BBgame at SemEval-2026 Task 12: Small Language Model Fine-tuning for Abductive Event Reasoning Task

Shu Li[1] **Huizhi Liang**[1]

[1]School of Computing, Newcastle University
Newcastle upon Tyne, UK
{c1041562, huizhi.liang}@newcastle.ac.uk

Abstract

We introduce a three-stage training framework for abductive event reasoning (AER). The task dataset was decomposed into 3 subsets, including causal judgment, cause generation, and multiple-choice answering (MCQA). Abductive reasoning requires understanding complex causal relationships between events. However, small language models typically struggle due to the multi-step inference required. Our approach combines supervised fine-tuning with group relative policy optimization (GRPO) to enhance the reasoning capabilities based on a 0.5B-parameter model. On the SemEval-2026 Task 12 development set, our Causal-Qwen-0.5B model achieves 64.75%, an absolute improvement of 55 percentage points over the native version baseline at 9.75%. Our ablation study reveals that binary causal judgment rather than cause generation or direct MCQA training is the key skill for AER task, with more complex stages significantly underperforming due to the task misalignment or task complexity.

1 Introduction

Understanding real world events why they occur is a fundamental aspect for natural language understanding. Everyday real world events unfold as the result of complex causal chains that often implicitly distributed across vast sources. This brings difficulty to reconstruct from single sourced evidence. The Abductive Event Reasoning(AER) is essential for identifying the causes behind the events. Utilizing contextual information and built-in knowledge inside the language model, to unveil the most plausible direct cause of the real world event. Currently, most advanced Large Language models tend to select answers that are semantically related to an event rather than causally responsible for it. In this case, it will be necessary to investigate the circumstances under real world textual evidence.

Specifically, given a specific event for AER task investigation in language models. Semeval2026-Task12 frames the task in the structured format. The input comprises a target event stated in natural language, alongside a contextual evidence set containing both relevant and irrelevant documents. The objective of the language model is to select the most plausible cause from four competing natural language explanations A to D. The final target is to identify the option which is causally responsible rather than merely semantically associated.

Table 1 presents a simplified example from the AER dataset. The target event describes Iran's missile attack. Four candidate causes are provided, ranging from background context to the direct causal trigger. Abbreviated documents provide supporting evidence for reasoning.

The most fundamental challenge is that the transformer decoder architecture for Large language models captures the correlations within sequential training data but does not inherently distinguish correlation from causation logically. This research also formalized this by defining two capability levels. Level-1 is for recalling memorized causal patterns. Level-2 is designed for genuine causal reasoning on novel problems. Most LLMs achieve only Level-1, showing significant performance drops on the CausalProbe 2024 benchmark when tested on fresh causal problems(Chi et al., 2025). This means strong benchmark performance may reflect memorization rather than reasoning. Causal hallucination compounds this problem. LLMs generate plausible sounding but incorrect causal explanations(Kiciman et al., 2023), and the black-box nature makes it difficult to trace the logical chain behind generated content. Top-performing approaches for causal and abductive reasoning overwhelmingly rely on proprietary models. Problems were identified on previous shared SemEval shared tasks. First, proprietary model versions change over time, making results non-reproducible. Second, model size was the dominant performance factor (Jullien et al., 2023). Increasing the number

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 894–898

Target Event	Iran launched ballistic missile attacks against Al Asad and Erbil air bases in Iraq used by US and coalition forces.
Context (excerpts from 20 retrieved documents)	
Doc 13	"Qassem Soleimani, head of Iran's elite IRGC Quds Force, was killed in a pre-dawn US air raid at Baghdad's international airport..."
Options	
A ✗	Dec 29 US airstrikes at Kataib Hezbollah facilities
B ✗	Muhandis founded Kataib Hezbollah after 2006
C ✗	Dec 27 Kataib Hezbollah attacked K1 base near Kirkuk
D ✓	US drone strike killed Gen. Soleimani and al-Muhandis
Answer	**D**

Table 1: Simplified example from the AER dataset. The target event describes Iran's missile attack on US bases in Iraq. Four candidate causes are provided, ranging from distant background context to the direct causal trigger. Document excerpts illustrate the textual evidence available for reasoning.

of model parameters leads to a direct increase in performance is far more significant than the effect of domain-specific training.

Research Questions:

1. How can we effectively train small language models for abductive causal reasoning through multi-stage supervised fine-tuning and reinforcement learning?
2. How can we decompose the abductive event reasoning task into learnable sub-tasks that small models can master through progressive training?

Contributions:

We introduce a progressive three-stage training framework that decomposes abductive reasoning into causal judgment, cause generation, and multiple-choice question answering, enabling small models to perform AER tasks.

2 Related Works

The αNLI dataset established foundational benchmarks for abductive reasoning in 2020 (Bhagavatula et al., 2020). The current task extends this from commonsense stories to real-world news events. E-CARE introduced an explainable causal reasoning dataset with 21K QA pairs, however the contextual document is absent (Du et al., 2022). Existing benchmarks focus on commonsense-level abduction or lack document-grounded contextual information. The SemEval-2026 Task12 uniquely requires reasoning over retrieved documents about real-world news events (Cao et al., 2026).

