SC-Refine: Option-Level Self-Consistency with Hybrid Retrieval for Abductive Event Reasoning

Lan Deng
Politecnico di Torino
s338219

Xueyufei Zhang
Politecnico di Torino
s336472

Yufei Chen
Politecnico di Torino
s338116

Chunyi Li
Politecnico di Torino
s338968

Javokhirbek Parpikhodjaev
Politecnico di Torino
s345099

Abstract

This paper presents our system for SemEval-2026 Task 12: Abductive Event Reasoning. We propose **SC-Refine**, a Self-Consistency with Refinement approach designed to optimize for the task's unique partial matching metric. Our system employs a hybrid retrieval strategy to ground reasoning in relevant documents. The core innovation is **option-level** voting, which aggregates votes independently for each option across multiple reasoning chains to better capture partial agreements in multi-label scenarios. Furthermore, we apply adaptive voting thresholds and rule-based post-processing to enforce logical consistency. Our **Balanced** prompting strategy explicitly optimizes for the asymmetric penalty structure. Experimental results demonstrate that our approach achieves **0.74** score (60.3% strict accuracy) on the development set and **0.86** on the official test platform, representing a **32%** relative improvement over single-shot baselines (0.56).

1 Introduction

Every day, the world is shaped by countless events, for example, economic fluctuations, policy decisions, natural disasters, technological breakthroughs. Yet behind every headline lies a deeper question: Why did this happen? Understanding causal relationships is fundamental to human cognition and critical for intelligent systems tasked with interpreting real-world events. While LLMs have demonstrated impressive capabilities in event extraction and summarization, they still struggle with abductive reasoning, inferring plausible causes from incomplete or ambiguous evidence.

SemEval-2026 Task 12 formalizes this challenge as Abductive Event Reasoning (AER), where systems must identify plausible causes for news events by selecting from multiple candidate options based on supporting documents. Unlike standard question answering that tests factual recall, AER requires synthesizing distributed evidence, distinguishing direct causes from correlations or consequences, and reasoning under uncertainty, capabilities essential for high-stakes applications such as misinformation detection, policy impact assessment, and crisis response.

The task presents three unique challenges. First, the evaluation metric imposes asymmetric penalties: selecting any incorrect option yields zero points, while missing some correct options still earns partial credit (0.5 points). This structure prioritizes precision over recall, reflecting real-world scenarios where false positives are more costly than false negatives. Second, causal evidence is often incomplete, scattered across multiple documents, or only implicitly stated, requiring systems to piece together context and background knowledge. Third, multi-label selection with zero to four correct answers demands aggregation strategies that can handle partial agreements rather than voting on complete answer sets.

To address these challenges, we propose a Self-Consistency with Refinement (SC-Refine) approach combining ensemble reasoning with option-level voting. Our key contributions include:

1. An option-level voting mechanism that aggregates votes independently for each option across 7 reasoning chains, better capturing partial agreements in multi-label scenarios;

2. A balanced prompting strategy optimized for asymmetric penalties with evidence-anchored analysis and causal chain validation;

3. A hybrid retrieval module combining BM25 and Sentence-BERT via RRF with adaptive per-option weighting;

4. An adaptive voting thresholds with rule-based post-processing to enforce logical consistency.

2 Background

2.1 Abductive Reasoning in NLP

Abductive reasoning, first formalized by Charles Sanders Peirce, involves inferring the best explanation for a set of observations. In computational settings, this translates to selecting hypotheses that maximize posterior probability given evidence. Recent work has explored abductive reasoning in various NLP tasks, including story understanding (Bhagavatula et al., 2020), common-sense reasoning (Talmor et al., 2019), and scientific hypothesis generation (Zhou et al., 2024).

The ART dataset (Bhagavatula et al., 2020) focuses on narrative understanding, requiring systems to select plausible middle sentences given beginning and ending contexts. However, these tasks typically involve clearly structured narratives with complete information. In contrast, SemEval-2026 Task 12 addresses real-world news events where causal evidence is incomplete, distributed across multiple documents, and often implicit, presenting additional challenges for abductive inference.

2.2 Document Retrieval for Question Answering

Effective question answering requires retrieving relevant documents from large corpora. Traditional approaches like BM25 (Robertson and Zaragoza, 2009) provide strong lexical matching based on term frequency and inverse document frequency. Neural methods using dense embeddings, such as Sentence-BERT (Reimers and Gurevych, 2019) and Dense Passage Retrieval (Karpukhin et al., 2020), capture semantic similarity beyond exact keyword matches.

Hybrid approaches combining lexical and semantic methods have shown superior performance across various domains. Reciprocal Rank Fusion (RRF) (Cormack et al., 2009) provides a simple yet effective method for merging rankings from multiple retrieval systems without requiring score normalization. Recent work (Bruch et al., 2022) has analyzed various fusion functions, demonstrating that RRF achieves robust performance across different ranking distributions.

