

SemEval-2026 Task 12: Abductive Event Reasoning: Towards Real-World Event Causal Inference for Large Language Models

Pengfei Cao^{1,2*}, Mingxuan Yang^{1,2*}, Yubo Chen^{1,2},
Chenlong Zhang^{1,2}, Mingxuan Liu^{1,2}, Kang Liu^{1,2}, Jun Zhao^{1,2†}

¹The Key Laboratory of Cognition and Decision Intelligence for Complex Systems,
Institute of Automation, Chinese Academy of Sciences, Beijing, China

²School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China

{yangmingxuan2025, zhangchenlong2024, liumingxuan2025}@ia.ac.cn

{pengfei.cao, yubo.chen, kliu, jzhao}@nlpr.ia.ac.cn

Abstract

Understanding why real-world events occur is important for both natural language processing and practical decision-making, yet direct-cause inference remains underexplored in evidence-rich settings. To address this gap, we organized SemEval-2026 Task 12: Abductive Event Reasoning (AER).¹ The task asks systems to identify the most plausible direct cause of a target event from supporting evidence. We formulate AER as an evidence-grounded multiple-choice benchmark that captures key challenges of real-world causal reasoning, including distributed evidence, indirect background factors, and semantically related but non-causal distractors. The shared task attracted 122 participants and received 518 submissions. This paper presents the task formulation, dataset construction pipeline, evaluation setup, and system results. AER provides a focused benchmark for abductive reasoning over real-world events and highlights challenges for future work on causal reasoning and multi-document understanding.

1 Introduction

Understanding why real-world events occur is important for both natural language processing and practical decision-making (Waldmann and Hagmayer, 2013). Identifying direct causes, however, remains difficult in realistic settings because evidence is often incomplete, distributed across multiple documents, and mixed with background conditions, consequences, or semantically related but non-causal events (Tu et al., 2019; Min et al., 2025; Yang et al., 2018; Trivedi et al., 2022). Despite progress in event causality identification (Cao et al., 2021b; Zuo et al., 2021; He et al., 2024; Mirza, 2014; Romanou et al., 2023), event relation extraction (Cao et al., 2021a; Hao et al., 2023; Sui et al.,

*Equal contribution.

†Corresponding author.

¹The task data is available at <https://github.com/s0oo66/semEval2026-task12-dataset.git>

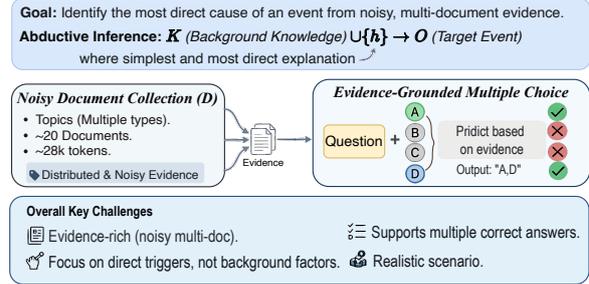


Figure 1: Overview of the Abductive Event Reasoning (AER) task. Given a noisy multi-document evidence collection, the goal is to identify the most plausible direct cause of a target event via evidence-grounded abductive inference. The task is challenging because the supporting evidence is distributed and noisy, systems must focus on direct triggers rather than background conditions, and multiple answer options may be correct.

2023; Tang et al., 2021; Wang et al., 2022), and abductive reasoning (Bhagavatula et al., 2020; Ponti et al., 2020), direct-cause inference under noisy multi-document evidence remains underexplored.

To address this gap, we organized SemEval-2026 Task 12: Abductive Event Reasoning (AER), a shared task that asks systems to identify the most plausible direct cause of a target event from supporting evidence (Figure 1). We formulate AER as an evidence-grounded multiple-choice benchmark for real-world event causal reasoning. Unlike traditional causal relation extraction, which often focuses on predefined or locally expressed event pairs, AER requires systems to reason over noisy multi-document evidence, distinguish direct causes from broader contextual factors, and avoid semantically plausible but causally incorrect distractors (Romanou et al., 2023; Trivedi et al., 2022).

To support this task, we constructed a benchmark through a multi-stage pipeline involving document collection, event extraction, timeline construction, multi-model causality scoring, and human verification. The resulting dataset contains 60 topics and

2,831 question instances derived from real-world reports spanning 2016 to 2025. On average, each topic contains 19.7 documents, with supporting evidence averaging approximately 28K tokens. The difficulty of the task arises not only from causal ambiguity but also from the scale and noisiness of the evidence, which place substantial demands on long-context understanding and cross-document integration (Bai et al., 2024; Trivedi et al., 2022).

The shared task attracted 122 participants and received a total of 518 submissions. Ultimately, 21 teams submitted system description papers. This paper presents the overall task design, dataset construction details, and key characteristics of the participating systems. The submitted systems demonstrate how current machine learning techniques perform when addressing complex causal reasoning under long, noisy multi-document evidence, and they provide valuable insights for future improvements in causal reasoning, multi-document understanding, and robust abductive reasoning.