Existing research indicates that current LLMs often fall short of genuine causal understanding. An increasing number of studies have examined the causal reasoning abilities of LLMs. For instance, Kiciman et al. (2023) performed an extensive behavioral study showing that while LLMs can generate text corresponding to correct causal arguments with high probability, they display unpredictable failure modes, especially in distinguishing causation from mere correlation. CausalProbe-2024 is a benchmark constructed from recent news to test whether LLMs are really using causal reasoning or just rely on superficial pattern matching from training data Chi et al. (2024). Their results reveal significant performance degradation when LLMs encounter novel causal scenarios, suggesting that much of their apparent causal ability stems from memorized associations rather than reasoning.

Additional evidence is provided by Miliani et al. (2025), who demonstrate that LLMs often confuse temporal precedence with causality. This finding is directly relevant to our task, where candidate options include temporally prior events that contributed to an escalation chain but are not the direct cause. The survey by Li et al. (2025) provides a comprehensive overview of methods for enhancing LLM causal reasoning, including retrieval-augmented generation and structured prompting strategies. Together, these studies motivate our approach: we explicitly split our dataset into three subsets to train a small language model to distinguish direct causes from background events through a staged fine-tuning framework.

Our work applied group relative policy optimization (GRPO) to abductive causal reasoning, a domain where reward signals are less straightforward than in mathematical verification. We adopt Qwen2.5-0.5B-Instruct (Qwen et al., 2024) as our base model and apply LoRA (Hu et al., 2022) for parameter-efficient adaptation. Crucially, we de-

compose the reasoning task into three progressive stages: causal judgment, cause generation, and MCQA, each trained with supervised fine-tuning followed by GRPO-based refinement.

3 Task and Dataset

SemEval-2026 Task 12 frames AER as a multiple-choice question answering problem. Each instance consists of three components. A target event described as a short natural language sentence. Context comprising a set of retrieved documents related to the event, which may include both causally relevant and distractor documents. Candidate options, each presenting a natural-language explanation for why the event occurred. The task is then reformulated as a multi-task classification problem among the four options. The model must output the correct options based on reasoning over the provided context.

We further split the dataset into three stages, which decomposes the abductive event reasoning into causal judgment, cause generation, and multiple-choice question answering, decoupling the AER task into judgement, causal reasoning and causal inference subproblems.

3.1 Subset 1: Causal Judgment

The causal judgment task is defined as, given a target event and a single candidate cause, predict whether the candidate is a plausible direct cause. We expand the original 1,819 questions into 7,276 binary classification instances as below:

> **User:** Target Event: `{event}`
> Candidate Cause: `{option}`
> Is this a direct and plausible cause? Answer Yes or No.
>
> **Assistant:** Yes / No

3.2 Subset 2: Cause Generation

We define Subset 2 as, given a target event, generate a plausible cause in free text. We hypothesized that learning to generate causal explanations would improve understanding of causal structure, which could transfer to MCQA performance.

3.3 Subset 3: MCQA

In this subset, we directly predict answer label from four options.

Dataset	Transformation	Size	Format
Subset 1	Each question 4 binary pairs	7,276	event, single option
Subset 2	Each question generation target(s)	1,819	event, cause text
Subset 3	Original format preserved	1,819	event, 4 options

Table 2: Training data construction for each stage.

4 Methodology

Small language models typically struggle with abductive reasoning. We hypothesize that decomposing the task into learnable sub components, consists of causal judgment, cause generation, and final prediction. This decomposition enables effective learning even at the 0.5B-parameter scale. Our approach, illustrated in Figure 1, consists of three training stages. First, we train a binary causal judgment classifier to distinguish valid from invalid causal relationships. This model is then refined through GRPO. We report the best-performing stage as our main methodology.

Our main framework employs Qwen2.5-0.5B-Instruct as the foundation model. For each stage, we first apply supervised fine-tuning with LoRA adapters for 5 epochs using the AdamW optimizer (learning rate 2×10^{-5}, batch size 8, LoRA rank $r = 16$, scaling factor $\alpha = 32$). We then apply Group Relative Policy Optimization for 2 additional epochs with a reduced learning rate (1×10^{-5}, batch size 2–4, KL penalty coefficient $\beta = 0.1$). Training uses the three subsets derived from the SemEval-2026 Task 12 training split, with hyperparameter tuning performed on the development set. All three stages share a chat-formatted template combining a system message that defines the causal-reasoning role with a user message whose content varies by stage. Stage 1 supplies the target event e paired with a single candidate cause c and asks for a Yes/No judgment. Stage 2 supplies e together with the retrieved context and asks the model to generate a plausible cause in free text. Stage 3 supplies e, the retrieved context, and the four options $\{A, B, C, D\}$, asking the model to select the correct letter(s). The maximum sequence length is set to 1024 tokens to accommodate the context length of the dataset. All training is conducted on a single NVIDIA RTX 3090.

