence boundaries.

Lexical Filtering. To ensure high recall of exact entity and event mentions, we first apply lexical filtering using BM25. For each event-option pair (E, O_i), we concatenate them into a single query $q_i = [E; O_i]$ and search against all chunks in the corpus $\mathcal{C}$. We retain the top $k_{\text{lex}} = 30$ chunks using standard parameters ($k_1 = 1.5$, $b = 0.75$).

Semantic Reranking. To further refine the candidate pool and improve the signal-to-noise ratio, we rerank the BM25 outputs using the bge-reranker-v2-m3 cross-encoder(Chen et al., 2024). By processing the query and document jointly, the cross-encoder captures fine-grained semantic interactions that lexical matching misses. For each candidate chunk c, we compute a relevance score $s(c) = \text{CrossEncoder}(q_i, c)$ and retain only the top $k_{\text{sem}} = 5$ chunks. This aggressively distilled set forms the final evidence D_i provided to the reasoning engine.

3.2 Multi-Perspective Reasoning Ensemble

Abductive reasoning necessitates evaluating hypotheses from multiple cognitive angles, as a single Chain-of-Thought (CoT) trajectory is highly susceptible to confirmation bias. To mitigate this, we introduce a Multi-Perspective Reasoning Ensemble that queries a single LLM using five distinct prompting strategies. Each strategy embodies a specific "reasoning persona": Standard CoT, Detailed CoT, Counterfactual Analysis, Natural Language Inference (NLI), and Detective Role-Play. By decomposing the reasoning process into these distinct logical perspectives, the model effectively self-verifies its logic and reduces its propensity to accept plausible but unsupported semantic distractors. Detailed descriptions of each persona's mechanism and their exact prompt templates are provided in Appendix B.

Each prompt P_j ($j \in \{1, \ldots, 5\}$) outputs a binary verdict $v_{ij} \in \{0, 1\}$ for option O_i, where 1 means "Valid Cause."

3.3 Vote Aggregation

The final prediction for each option O_i is determined via majority voting across the five prompts. We count the valid (V_{valid}) and invalid (V_{invalid}) verdicts. An option is initially deemed a candidate cause if it secures a strict majority of valid votes:

$$\text{IsCandidate}(O_i) = \mathbb{I}\left(V_{\text{valid}}(O_i) > V_{\text{invalid}}(O_i)\right) \quad (1)$$

We select all options where $\text{IsCandidate}(O_i) = \text{True}$ as predicted causes. If no option gets a strict majority, the system falls back: it either chooses the dataset's "None" option (if available)

Method	Full Match	Partial Match	Incorrect	Average Score
System (Ensemble-5)	**73.86%**	**4.58%**	**21.57%**	**0.7614**
Ensemble (Top-3)	71.73%	6.05%	22.22%	0.7475
Standard CoT	69.77%	7.52%	22.71%	0.7353
NLI Check	68.63%	5.23%	26.14%	0.7124
Detailed CoT	62.25%	5.56%	32.19%	0.6503
Counterfactual	55.56%	4.08%	40.36%	0.5760
Detective	48.69%	3.43%	47.88%	0.5041

Table 1: Official evaluation results on the SemEval-2026 Task 12 test set. Our 5-prompt ensemble achieved the highest performance among all our tested variants. Top-3 Ensemble combines Standard CoT, NLI Check, and Detailed CoT.

or picks the option with the most valid votes ($\arg\max_{O_i} V_{\text{valid}}(O_i)$). This aggregation mechanism obviates the need for continuous hyperparameter threshold tuning and enables the system to adapt effectively to both open-world scenarios and forced-choice constraints.

4 Experimental Setup

4.1 Dataset and Evaluation

We evaluate our system on the official dataset provided by SemEval-2026 Task 12, which consists entirely of English-language texts. The corpus comprises 1,819 training instances, 400 development instances, and 612 blind test instances. Each instance consists of a real-world event description, four candidate causal options, and a context of 10 to 20 retrieved documents. To avoid data leakage, we strictly limited prompt tuning and hyperparameter searches to the development set.

The official metrics penalize guessing: Full Match (1.0 point for finding the exact set of correct causes), Partial Match (0.5 points for a proper subset), and Incorrect (0.0 points if any wrong option is included).