2 Related Work

Event Causality Reasoning. A substantial body of prior work has studied causality in text, including temporal and causal relation extraction (Mirza and Tonelli, 2014), unified event relation extraction (Wang et al., 2022), and the construction of causal reasoning benchmarks (Romanou et al., 2023). Much of this literature focuses on relatively local settings, such as sentence-level or document-level relation classification (Mirza and Tonelli, 2014; Zhao et al., 2021; Liu et al., 2024), often assuming explicit relation instances or predefined event pairs. More recent work has moved toward richer causal benchmarks that consider contextualized event relations and graded causal strength (Romanou et al., 2023; Yuan et al., 2023). In particular, recent studies highlight that causal judgments over real-world events are often context-dependent, temporally constrained, and more complex than simple binary cause-effect detection (Halpern, 2016; Pearl, 2009). Our task is closely related to this line of work but differs in formulation: rather than classifying predefined event pairs, AER asks systems to select the most plausible direct cause of a target event from multiple evidence-grounded candidates.

Multi-document Reasoning. Our task is also related to research on reasoning over information distributed across multiple documents. Prior work has shown that language understanding becomes

significantly more difficult when evidence must be aggregated from multiple sources rather than read from a single passage (Tu et al., 2019; Yang et al., 2018). This challenge is particularly relevant for causal reasoning over real-world events, where supporting evidence may be scattered across reports published at different times and mixed with noise or irrelevant background information. Although recent work has explored cross-document or real-world causal reasoning benchmarks (Romanou et al., 2023; Min et al., 2025), these datasets are often smaller or more structured. Related work on multi-hop QA and long-context evaluation has likewise shown that models remain vulnerable to shortcut reasoning and degradation over long, compositional evidence chains (Trivedi et al., 2022; Bai et al., 2024). In contrast, AER introduces a larger and more realistic noisy multi-document environment, significantly increasing the difficulty of long-context comprehension, noise robustness, and fine-grained causal discrimination. This makes AER a more challenging and representative benchmark for real-world event causal reasoning.

3 Task Description

3.1 Task Overview

SemEval-2026 Task 12: Abductive Event Reasoning (AER) aims to evaluate whether systems can identify the most plausible and most direct cause of a target event from supporting evidence. The task is grounded in the concept of *abductive reasoning*, which involves finding the best explanation for an observed outcome (Walton, 2014). Formally, let O denote the observed event, K denote background knowledge, and H denote the set of candidate hypotheses derived from the available context. The goal of abductive reasoning is to select the simplest and most direct hypothesis $h \in H$ such that

$$K \cup \{h\} \models O$$

That is, the hypothesis together with the background knowledge forms an explanation for the observed outcome. In our setting, the *target event* corresponds to the observed outcome, the *candidate causes* correspond to explanatory hypotheses supported by the provided evidence, and the model’s internal knowledge and reasoning capability correspond to the background knowledge.

A central design principle of AER is the distinction between **direct causes** and broader background factors. Real-world events are often associ-

ated with many temporally related or semantically similar developments, but not all of them provide explanations of equal strength (Halpern, 2016). For example, when the target event is “severe flooding in a certain region,” the evidence may include both “prolonged heavy rainfall” and “global climate warming trends.” The task requires the system to prioritize identifying “heavy rainfall” as the immediate triggering factor rather than treating climate warming as an equally direct cause, even though it may be relevant in a broader context. In short, AER focuses on causes that **directly trigger** the target event, rather than factors that merely provide background conditions. This distinguishes the task from simple topic matching and also makes it more challenging than traditional sentence-level causal relation extraction.

3.2 Task Formalization

We formulate AER as an **evidence-based multiple-choice task** that allows **multiple correct answers**. Each instance is defined as a triple $I = (e_t, D, C)$, where e_t is the target event, $D = \{d_1, d_2, \dots, d_n\}$ is an unordered set of supporting documents, and $C = \{c_A, c_B, c_C, c_D\}$ is a set containing exactly four candidate options. The system is required to output a subset of option labels indicating all correct direct causes.

Allowing multiple correct answers is an intentional design choice rather than an annotation artifact. Because causality is inherently non-binary, a target event in real-world event chains may have more than one directly relevant triggering factor that is clearly supported by the evidence (Halpern, 2016; Romanou et al., 2023). Therefore, this benchmark evaluates whether systems can recover a *complete and plausible set of direct causes*, rather than forcing the problem into a single-answer format.

The official shared task consists of a single track. Participants are provided with the *full document collections* rather than curated timelines or summarized evidence. On average, each topic contains 19.7 documents, with an average topic-level evidence length of 28,047 tokens. Importantly, these document collections are **intentionally noisy**: besides relevant reports, they may also include documents that are topically similar but causally irrelevant. The inclusion of such distractors is designed to better approximate real-world evidence conditions and to reduce the effectiveness of shallow lexical matching strategies.