We also evaluated logistic regression, the API-based LLM DeepSeek-Chat-V3.2, the T5-based QA model UnifiedQA-t5-small, and Qwen2.5-0.5B

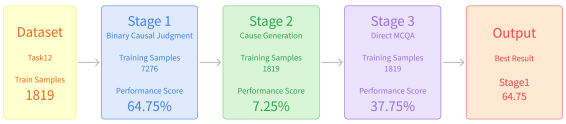

Figure 1: The split datasets for the three stages of fine-tuning.

without fine-tuning as baselines.

5 Experiments

Method	Type	Dev	Test
UnifiedQA	Zero-shot	0.0625	–
Qwen2.5-0.5B	Local LLM	0.0975	–
Logistic Regression	ML	0.5549	–
DeepSeek-V3.2	API-based	0.5713	–
Causal-Qwen-0.5B (Ours)	**Staged**	**0.6475**	**0.5800**

Table 3: Comparison with baselines on the SemEval-2026 Task 12 dev and test set. Score: 1.0 exact, 0.5 partial, 0.0 incorrect. Test scores are from the official CodaLab leaderboard; dashes indicate runs not submitted to the test server before its closure.

As shown in Table 3, our staged training approach achieves 64.75 percentage-point on the development set, representing a 7.62 percentage improvement over the best API-based baseline DeepSeek-V3.2 and a 9.26 percentage-point improvement over logistic regression with multimodal embeddings at 55.49 percentage. Compared with traditional zero-shot generative model UnifiedQA-t5-small, which achieves 6.25 percentage.

5.1 Ablation Study

Table 4 presents ablation results across our three-stage framework. The subset 1 causal judgment achieves the best performance (64.75% with GRPO), significantly outperforming the more complex Stage 3 MCQA training (37.75%). This suggests that binary causal discrimination is the core skill for this task, and that direct MCQA training may actually harm performance by encouraging overfitting to spurious patterns.

5.2 Error Analysis

We analyzed errors made by the Causal-Qwen model on the development set. Error types fall into two categories: correct subset predictions representing partial understanding where the model misses one option in multi-label cases, and incorrect predictions representing complete failure.

Stage	Method	Dev	Test
Qwen2.5-0.5B	None	0.0975	–
S1: Causal Judgment	SFT	0.6338	–
S1: Causal Judgment	GRPO	**0.6475**	0.5800
S2: Cause Generation	SFT	0.0838	–
S2: Cause Generation	GRPO	0.0725	–
S3: MCQA	SFT	0.3337	–
S3: MCQA	GRPO	0.3775	–
Fusion Subsets	SFT	0.4675	–

Table 4: Trainable approaches across stages on the development and test sets.

Common causes of incorrect predictions include ambiguous causal chains where multiple plausible causes exist, background context mistaken for direct cause, and temporal reasoning failures where the model misidentifies causal order.

For example, given the target event "Videos of the assassination circulated on social media," the gold answer is D (*A man fired twice at Shinzo Abe*), but the model predicted B (*Security arrested the suspected gunman*). This error indicates that the model selected the subsequent event rather than the preceding direct cause, revealing challenges with temporal reasoning that could be addressed in future work.

6 Conclusion

We introduced a three-stage training framework for abductive event reasoning that achieves 64.75% on the SemEval-2026 Task 12 development set using a 0.5B parameter model. Our key finding is that binary causal judgment serves as the most effective formulation for the AER task at small language model scale.

The surprising failure of Stage 2 cause generation and the underperformance of Stage 3 direct MCQA reveal important insights. Task alignment between training objective and evaluation metric is critical. Text generation optimizes for fluency rather than precise causal discrimination. Simpler formulations with strong inductive biases can outperform complex end-to-end approaches refer to the Stage 3 result. Multi-label complexity in MCQA introduces challenges for 0.5B scale models.

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*.

Pengfei Cao, Yubo Chen, Mingxuan Yang, 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.

Haoang Chi, He Li, Wenjing Yang, Feng Liu, Long Lan, Xiaoguang Ren, Tongliang Liu, and Bo Han. 2024. Unveiling causal reasoning in large language models: Reality or mirage? In *Advances in Neural Information Processing Systems*.

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

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-care: a new dataset for exploring explainable causal reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–446, Dublin, Ireland. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Ma"el Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Donal Landers, and André Freitas. 2023. Semeval-2023 task 7: Multi-evidence natural language inference for clinical trial data. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2216–2226, Toronto, Canada. Association for Computational Linguistics.

Emre Kiciman, Robert Osazuwa Ness, Amit Sharma, and Chenhao Tan. 2023. Causal reasoning and large language models: Opening a new frontier for causality. *Preprint*, arXiv:2305.00050.

Xin Li, Zhuo Cai, Shoujin Wang, Kun Yu, and Fang Chen. 2025. A survey on enhancing causal reasoning ability of large language models. *Preprint*, arXiv:2503.09326.

Martina Miliani, Serena Auriemma, Alessandro Bondielli, Emmanuele Chersoni, Lucia Passaro, Irene Sucameli, and Alessandro Lenci. 2025. Explica: Evaluating explicit causal reasoning in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17335–17355, Vienna, Austria. Association for Computational Linguistics.

Qwen and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.