4.2 Implementation Details

To guarantee deterministic reasoning and minimize variance across prompt evaluations, we apply near-greedy decoding (temperature = 0.1) for the Qwen3-8B model. During the retrieval phase, document processing is batched to optimize throughput. The reasoning prompts were developed through iterative manual refinement on the development set to systematically mitigate causal hallucination. For the final vote aggregation, a simple majority rule replaces the need for continuous threshold tuning, generalizing robustly across the test set.

Appendix A lists our full hyperparameter settings for retrieval and inference, along with vLLM deployment details.

5 Results and Analysis

5.1 Main Results

Table 1 shows how our system and its variants performed on the 612 blind test instances. Among all our tested configurations, the 5-prompt ensemble achieved the highest score, averaging 0.7614.

The ensemble approach yields an absolute improvement of 4.09% in Full Match accuracy over the best-performing single prompt (Standard CoT), while simultaneously reducing the Incorrect rate from 22.71% to 21.57%.

5.2 Analysis of Prompt Strategies

Weak prompts help the ensemble: Surprisingly, the individually weak prompts contributed heavily to the final score. The *Counterfactual* and *Detective* personas scored poorly on their own (0.5760 and 0.5041), but adding them pushed the Top-3 ensemble score from 0.7475 to 0.7614. They seem to catch causal signals—like missing necessary conditions or negative evidence—that the standard generative prompts miss.

Complexity doesn't mean better reasoning: Making prompts more complex didn't always help. *Standard CoT* (0.7353) beat both *Detailed CoT* (0.6503) and *NLI Check* (0.7124). The explicit checklist in Detailed CoT might have been too rigid, restricting the model's reasoning, while Standard CoT gave it enough room to think.

Voting smooths out errors: The voting mechanism balanced the precision and recall of different prompts. *NLI Check* had high precision but missed some valid causes; the ensemble covered those gaps. The *Detective* persona was especially useful as a regularizer. Even with its low standalone accuracy, it acted as a "devil's advocate," rejecting false correlations that the affirmative prompts accepted.

Persona	Option A	Option B	Option C	Option D
cot_standard	None	✗	✗	✗
nli_check	None	✓	✗	✗
counterfactual	None	✓	✓	✗
detective	None	✓	✗	✗
cot_detailed	None	✗	✗	✗
Count (Valid/Invalid)	None	**3/2**	**1/4**	**0/5**

Table 2: Persona voting breakdown for the failure case (q-2470). Three personas incorrectly accepted Option B as a valid cause. The majority vote led to the wrong prediction.

5.3 Error Analysis

To investigate the limitations of our multi-perspective ensemble, we analyze a representative failure case (Question ID: q-2470). The target event was "Chang'e 5 launched from the lunar surface Dec. 3." The correct answer was Option A ("None of the others are correct causes"), yet our system incorrectly predicted Option B ("Chang'e 5 launched on November 23 atop a Long March 5 rocket").

Table 2 details the voting breakdown across the five personas.

Our analysis of the reasoning traces reveals that the model struggled to distinguish between a necessary precondition and a direct cause. Option B is factually correct and represents the initial Earth launch, which is undeniably a prerequisite for the entire lunar mission. Three out of our five personas (nli_check, counterfactual, and detective) fell into the trap of "causal hallucination" by conflating this distant root cause with the immediate cause of the specific sub-event. For example, the counterfactual persona correctly deduced that without the November 23 launch, the December 3 lunar ascent would not have occurred. However, it erroneously concluded that this counterfactual dependency implies direct causality. Similarly, the detective persona argued that the initial launch "initiated the entire mission," thereby making it the "true cause."

Conversely, the cot_standard and cot_detailed personas correctly identified the logical flaw. They noted that while the November 23 launch placed the spacecraft into orbit, it was merely a precursor; the direct cause of the December 3 launch was the completion of lunar surface operations.

Because three personas voted valid for Option B, the majority voting mechanism aggregated these flawed judgments and selected the incorrect option. This case highlights a key vulnerability in LLM-based causal inference: models often struggle with "strong abduction" in complex, sequential event chains, frequently equating historical necessity with direct, immediate causality. It also demonstrates that while ensemble voting mitigates many errors, it can be overpowered when multiple personas share the same underlying logical bias.

6 Conclusion

We built a training-free retrieve-verify-aggregate pipeline for SemEval-2026 Task 12. LLMs often hallucinate causal links, so we paired strict passage-level retrieval with an ensemble of five reasoning personas. Running these different personas on a single Qwen3-8B model gave us diverse logical checks without the cost of loading multiple models. The system scored 0.7614 on the official test set. The results show that forcing the model to use different reasoning paths helps it ignore irrelevant text and reject false correlations.