3.3 Input and Output Format

Each problem instance contains exactly four candidate options labeled A to D. System outputs should be a comma-separated string of option labels, for example B, C. Since multiple answers may be correct, predictions may include one or more labels. Predictions must only contain valid labels from A–D; invalid outputs receive a score of 0 for the corresponding instance.

For each instance, the evidence is provided as a collection of documents associated with the corresponding topic. Documents are treated as an unordered set, and participating systems must determine which options are directly supported as causes of the target event based on these documents.

3.4 Evaluation Metrics

The official evaluation metric is **instance-based accuracy**, which accounts for both exact matches and partial matches. Let G denote the gold set of correct options and P denote the system prediction set. The score for a single instance is defined as follows:

- **1.0 (Exact Match)**: if $P = G$;
- **0.5 (Partial Match)**: if P is a non-empty proper subset of G ;
- **0.0 (Incorrect)**: all other cases, including when P contains any incorrect option or when P is empty.

The final system score is the average of the instance-level scores across all evaluation instances. This scoring scheme encourages precise causal predictions and penalizes over-prediction, which is particularly important in AER since some instances contain multiple correct answers.

3.5 Task Example

Table 1 presents an example from the benchmark. The target event corresponds to an observed outcome, the evidence is provided by a set of retrieved documents, and the candidate options include both plausible and implausible explanations. In this example, the direct cause is the shooting incident, while the other options correspond to indirect factors, subsequent events, or unrelated distractors.

Target Event	Videos of the assassination circulated on social media.
Context	A collection of documents containing mixed relevant and irrelevant evidence
Options	
A ✗	The shooter used a handmade gun.
B ✗	Security arrested the suspected gunman, Tetsuya Yamagami.
C ✗	Shinzo Abe became the deputy chief cabinet secretary in the early 2000s.
D ✓	A man fired twice at Shinzo Abe.
Answer	D

Table 1: An illustrative example from the AER dataset. The target event concerns the circulation of assassination videos on social media, and the candidate causes range from peripheral background information to the most immediate precipitating event. This example highlights the task’s focus on selecting the most plausible direct cause rather than merely related context.

4 Dataset Construction

The AER benchmark is derived from real-world news reports published between 2016 and 2025. The dataset was constructed through a multi-stage pipeline that combines large language models with human verification, with the goal of producing evidence-based multiple-choice questions for direct cause inference.

4.1 Overall Pipeline

Inspired by the framework of Romanou et al. (2023), we designed the data construction pipeline shown in Figure 2. We first collected topic-centered document sets using the Google News API. We then extracted sentence-level events from the retrieved documents and organized them into topic-level timelines. Next, we scored candidate causal relations using multiple language models and combined these scores with human verification. Finally, the verified event pairs were converted into multiple-choice question instances.

4.2 Data Collection and Event Extraction

Our source documents were collected through the Google News API.² We retrieved news reports centered on specific topics and filtered the results using fixed time windows. The resulting document sets were intended to cover both the core trajectory of each event and its surrounding background context. To make the task more realistic and reduce the effectiveness of shallow lexical matching strategies, we deliberately introduced distractor documents. These documents were obtained by querying with keywords that were semantically related

to the topic but tended to lead to reports outside the target event chain.

For the retrieved documents, we used GPT-4.1 (OpenAI, 2025a) to extract sentence-level event mentions. The extracted outputs were then manually inspected to remove malformed or low-quality results. In total, the initial collection contained 913 unique event mentions, which were subsequently normalized into 865 timeline-level events. Sentence-level event extraction provides a practical balance between granularity and scalability: it is fine-grained enough to support causal reasoning, while keeping the complexity of subsequent alignment and annotation stages manageable.

4.3 Timeline Construction

After event extraction, we used the then state-of-the-art large language model GPT-4.5 (OpenAI, 2025b) to organize event mentions into topic-level timelines, followed by human verification. During this process, mentions that referred to the same real-world event but were expressed differently were merged and disambiguated through a combination of prompt-based LLM processing and manual validation.

4.4 Causality Scoring and Human Verification

For each target event, we treated all events occurring earlier on the timeline as candidate causes. Each candidate event pair was then scored by three models: GPT-4.1 (OpenAI, 2025a), Gemini-2.0-Flash (Gemini Team and Google, 2025), and Claude-3.7-Sonnet (Anthropic, 2025). Each model produced a causality score between 0 and 100. We then computed the variance across the three model scores for each candidate pair and used this signal to identify uncertain cases prior to human annota-