2.3 Large Language Models for Reasoning

The emergence of large language models (LLMs) has demonstrated impressive reasoning capabilities through carefully designed prompting techniques. Chain-of-Thought (CoT) prompting (Wei and et al., 2022) encourages models to generate intermediate reasoning steps, significantly improving performance on complex reasoning tasks including arithmetic, symbolic manipulation, and commonsense reasoning.

Self-consistency (Wang et al., 2022) extends CoT by sampling multiple reasoning paths and selecting the most consistent answer through majority voting. This approach improves reliability by aggregating diverse reasoning chains, reducing the impact of individual errors or hallucinations.

However, traditional self-consistency votes on complete answer sets, treating different combinations (e.g., {A, B} versus {A, C}) as entirely distinct predictions. This approach suffers from vote fragmentation in multi-label scenarios where reasoning chains may partially agree. Our work addresses this limitation through option-level voting.

3 System overview

Our system employs a **Self-Consistency with Refinement (SC-Refine)** approach that combines ensemble reasoning with option-level voting to address the abductive event reasoning task. The pipeline consists of three main stages: document retrieval, multi-sample generation with self-consistency, and option-level aggregation with post-processing. The "refinement" in SC-Refine refers to deterministic post-processing rules rather than additional LLM-based refinement, maintaining computational efficiency.

3.1 Model Selection

We utilize **DeepSeek-R1** as our primary large language model, selected for its strong reasoning capabilities demonstrated in complex logical tasks. The model is deployed with temperature $= 0.5$ and top-p $= 0.95$ to balance diversity and coherence in generated responses across multiple reasoning chains.

3.2 Document Retrieval Strategy

We employ a **hybrid retrieval system** combining lexical matching (BM25) and semantic search (Sentence-BERT: all-MiniLM-L6-v2) with Reciprocal Rank Fusion (RRF, $k = 60$) to identify the

top-10 most relevant documents from the corpus. The RRF score for a document d is computed as:

$$\mathrm{RRF}(d) = \sum_{m \in \{\mathrm{BM25}, \mathrm{SBERT}\}} \frac{1}{k + \mathrm{rank}_m(d)}.$$

We configure the retrieval system with two key settings: (1) **Full content processing**—we index and retrieve complete document texts rather than title snippets only, as preliminary experiments showed that causal evidence frequently appears in body content rather than titles; (2) **Per-option retrieval**—documents are retrieved separately for the target event (weight $2\times$) and each candidate option (weight $1\times$), then merged using weighted RRF to ensure comprehensive coverage of option-specific evidence.

3.3 Prompt Engineering

We develop a **balanced prompting strategy** that explicitly optimizes for the task's evaluation metric:

- Perfect match ($P = G$): 1.0 points
- Partial match ($P \subset G$, no errors): 0.5 points
- Any wrong selection: 0.0 points

The asymmetric penalty structure (wrong selection $= 0$ points, missing correct $= 0.5$ points) fundamentally shapes our prompt design. The prompt incorporates three critical components:

1. **Evidence-anchored analysis:** The model is required to quote direct textual evidence from documents for each candidate option in a structured table format, marking options with "NONE" when evidence is absent. This structured format reduces hallucination and grounds reasoning in retrieved documents.

2. **Causal chain validation:** The prompt includes explicit warnings to distinguish direct causes from consequences, correlations, and background events (e.g., asking whether an option is a direct cause, a consequence of another option, or merely correlated).

3. **Conservative decision rules:** The prompt explicitly states that selecting any wrong option yields 0 points (complete failure), whereas missing some correct options can still earn partial credit, instructing the model to prioritize precision over recall.

3.4 Self-Consistency Framework

Our SC-Refine approach generates $N = 7$ independent reasoning chains for each question using temperature sampling ($T = 0.5$). Unlike traditional self-consistency methods that vote on complete answer sets (e.g., "A,B" vs. "A,C"), we implement **option-level voting** to better handle multi-label selection:

1. Generate 7 responses with identical prompts but stochastic sampling.

2. Parse each response to extract selected options using the regex pattern: `Final Answer I Reasoned: [A-D](,[A-D])*`.

3. Count votes for each option independently: $V_A, V_B, V_C, V_D \in [0, 7]$.

4. Apply adaptive thresholds: general options (A/B/C) are selected if votes $\geq 4/7$ (57%), while option D is selected if votes $\geq 5/7$ (71%).

The stricter threshold for option D ($\theta_D = 5/7$ versus $\theta_{\text{gen}} = 4/7$) is motivated by empirical observations, requiring higher confidence to avoid false positives. If no option reaches the threshold, we select the option(s) with the highest vote count to ensure at least one answer, as required by the task rules.