Limitations

Running five prompts per option means inference takes five times longer than a standard pass. This latency makes the system unsuitable for real-time use. Our voting method also heavily favors precision. Because it requires a strict majority to accept a cause, it misses valid answers in complex scenarios, hurting our Partial Match score. The voting is also rigid: it forces a yes/no decision even when the personas are completely split, rather than outputting a confidence score. Lastly, the entire pipeline depends on Qwen3-8B's baseline knowledge. If the model simply doesn't know a fact, no amount of prompt engineering will fix it. A natural next step would be an adaptive system that only triggers the full ensemble when the initial prompt is uncertain, balancing accuracy with compute cost.

References

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Han-

nah Rashkin, Doug Downey, Scott Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. *ArXiv*, abs/1908.05739.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, and 9 others. 2022. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, John A. Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuan-Fang Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *ArXiv*, abs/2303.12712.

Pengfei Cao, Mingxuan Yang, Yubo Chen, Chenlong Zhang, Mingxuan Liu, Kang Liu, and Jun Zhao. 2026. Semeval-2026 task 12: Abductive event reasoning: Towards real-world event causal inference for large language models. *Preprint*, arXiv:2603.21720.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, volume ACL 2024 of *Findings of ACL*, pages 2318–2335. Association for Computational Linguistics.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Jun-Mei Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiaoling Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 179 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645:633 – 638.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pages 11733–11763. PMLR / OpenReview.net.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang (Lorraine) Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena Hwang, Soumya Sanyal, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. 2023. Faith and fate: Limits of transformers on compositionality. In *Advances in Neural Information Processing Systems*, volume 36, pages 70293–70332. Curran Associates, Inc.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 113–122, New York, NY, USA. Association for Computing Machinery.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. *ArXiv*, abs/2310.01798.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

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. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Xin Li, Zhuo Cai, Shoujin Wang, Kun Yu, and Fang Chen. 2025. A survey on enhancing causal reasoning ability of large language models. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S.

Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*.

Stephen E. Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3:333–389.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 8634–8652. Curran Associates, Inc.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2022. Efficient transformers: A survey. *ACM Comput. Surv.*, 55(6).

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, S. H. Cai, Yuan Cao, Y. Charles, H. S. Che, Cheng Chen, Guanduo Chen, Huarong Chen, Jia Chen, Jiahao Chen, Jianlong Chen, Jun Chen, Kefan Chen, Liang Chen, Ruijue Chen, Xinhao Chen, and 307 others. 2026. Kimi k2.5: Visual agentic intelligence. *Preprint*, arXiv:2602.02276.

Kangsheng Wang, Xiao Zhang, Zizheng Guo, Tianyu Hu, and Huimin Ma. 2024. Csce: Boosting llm reasoning by simultaneous enhancing of causal significance and consistency. *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023a. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2023b. Unleashing cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. *ArXiv*, abs/2307.05300.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. 2024. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. 2021. Pretrained transformers for text ranking: Bert and beyond. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, page 1154–1156, New York, NY, USA. Association for Computing Machinery.

A Reproducibility Details

We ran Qwen3-8B using vLLM for faster inference. Table 3 shows the exact hyperparameters used for retrieval and generation.

B Multi-Perspective Prompt Templates

To mitigate confirmation bias, our ensemble utilizes five distinct reasoning personas. The exact prompt templates used for each persona are detailed below. Each template is populated with the target event (`{event}`), candidate option (`{option}`), option key (`{option_key}`), and retrieved evidence (`{evidence}`).

Component	Parameter	Configuration
LLM Inference	Base Model	Qwen3-8B
	Framework	vLLM
	Temperature	0.1
	Max Context Window	10,240 tokens
Lexical Retrieval	Algorithm	BM25
	Parameters (k_1, b)	1.5, 0.75
	Recall Limit (k_{lex})	Top-30 passages
	Chunk Size ($L_{\max}$)	1,000 characters
Semantic Reranking	Model	bge-reranker-v2-m3
	Max Sequence Length	1,024 tokens
	Selection Limit (k_{sem})	Top-5 passages

Table 3: Comprehensive hyperparameters for the retrieval-verification pipeline.