²<https://serper.dev/>

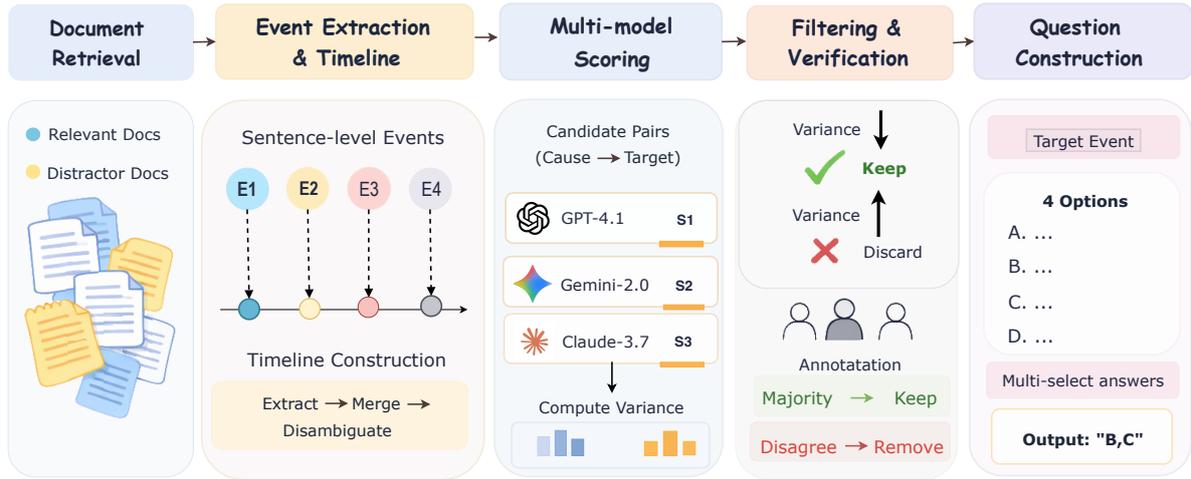


Figure 2: Construction pipeline of the AER benchmark.

tion.

The remaining candidate pairs were manually verified by three annotators. Each event pair was independently labeled by all three annotators. Samples with complete disagreement among the three annotators were discarded, whereas samples with a majority decision were retained. In the final release, we use Krippendorff’s α as the primary inter-annotator agreement metric (Krippendorff, 2011). Final agreement statistics are reported in Section 4.7.

4.5 Question Generation

Strictly verified causal event pairs were used to construct the final multiple-choice questions. To ensure task difficulty and prevent shortcut learning, we adopted a difficulty-aware stratified distractor design strategy. Specifically, for each target event E_t , we constructed a positive candidate pool (P) and a negative distractor pool (N) based on the timeline and causal labels.

The distractor pool N was further divided into three categories: (1) **temporal distractors**: events occurring after E_t on the timeline, i.e., outcomes or subsequent developments; (2) **semantic distractors**: events that share salient entities or keywords with E_t but do not bear a substantive causal relation in the given context; and (3) **indirect/background factors**: events that received relatively low scores during the initial scoring stage and typically correspond to long-term background conditions or indirect causes.

During instance assembly, we sampled options to create four-choice (A/B/C/D) test items for each

E_t . In addition, we introduced *None of the others are correct causes* instances to test model abstention ability and mitigate hallucination.

4.6 Dataset Statistics

The final benchmark contains 60 topics and 2,831 question instances. The training, development, and test sets contain 1,819, 400, and 612 instances, respectively. The benchmark is built on topic-organized document collections spanning real-world reports from June 2016 to November 2025, and includes 913 unique event mentions and 865 timeline events.

On average, each topic contains 19.7 documents, and the average topic-level evidence length is 28,047 tokens; the longest topic reaches 91,417 tokens. The average document length is 1,088.6 tokens. These statistics indicate that AER not only covers a wide range of topics, but also involves lengthy and highly variable contextual evidence, posing a substantial challenge for long-context understanding.

AER is designed as a genuinely multi-answer benchmark rather than a traditional single-label multiple-choice task. In the training and development sets, 56.42% of instances have only one correct answer, whereas 43.58% contain multiple correct answers; the average number of gold labels per instance is 1.57. This shows that multi-answer reasoning is not a marginal phenomenon, but a core property of the benchmark.

Regarding option distribution, after aggregating all positive labels, the positions of correct answers across A–D are relatively balanced, indicating that

Statistic	Value
Number of topics	60
Total number of questions	2,831
Training questions	1,819
Development questions	400
Test questions	612
Time span	2016-06 to 2025-11
Unique event mentions	913
Timeline events	865
Average documents per topic	19.7
Average topic evidence length	28,047 tokens
Average document length	1,088.6 tokens

Table 2: Core statistics of the AER benchmark. Agreement statistics will be added after the final annotation analysis is completed.