This option-level voting strategy provides two key advantages over answer-set voting: (1) it captures partial agreements between reasoning chains (e.g., chains agreeing on option A but disagreeing on B/C), and (2) it provides finer-grained confidence signals for individual options, enabling adaptive thresholding.

3.5 Post-Processing Logic

We implement logical consistency checks to handle edge cases and enforce task constraints:

- **Duplicate option handling:** When multiple options contain identical or nearly identical text (detected via string matching after normalization), they are treated as a single logical unit. If one is selected by voting, all duplicates are automatically selected to maintain logical consistency.

- **Mutual exclusivity enforcement:** If a "None of the others are correct" option is selected alongside substantive options, we remove the

"None" option as they are logically contradictory.

- **Four-option anomaly detection:** When all four options receive votes above threshold, we check for clear weak options (e.g., a single vote versus 5+ for others); otherwise, all four are retained.
- **Empty answer prevention:** The system ensures at least one option is selected in the final output.

3.6 Implementation Details

The system is implemented in Python with parallel processing using ThreadPoolExecutor. For the dev set (400 questions), the SC-Refine approach requires 2,800 LLM API calls (7 per question). Note that actual processing time is highly variable, depending on API response latency, rate limiting, and network conditions. The computational overhead compared to single-shot baselines is 7× in terms of API calls, but the performance gain (0.74 vs 0.56 official score) justifies the additional cost for accuracy-critical applications.

4 Experimental results

We conduct a series of experiments to evaluate the performance of our proposed framework on the SemEval 2026 Task 12: Abductive Event Reasoning.

Our evaluation primarily focuses on the model's inherent capability to identify complex and multi-layered causal relationships, as well as its ability to infer reasonable antecedent explanations from evidence that is incomplete and sometimes inconsistent.

4.1 Exprimental Setup

Dataset We utilize the official AER development set, which consists of a curated set of real-world scenarios drawn from a wide range of domains, including politics, finance, and public emergencies. Each instance includes a target event, referred to as the "effect," along with a large collection of evidence documents from which the corresponding "cause(s)" must be abduced.

Models To ensure a broad and representative evaluation, **DeepSeek-R1**, which serves as our primary backbone due to its strong performance in long-context handling and logical reasoning. All experiments employ a semantic retrieval pipeline based on BM25 and Scentence-BERT, configured to retrieve the top-$k = 10$ document snippets in order to construct a focused yet adequate context window. For our main proposed setting (**SC-Refine + Balanced**), we use a sampling temperature of $T = 0.5$ and generate $N = 7$ independent reasoning paths to enable majority voting, while all baseline settings use $T = 0$ to maintain deterministic outputs.

Evaluation We report three evaluation metrics: the *Official Score* (1.0/0.5/0 scale), *Strict Accuracy (SA)*, and *Macro F1*. Among them, the Official Score is treated as the primary metric, as its design assigns partial credit to partially correct predictions while penalizing over-inference with zero scores, making it well suited for assessing abductive reasoning where precision is more important than recall.

4.2 Main Results

Table 1 shows the results of the best-performing approach on the AER development set. The model correctly predicts all causal options in 241 cases and achieves partial correctness in 108 cases, while only 51 instances are fully incorrect. This result indicates that the system is able to identify correct causes without frequently selecting incorrect options. As a consequence, it achieves an official score of 0.74, demonstrating effective performance under the task's asymmetric evaluation metric.

Table 1: Best performance on the AER development set.

Metric	Value
Correct Answers	241
Partial Correct	108
Incorrect Answers	51
Total Questions	400
Official score	0.74

Table 2 summarizes the performance across different reasoning paradigms, illustrating the significant performance gap between standard baseline approaches and our refined methodology.

The empirical results indicate a clear differences among different reasoning strategies. The proposed **SC-Refine + Balanced** achieves an official score of **0.74** on development set (and **0.86** on official test set), corresponding to a relative improvement of **32%** over the zero-shot CoT baseline. This improvement suggests that combining a structured reasoning framework with an ensemble-based refinement mechanism plays an important role in

Table 2: Performance comparison on the AER development set.

Approach	Prompt	Model†	Official Score	Strict Acc.	Partial %	Macro F1
Baseline	Zero-shot CoT	deepseek-r1	0.56	48.00%	15.00%	0.73
	Conservative	deepseek-v3.2	0.42	39.25%	4.75%	0.69
	Balanced	deepseek-v3.2	0.47	41.25%	11.00%	0.67
Conservative	Zero-shot CoT	deepseek-v3.2	0.66	63.00%	6.25%	0.82
	Conservative	deepseek-r1	0.56	46.50%	19.00%	0.71
	Balanced	deepseek-r1	0.52	45.75%	11.75%	0.71
SC-Refine	Zero-shot CoT	–	–	–	–	–
	Conservative	deepseek-r1	0.53	48.00%	10.5%	0.75
	Balanced	**deepseek-r1**	**0.74**	**60.25%**	**27.00%**	**0.79**

† Due to the expiration of the DeepSeek-R1 model provided by the API vendor, some later experiments were performed with DeepSeek-V3.2 as an alternative.

handling complex evidence.