Prompt: Standard CoT

You are an expert in causal reasoning. Your task is to determine whether a given option is the direct cause of an observed event.
Event (What happened):
{event}
Option {option_key}:
{option}
Evidence (Retrieved documents):
{evidence}
Instructions:

1. Analyze the evidence carefully.
2. Determine if Option {option_key} directly caused the Event.
3. A cause must happen BEFORE the effect and have a direct causal link.
4. Correlation or temporal proximity alone is NOT sufficient.

Your Analysis:
Think step-by-step:

- What does the evidence say about Option {option_key}?
- Is there a clear causal mechanism connecting Option {option_key} to the Event?
- Did Option {option_key} happen before the Event?

Provide your reasoning, then conclude with either [Valid] or [Invalid].
Reasoning:

Figure 2: Prompt template for the Standard CoT persona.

Prompt: Detailed CoT

You are analyzing causal relationships between events. Determine if the given option is the DIRECT CAUSE of the observed event.
Observed Event:
{event}
Candidate Cause (Option {option_key}):
{option}
Supporting Evidence:
{evidence}
Causal Verification Checklist:
Answer each question based on the evidence:

1. **Temporal Order:** Did Option {option_key} occur BEFORE the Event?
   [ ] Yes, clearly before
   [ ] No, or unclear
2. **Direct Link:** Is there explicit evidence that Option {option_key} led to the Event?
   [ ] Yes, explicit causal language (e.g., "because of", "led to", "caused")
   [ ] No, only correlation or coincidence
3. **Mechanism:** Is there a plausible causal mechanism?
   [ ] Yes, the connection makes logical sense
   [ ] No, the connection is implausible
4. **Alternative Causes:** Could something else better explain the Event?
   [ ] No, Option {option_key} is the best explanation
   [ ] Yes, other factors are more likely

Your Evaluation:
Based on the checklist above, provide your detailed reasoning and final verdict. Conclude with [Valid] if Option {option_key} is a direct cause, or [Invalid] if not.
Analysis:

Figure 3: Prompt template for the Detailed CoT persona.

Prompt: Counterfactual Analysis

You are an expert in counterfactual causal analysis. Evaluate whether the given option caused the observed event.
The Event That Occurred:
{event}
Proposed Cause (Option {option_key}):
{option}
Evidence:
{evidence}
Counterfactual Analysis:
To determine causality, ask yourself: "If Option {option_key} had NOT happened, would the Event still have occurred?"
Step 1: Understand the proposed causal chain
What is the claimed mechanism by which Option {option_key} would lead to the Event?
Step 2: Counterfactual test

- If Option {option_key} never happened, what would be different?
- Would the Event still occur due to other factors?

Step 3: Evidence check

- Does the evidence support this counterfactual dependency?
- Is Option {option_key} necessary (not just correlated) for the Event?

Your Verdict:
Based on counterfactual reasoning, is Option {option_key} a direct cause? Provide your analysis, then conclude with [Valid] or [Invalid].
Reasoning:

Figure 4: Prompt template for the Counterfactual Analysis persona.

Prompt: NLI Check

Task: Determine if the evidence supports that Option {option_key} CAUSED the Event.
Event: {event}
Option {option_key}: {option}
Evidence:
{evidence}
Question: Based on the evidence, does Option {option_key} directly cause the Event?
Consider:

- **Entailment:** Evidence clearly states Option {option_key} caused Event → [Valid]
- **Contradiction:** Evidence shows Option {option_key} did NOT cause Event → [Invalid]
- **Neutral:** Evidence does not establish direct causation → [Invalid]

Your answer must end with [Valid] or [Invalid].
Analysis:

Figure 5: Prompt template for the NLI Check persona.

Prompt: Detective Role-Play

You are a detective investigating the cause of an incident. Your job is to determine if the suspect (Option {option_key}) is the TRUE CAUSE of what happened.
The Incident (Event):
{event}
Suspect (Option {option_key}):
{option}
Evidence Collected:
{evidence}
Detective's Investigation:
As a seasoned detective, you know that:

- Correlation is not causation
- The cause must precede the effect
- There must be a clear causal mechanism
- You need evidence, not just speculation

Examine the evidence:

- Does the evidence place the suspect at the scene before the incident?
- Is there a clear motive and mechanism?
- Could the incident have happened without the suspect?

Your verdict:
Is Option {option_key} guilty of causing the Event? Provide your detective's reasoning, then deliver your verdict: [Valid] for guilty (is the cause), [Invalid] for not guilty (is not the cause).
Detective's Report:

Figure 6: Prompt template for the Detective Role-Play persona.