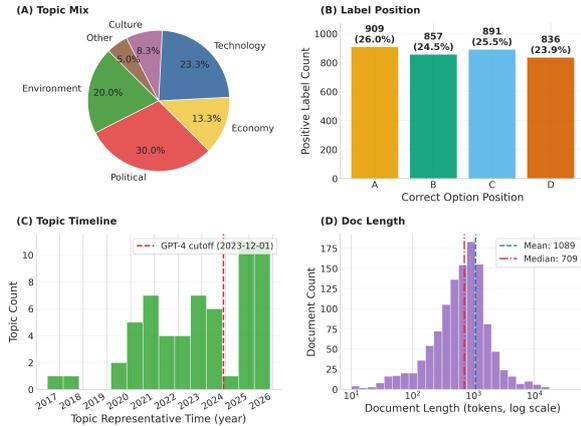


Figure 3: Distributional overview of the dataset: (A) topic composition across six categories, (B) frequency of correct labels at options A–D, (C) topic representative-time distribution with the GPT-4 knowledge cutoff (2023-12-01) marked by a dashed line, and (D) document length distribution in tokens, with mean and median indicated.

models cannot benefit from fixed positional biases. At the topic level, representative topic timestamps span multiple years, showing that the benchmark is not concentrated within a narrow time range, but is instead grounded in news collections reflecting long-term real-world developments.

The document length distribution further shows that although most documents remain within a manageable range, a substantial proportion of long documents still exists, increasing the difficulty of evidence retrieval and direct cause identification.

Table 2 summarizes the core statistics of the AER benchmark. Figure 3 further illustrates several key dataset distributions through four subfigures, including document category proportions, the positional distribution of correct answers, the temporal distribution of topic representation, and the document length distribution.

4.7 Annotation Quality

Constructing an evidence-based benchmark for direct cause identification is inherently challenging. In real-world events, causal relations are often not fully determinate. On the one hand, an outcome event may plausibly admit multiple candidate causes. On the other hand, the causal strengths of different candidates may be highly similar, making them difficult to distinguish consistently (Halpern, 2016).

Moreover, topic-level document retrieval inevitably introduces some noise, and several stages of the pipeline—including event extraction, timeline construction, and candidate event selection—partly rely on LLM outputs. Human verification is therefore essential for ensuring dataset quality.

To improve annotation reliability, we asked three annotators to independently assign a three-way label to each candidate event pair. Specifically, given the available evidence, annotators labeled the relation between a candidate cause event and a target event as one of the following:

- (1) *No causal relation*: the candidate event cannot be regarded as a direct cause of the target event.
- (2) *Moderate causal relation*: there is some direct causal connection between the two events, but the supporting evidence is limited in strength or exclusivity.
- (3) *Strong causal relation*: the candidate event is a clear and strongly supported direct cause of the target event.

We adopt Krippendorff’s α as the primary inter-annotator agreement metric (Krippendorff, 2022). The overall agreement among the three annotators on the three-way classification task is $\alpha = 0.51$, indicating moderate agreement.

5 Baselines and Pilot Analysis

Before the formal evaluation phase, we conducted a pilot study to assess the difficulty of the benchmark and establish reference baselines. The purpose of this study was not to define additional shared-task tracks, but rather to understand how current strong language models perform under different evidence conditions.

5.1 Baseline Models

We evaluated three strong language models in a zero-shot prompting setting: GPT-4 (OpenAI, 2023), Qwen-2.5-72B-Instruct (Qwen Team, 2025), and GLM-4 (Zeng et al., 2024). These models were

selected to cover both leading proprietary systems and open-weight models with strong general reasoning capabilities. All models were asked to predict answers directly from the provided evidence and candidate options, without task-specific fine-tuning.

5.2 Experimental Setup

We tested the models on a subset of 200 instances under two input conditions:

(1) Original text (Ori_T): the full set of retrieved documents, including distractor documents;

(2) Summarized text (Sum_T): timeline-style summaries generated by prompting Gemini 2.0 (Gemini Team and Google, 2025), which preserve the core event evolution while reducing redundancy and removing much irrelevant material.

These settings were used only for pilot analysis and baseline comparison, and do not constitute separate official tracks of the shared task.

5.3 Pilot Results

Table 3 presents the pilot results. Across all tested models, performance drops when the input is changed from summarized evidence to the original document collections, indicating that evidence noise and long-context burden substantially affect direct cause inference.

Model	Ori_T	Sum_T
GPT-4	68.66	70.35
Qwen-2.5-72B-Instruct	53.65	60.72
GLM-4	58.12	60.36

Table 3: Pilot results on a 200-instance sample under original-document input (Ori_T) and summarized input (Sum_T).

5.4 Preliminary Findings

The pilot study reveals two central difficulties captured by AER. First, systems struggle to locate and integrate causal evidence within long, noisy multi-document contexts. Even strong models perform noticeably better when the evidence is compressed into cleaner summaries. Second, systems are vulnerable to semantic confounds: when exposed to the full evidence collection, they are more easily misled by background conditions or related events that are semantically similar but not actually causal.

These findings support the motivation behind the benchmark design. The difficulty of AER does not arise merely from asking systems to select from a

set of options; rather, the real challenge lies in the fact that relevant evidence is scattered, noisy, and mixed with plausible distractors.

