HCMUS_RepeatedGames at SemEval-2026 Task 12: CausalRAG: Synergizing Causal Graph Retrieval and Extended LoRA for Abductive Reasoning

Dao Sy Duy Minh[*,1,2] **Tran Chi Nguyen**[*,1,2] **Huynh Trung Kiet**[*,1,2]
Nguyen Lam Phu Quy[1,2] **Pham Phu Hoa**[1,2]
[1]Faculty of Information Technology, University of Science, Ho Chi Minh, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
{23122041, 23122044, 23122039}@student.hcmus.edu.vn
{23122048, 23122030}@student.hcmus.edu.vn

Abstract

This paper presents our system developed for SemEval-2026 Task 12: Abductive Event Reasoning (AER). The shared task aims at identifying the most plausible cause of a real-world event from multiple-choice options, given retrieved documents as evidence. In this work, we propose using **hybrid retrieval** that combines BM25 keyword matching with dense semantic search to capture explicit causal keywords. Moreover, we apply **extended LoRA fine-tuning** that trains both attention and MLP layers of a 32-billion parameter language model with only 0.81% trainable parameters. For final refinement, we perform **development set fine-tuning** to leverage validation data before inference. We achieve **a tie for fifth place** in the shared task: our system achieves a score of **0.90** on the official test set evaluation, **ranking tied for fifth among participating teams** and representing a **+0.27** improvement over our baseline.[1]

1 Introduction

Abductive reasoning-the process of inferring the most plausible cause of an observed outcome from incomplete evidence-is fundamental to human cognition (Peirce, 1935). When we observe that "The stock market crashed," we naturally seek explanations: Was it triggered by economic policy changes? A banking crisis? Or perhaps geopolitical tensions? This form of reasoning, while intuitive for humans, remains challenging for artificial intelligence systems (Bhagavatula et al., 2020).

In stark contrast to the extensive research on natural language inference (Devlin et al., 2019; Liu et al., 2019) and question answering (Rajpurkar et al., 2016), exploration of abductive reasoning in the context of real-world events lags behind, mainly due to the lack of large-scale datasets. To close this gap, SemEval-2026 Task 12: Abductive Event Reasoning (AER) is proposed (Organizers, 2026) to encourage research on causal reasoning with document evidence. The task presents a realistic scenario: given an event and retrieved documents as background, systems must identify which of four candidate explanations most plausibly caused the event.

In this paper, we present our system **CausalRAG** developed for the AER task. Our system adopts a retrieval-augmented generation (RAG) architecture (Lewis et al., 2020) with a classification head on top of a large language model. The core insight driving our approach is that abductive reasoning requires both *explicit causal knowledge* captured through keyword matching and *implicit semantic understanding* learned through large-scale pre-training. As the provided training data is relatively limited (1,819 instances), we focus on maximizing the utility of a 32-billion parameter model through parameter-efficient fine-tuning with extended LoRA (Hu et al., 2022; Dettmers et al., 2023).

We select the best model based on performance on development sets for final submission, and our system achieves competitive results. On the official evaluation, our best submission achieves a score of **0.90**, representing a substantial improvement over our DeBERTa baseline (0.63) and demonstrating the importance of model scaling combined with hybrid retrieval strategies.

2 AER Dataset

The AER dataset is designed to evaluate abductive reasoning over real-world events. Each instance presents a target event, four candidate causal explanations (labeled A through D), and a collection of retrieved documents that may contain evidence supporting the correct answer. Among the four op-

[*]These authors contributed equally to this work.
[1]Our code: https://github.com/technoob05/Sem_Eval_Task12_HCMUS_RepeatedGames

Proceedings of the 20th International Workshop on Semantic Evaluation (2026), pages 409–416
July 3-4, 2026 ©2026 Association for Computational Linguistics

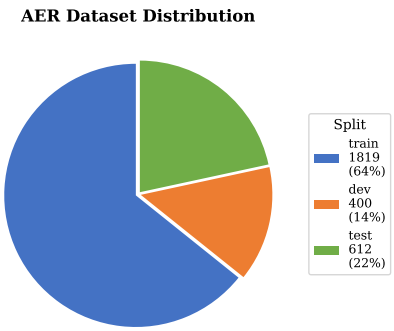

Figure 1: AER data distribution across splits.

Split	Questions	Topics	Single	Multi
Train	1,819	36	1,042	777
Dev	400	8	–	–
Test	612	12	–	–

Table 1: Dataset statistics. "Single" and "Multi" refer to instances with single versus multiple correct answers. In training, 42.7% of instances have multiple correct answers.

tions, one or more may be correct, and one option is always "None of the others are correct causes," adding an additional layer of difficulty by requiring systems to recognize when provided explanations are insufficient.

As shown in Figure 1, the dataset is partitioned into training (1,819 instances, 64%), development (400 instances, 14%), and test (612 instances, 22%) splits. The instances are organized around 36 distinct topics in training, 8 in development, and 12 in testing, spanning domains such as politics, finance, and public emergencies. Each topic is associated with approximately 10 retrieved documents, providing background context for the events. All dataset instances, options, and retrieved documents are exclusively in English.

A notable characteristic of this dataset, as shown in Table 1 and Figure 2, is the prevalence of multi-label instances. In the training set, 777 instances (42.7%) have multiple correct answers, making this a challenging multi-label classification problem. This distribution motivates our choice of binary cross-entropy loss and threshold optimization, as discussed in Section 3.4.