In comparison, while the *Conservative* approach achieves high precision due to its risk-averse design, it is still limited by lower recall, as the use of strict confidence thresholds often causes the model to exclude valid causal relationships in order to maintain certainty.

4.3 Ablation Study

To analyze the overall performance of our system and measure the specific impact of each architectural part, we conduct a detailed ablation study. This step is necessary to determine whether our improvements are driven by better data grounding or by improved reasoning logic.

The largest decrease in performance happens when Document Retrieval is removed, resulting in a **27.0%** drop. This shows that abductive reasoning is highly sensitive to noisy information contained in the context documents. If there is no specific process to handle these noises, the performance will be greatly reduced because the model may rely on false evidence instead of facts.

Additionally, there is a clear drop after removing Balanced prompting suggests that even with good evidence, the model still needs a clear structure for reasoning to tell the difference between actual causes and events that just happen first.

4.4 Case Study Analysis

To show in detail how our approach moves the model away from simple associations toward a better understanding of causes, we examine the "2021 Texas Power Crisis". This specific example demonstrates the system's capacity to tell the difference between the initial trigger of the crisis and the many serious problems that occurred as a result.

Scenario and Conflict Target Event: *"The Electric Reliability Council of Texas (ERCOT) initiated statewide rolling outages in February 2021."* Evidence Analysis: Document A explains how a very strong cold front made power equipment freeze and stop working; Document B describes the massive power failures that happened next and the many problems they caused for the people in the area.

The core challenge is to isolate the structural failure from the resulting outage, as they are semantically nearly in daily life.

Deep Trace Comparison

- **Baseline Performance**: The Baseline CoT model picked both "extreme cold" and "outages" as the causes. It reasoned that "the power was out because there was an outage." This is circular reasoning, where the model mixes up the final result (the outages) with the actual starting cause (the equipment failure). Because the model included incorrect options in its final answer, it received a **score of 0**.
- **Proposed Framework Performance**:
  1. **The Balanced Phase**: The model explicitly constructed a timeline, noting that the "outages" were the *result* of the grid's inability to meet demand triggered by the cold.
  2. **The SC-Refine Phase**: Across 7 independent paths, the system identified that the outages occurred *t+1* relative to the equipment failure. 6 out of 7 paths correctly rejected the "outages" option as a

Table 3: Ablation results on the AER development set.

Configuration	Score
Full System (SC-Refine + Balanced)	**0.74**
− w/o Balanced Prompting (Unoptimized Balanced)	0.58 (−22.6%)
− w/o Document Retrieval (Unprocessed Knowledge)	0.54 (−27.0%)
− w/o Per Option Retrieval (Options Excluded from Retrieval)	0.71 (−4.1%)

† Due to training cost considerations, the ablation study was conducted based on a sampling rate of $N = 5$.

cause. Final selection: "extreme cold" and "natural gas infrastructure failure." **Score: 1.0**.

5 Conclusion

This paper presented a Self-Consistency with Refinement approach for abductive event reasoning, achieving 0.74 official score on the development set and 0.86 on the test platform, a 32% relative improvement over baselines. Our core innovation, option-level voting, aggregates votes independently for each option across 7 reasoning chains, effectively capturing partial agreements in multi-label scenarios. Combined with balanced prompting, hybrid retrieval, and adaptive thresholds, the system demonstrates strong performance while maintaining interpretability through explicit voting statistics.

However, limitations include 7× computational overhead, fixed voting thresholds that may not generalize across domains, brittle rule-based post-processing, and persistent challenges in distinguishing causes from correlations under high semantic overlap. Future work could explore adaptive thresholding, LLM-based critique phases, learned post-processing models, structured causal graph reasoning, and uncertainty quantification for selective prediction.

Despite these limitations, our work demonstrates that option-level self-consistency provides an effective framework for tasks with partial matching metrics, where decomposing multi-label voting to individual options better exploits ensemble information. As LLMs advance, techniques for structured ensemble reasoning will become increasingly important for accuracy-critical applications with asymmetric error costs.

References

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *International Conference on Learning Representations (ICLR)*.

Sebastian Bruch, Siyu Gai, and Amir Ingber. 2022. An analysis of fusion functions for hybrid retrieval. *arXiv preprint*.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. 2009. Reciprocal rank fusion outperforms condorcet and individual ranker methods. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759. ACM.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of EMNLP-IJCNLP*.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. In *Advances in Neural Information Processing Systems*.

Jason Wei and et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. 2024. Hypothesis generation with large language models. In *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*.