6 Participation and Results

6.1 Participation Overview

The shared task was hosted on the Codabench platform and attracted broad participation throughout the evaluation period. In total, 122 participants registered for the task and made 518 submissions to the single official track. After the evaluation concluded, 21 teams submitted system description papers, among which 19 systems were retained as valid entries on the final leaderboard.

6.2 Overview of Submitted Systems

The submitted systems can be broadly categorized into four methodological families.

Retrieval-centered, evidence-grounded pipeline systems. These systems typically separate evidence selection from final decision making explicitly, often combining sparse retrieval, dense retrieval, reranking, and evidence verifiers or cross-encoder reasoning modules. Representative examples include the winning system AILS-NTUA, which uses graph-based retrieval followed by large language model reasoning and post-hoc correction through consistency constraints, and the University of Tübingen system, which combines option-oriented retrieval with an independent cross-encoder verifier. Similar retrieval-verification designs also appear in systems such as uir-cis, sutta, thiyaga6851, AI4PC-Howard University, and X-NLP. Overall, these systems share the intuition that the main bottleneck of AER lies in filtering out distractors and locating direct causal evidence before reasoning begins.

LLM prompting-based methods. These methods rely on large instruction-tuned models and often incorporate carefully designed prompts, role-based reasoning, self-consistency, or constrained decoding. For example, AILS-NTUA combines structured prompting, reflective prompt evolution, and self-consistency; uir-cis emphasizes high-precision decoding strategies to suppress over-selection. Paradise, pushkar, and X-NLP likewise treat prompt engineering, chain-of-thought reasoning, or retrieval-augmented prompting as core components. Such systems generally frame abductive reasoning as a controlled inference problem in

which output calibration is at least as important as raw reasoning ability.

Supervised fine-tuning and distillation-driven systems. Unlike methods that mainly rely on inference-time prompting, these systems directly train discriminative models or compact student models on the task data. CausalMinds is the clearest example: it reports that simply fine-tuning GPT-4.1-mini with option-shuffling augmentation outperforms more complex prompting and verification pipelines. HCMUS_RepeatedGames combines hybrid retrieval with extended LoRA fine-tuning (Hu et al., 2022) of a 32B model, while YNU-HPCC uses a DeBERTa-based (He et al., 2020) evidence-aware classifier together with teacher distillation. KDW also partially belongs to this category, as it employs a lightweight student model obtained through knowledge distillation. Taken together, these systems suggest that targeted supervision can sometimes compensate for architectural simplification on this benchmark.

Knowledge-enhanced and neuro-symbolic methods. These methods attempt to strengthen textual reasoning with explicit causal structure, such as evidence graphs, knowledge graphs, or theorem-proving-style verification processes. KDW combines knowledge-graph-based evidence extraction with distillation; OseiBrefo-Liang integrates hybrid retrieval, causal graphs, and neuro-symbolic policy optimization; and the Newcastle submission explores COMET-enhanced causal knowledge graphs together with Euclidean and hyperbolic embeddings. Although methodologically interesting, these approaches are, overall, less consistently competitive than the strongest retrieval-centered or fine-tuning-based systems.

Overall, the submitted systems exhibit a clear methodological pattern: most teams did not treat the task as ordinary multiple-choice question answering, but instead decomposed it into subproblems such as evidence selection, per-option verification, calibration, and structured set prediction.

6.3 Leaderboard Results

Table 4 presents the final leaderboard. The best-performing system, AILS-NTUA, achieved a score of 0.95, establishing a clear lead over the rest of the field. The next two teams, d-itlab and HCMUS_RepeatedGames, reached 0.91 and 0.90, respectively, showing that multiple methodological paths can achieve strong performance. Beyond this

Rank	Team Name	Score
1	AILS-NTUA	0.95
2	d-itlab	0.91
3	HCMUS_RepeatedGames	0.90
4	KDW	0.88
5	CausalMinds	0.88
6	X-NLP	0.84
7	Youngee Jeong et al.	0.84
8	uir-cis	0.80
9	Paradise	0.79
10	Bolun Liang et al.	0.78
11	sutta	0.76
12	pushkar	0.74
13	YNU-HPCC	0.73
14	Clutch or Cry	0.72
15	OseiBrefo-Liang	0.61
16	Thiyaga6851	0.56
17	BBgame	0.54
18	AI4PC-Howard University	0.53
19	REGLAT	0.30

Table 4: Final leaderboard of the shared task.

leading group, performance declines more gradually, forming a relatively dense middle tier roughly between 0.73 and 0.84, with lower-ranked systems forming a long tail.

This distribution supports two important observations. First, the task is largely solvable when systems can explicitly control distractors and define clear prediction boundaries. Second, the substantial gap between top and lower-ranked systems indicates that seemingly small design choices—especially in evidence selection, option isolation, and decoding calibration—can have major effects under the official evaluation metric.

The strongest systems exhibit distinct but convergent strategies. The top-ranked AILS-NTUA system combines graph-based retrieval, structured LLM reasoning, self-consistency, and iterative post-hoc consistency heuristics. This design directly addresses several failure modes in AER, including distractor documents, repeated options, and logically inconsistent multi-label outputs. d-itlab adopts a highly precision-oriented architecture in which candidate options are evaluated independently via multi-stage LLM gating and then combined with surprisal features in a final XGBoost ensemble (Chen and Guestrin, 2016). This system is explicitly optimized to avoid false positives, which aligns closely with the task metric. In contrast, HCMUS_RepeatedGames combines hybrid retrieval with extended LoRA fine-tuning of a 32B model, showing that large models can be highly effective when strong supervised adaptation is paired with evidence selection.

Among the remaining high-performing systems, KDW employs a logic-driven pipeline that combines knowledge-graph-based evidence extraction, distillation, and explicit handling of the None of the others option, whereas CausalMinds demonstrates that simple fine-tuning with option-shuffling augmentation can match or surpass more complicated multi-stage pipeline systems. Notably, KDW and CausalMinds achieved the same score as some more complex multi-stage systems. This suggests that success on this task does not depend on a single architecture, but rather on whether a system can reliably distinguish direct causes from semantically plausible distractors while avoiding over-prediction.

6.4 Methodological Trends

The dominance of retrieval and evidence compression. Many teams observed that the document packages provided by the task contain substantial irrelevant or only weakly relevant material. As a result, an increasing number of systems adopted hybrid retrieval, chunking, reranking, graph traversal, or option-oriented retrieval to reduce attention to distractors before final reasoning. This pattern is especially visible in top-ranked and upper-middle-ranked systems.

The importance of precision-oriented decision mechanisms. Because the evaluation setup awards no credit when incorrect options are included, many teams explicitly biased their systems toward more conservative prediction behavior. This tendency is reflected in threshold tuning, per-option verification, majority voting, constrained decoding, and post-hoc logical filtering. In other words, the evaluation setting encourages systems to behave more like cautious verifiers than aggressive generators.

The relative roles of prompting and fine-tuning. Prompt-based systems are widespread and remain competitive when combined with retrieval and calibration. However, the results also suggest that prompting alone is not always sufficient. Some teams reported that, in certain settings, chain-of-thought or multi-stage prompting could even hurt performance by amplifying spurious reasoning paths. By contrast, fine-tuned systems such as CausalMinds and HCMUS_RepeatedGames show that direct supervision, even when combined only with relatively simple augmentation strategies, can yield highly competitive performance.

The promise and current limits of knowledge-enhanced reasoning. remains a promising but comparatively less mature direction. Several teams explored knowledge graphs, causal graphs, neuro-symbolic verification, or theorem-proving-style reasoning, but these approaches did not consistently outperform strong retrieval-plus-verification or fine-tuning baselines. This suggests that explicit structure is useful only when it materially improves evidence grounding; otherwise, it may simply introduce an additional layer of modeling complexity.

6.5 Discussion

The shared task suggests three main conclusions. First, AER is fundamentally constrained by evidence selection: systems must identify evidence for a *direct* cause before effective reasoning is possible. Second, leaderboard performance reveals a trade-off between expressive generation and calibrated decision making; systems with explicit verification and conservative prediction often proved more reliable than unconstrained generative approaches. Third, the strongest submissions treated AER as a pipeline task, decomposing it into retrieval, evidence filtering, candidate verification, and constrained answer selection.

Overall, the results suggest that progress on AER depends less on model scale alone than on evidence grounding, robustness to distractors, and careful calibration under multi-label evaluation.

7 Conclusion

This paper introduced SemEval-2026 Task 12, Abductive Event Reasoning (AER), a shared task for direct-cause inference over noisy multi-document evidence. We presented the task formulation, dataset construction pipeline, evaluation protocol, pilot baseline analysis, and an overview of submitted systems and leaderboard results.

The shared task results show that strong performance depends not only on model capacity, but also on effective evidence selection, distractor filtering, and calibrated multi-label prediction. The strongest systems consistently treated AER as an evidence-grounded reasoning pipeline rather than as standard multiple-choice classification.

We hope AER will serve as a useful benchmark for future research on causal reasoning and multi-document understanding.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. U24A20335, No. 62406321). This work is also supported by Beijing Natural Science Foundation (L243006).

References

- Anthropic. 2025. Claude 3.7 sonnet system card. Accessed: 2026-03-13.
- Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. [LongBench: A bilingual, multi-task benchmark for long context understanding](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137, Bangkok, Thailand. Association for Computational Linguistics.
- 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*.
- Pengfei Cao, Xinyu Zuo, Yubo Chen, Kang Liu, Jun Zhao, and Wei Bi. 2021a. Uncertainty-aware self-training for semi-supervised event temporal relation extraction. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 2900–2904.
- Pengfei Cao, Xinyu Zuo, Yubo Chen, Kang Liu, Jun Zhao, Yuguang Chen, and Weihua Peng. 2021b. Knowledge-enriched event causality identification via latent structure induction networks. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)*, pages 4862–4872.
- Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.
- Gemini Team and Google. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Cites the Gemini 2.X family including Gemini 2.0 Flash. Accessed: 2026-03-13.
- Joseph Y Halpern. 2016. *Actual causality*. MIT Press.
- Yupu Hao, Pengfei Cao, Yubo Chen, Kang Liu, Jiexin Xu, Huaijun Li, Xiaojian Jiang, and Jun Zhao. 2023. Complex event schema induction with knowledge-enriched diffusion model. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4809–4825.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.
- Zhitao He, Pengfei Cao, Zhuoran Jin, Yubo Chen, Kang Liu, Zhiqiang Zhang, Mengshu Sun, and Jun Zhao. 2024. Zero-shot cross-lingual document-level event causality identification with heterogeneous graph contrastive transfer learning. In *Proceedings of the 2024 Joint International Conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)*, pages 17833–17850.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.
- Klaus Krippendorff. 2011. Computing krippendorff’s alpha-reliability.
- Klaus Krippendorff. 2022. *Content Analysis: An Introduction to Its Methodology*, 4 edition. SAGE Publications.
- Cheng Liu, Wei Xiang, and Bang Wang. 2024. [Identifying while learning for document event causality identification](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3815–3827, Bangkok, Thailand. Association for Computational Linguistics.
- Qingkai Min, Zitian Qu, Qipeng Guo, Xiangkun Hu, Zheng Zhang, and Yue Zhang. 2025. [Multi-document event extraction using large and small language models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19265–19296, Suzhou, China. Association for Computational Linguistics.
- Paramita Mirza. 2014. [Extracting temporal and causal relations between events](#). In *Proceedings of the ACL 2014 Student Research Workshop*, pages 10–17, Baltimore, Maryland, USA. Association for Computational Linguistics.
- Paramita Mirza and Sara Tonelli. 2014. [An analysis of causality between events and its relation to temporal information](#). In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2097–2106, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.
- OpenAI. 2023. [Gpt-4 technical report](#). *arXiv preprint arXiv:2303.08774*.
- OpenAI. 2025a. Introducing gpt-4.1 in the api. Accessed: 2026-03-13.
- OpenAI. 2025b. Openai gpt-4.5 system card. Accessed: 2026-03-13.
- Judea Pearl. 2009. *Causality: Models, Reasoning, and Inference*, 2 edition. Cambridge University Press.

- Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. [XCOPA: A multilingual dataset for causal common-sense reasoning](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.
- Qwen Team. 2025. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2412.15115*.
- Angelika Romanou, Syrielle Montariol, Debjit Paul, Leo Laugier, Karl Aberer, and Antoine Bosselut. 2023. [CRAB: Assessing the strength of causal relationships between real-world events](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15198–15216, Singapore. Association for Computational Linguistics.
- Dianbo Sui, Xiangrong Zeng, Yubo Chen, Kang Liu, and Jun Zhao. 2023. Joint entity and relation extraction with set prediction networks. *IEEE transactions on neural networks and learning systems*, 35(9):12784–12795.
- Jialong Tang, Hongyu Lin, Meng Liao, Yaojie Lu, Xi-pei Han, Le Sun, Weijian Xie, and Jin Xu. 2021. [From discourse to narrative: Knowledge projection for event relation extraction](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 732–742, Online. Association for Computational Linguistics.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.
- Ming Tu, Guangtao Wang, Jing Huang, Yun Tang, Xiaodong He, and Bowen Zhou. 2019. [Multi-hop reading comprehension across multiple documents by reasoning over heterogeneous graphs](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2704–2713, Florence, Italy. Association for Computational Linguistics.
- Michael R Waldmann and York Hagmayer. 2013. Causal reasoning.
- Douglas Walton. 2014. *Abductive reasoning*. University of Alabama Press.
- Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022. [MAVEN-ERE: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 926–941, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.
- Xiaosong Yuan, Renchu Guan, Wanli Zuo, and Yijia Zhang. 2023. [The causal strength bank: A new benchmark for causal strength classification](#). In Hisashi Kashima, Tsuyoshi Ide, and Wen-Chih Peng, editors, *Advances in Knowledge Discovery and Data Mining: 27th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2023, Osaka, Japan, May 25–28, 2023, Proceedings, Part I*, volume 13935 of *Lecture Notes in Computer Science*, pages 111–122. Springer, Cham.
- Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadao Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Jingyu Sun, and 38 others. 2024. [Chatglm: A family of large language models from glm-130b to glm-4 all tools](#). *arXiv preprint arXiv:2406.12793*.
- Kun Zhao, Donghong Ji, Fazhi He, Yijiang Liu, and Yafeng Ren. 2021. [Document-level event causality identification via graph inference mechanism](#). *Information Sciences*, 561:115–129.
- Xinyu Zuo, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Weihua Peng, and Yuguang Chen. 2021. Improving event causality identification via self-supervised representation learning on external causal statement. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2162–2172.