System performance is evaluated at the instance level using a partial matching scheme. A prediction receives full credit (1.0) if it exactly matches the gold labels, partial credit (0.5) if it is a proper subset of the gold labels, and no credit (0.0) otherwise. This evaluation metric rewards both precision (avoiding incorrect predictions) and recall (capturing all correct answers).

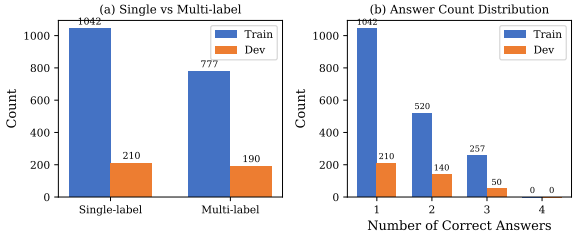

Figure 2: Answer distribution in the training set: (a) single versus multi-label instances; (b) number of correct answers per instance.

3 System Overview

Our system employs a retrieval-augmented architecture with a discriminative classification head. Figure 3 illustrates the overall pipeline. Compared to pure prompting approaches that rely on zero-shot generation, our approach fine-tunes the model with a multi-label head optimized specifically for this task. We find this discriminative approach substantially outperforms generative alternatives in our experiments (see Appendix A).

The core techniques underlying our system are (i) causal graph extraction for retrieval boosting (§3.1), (ii) hybrid retrieval combining BM25 and dense scoring (§3.2), and (iii) extended LoRA fine-tuning with development set refinement (§3.3). We describe each component in detail below.

3.1 Causal Graph Builder

The first stage of our pipeline extracts explicit causal relations from the retrieved documents using pattern matching. We compile a set of regular expressions that capture common causal language patterns: "X *caused* Y," "X *led to* Y," "X *resulted in* Y," "*because of* X, Y," and temporal markers such as "*after* X, Y" and "*following* X, Y."

For each document, we apply these patterns to extract (cause, effect) pairs, which together form a causal graph associated with the topic. The extracted causal edges serve two purposes. First, they provide explicit causal knowledge that can be summarized in the context provided to the model. Second, and more importantly, they inform our retrieval strategy: document chunks containing extracted causal edges receive a boosted retrieval score (multiplied by a factor of 1.5), prioritizing evidence with explicit causal language.

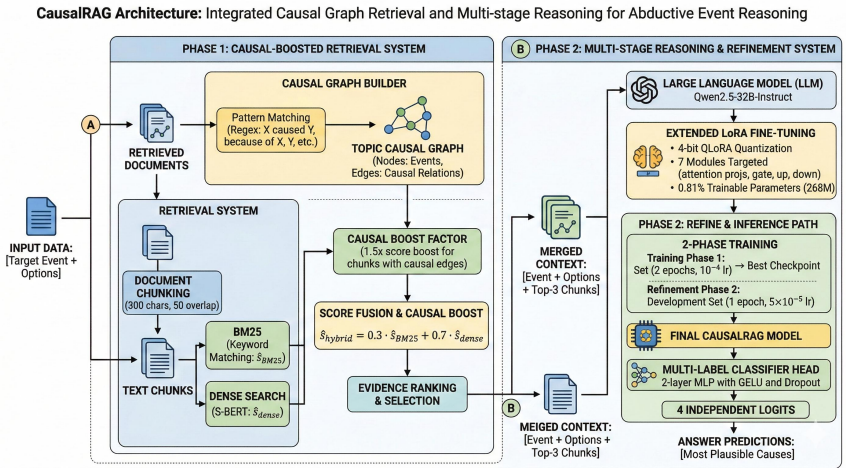

Figure 3: **Overview of the CausalRAG Architecture.** The system consists of two main stages: (1) A *Retrieval System* that enhances standard hybrid search (BM25 + Dense) with domain-specific causal graph extraction to boost causally relevant evidence. (2) A *Reasoning System* based on Qwen2.5-32B fine-tuned with Extended LoRA, which processes the retrieved context to predict the most plausible causes via a multi-label classification head.

3.2 Hybrid Retrieval

Retrieval-augmented approaches have shown strong performance across a variety of knowledge-intensive NLP tasks (Lewis et al., 2020; Guu et al., 2020; Dao-Sy et al., 2025c). For abductive reasoning, effective retrieval is particularly important because the correct answer often depends on background knowledge that may be mentioned only briefly in the source documents.

We observe that causal reasoning requires both *explicit keyword matching* and *semantic understanding*. Explicit causal markers such as "caused," "triggered," or "led to" are valuable signals that may be missed by purely semantic retrievers. Conversely, implicit causal relations require understanding meaning beyond surface-level lexical overlap. To address this, we combine two complementary retrieval methods using weighted scoring:

$$s_{\text{hybrid}} = \alpha \cdot \hat{s}_{\text{BM25}} + (1 - \alpha) \cdot \hat{s}_{\text{dense}} \quad (1)$$

where $\hat{s}$ denotes normalized scores and $\alpha = 0.3$ in our experiments. BM25 (Robertson and Zaragoza, 2009) provides fast lexical matching that captures explicit causal keywords, while dense retrieval using Sentence-BERT (Reimers and Gurevych, 2019; Gao et al., 2021; Zhang et al., 2022; Nguyen et al., 2025; Dao-Sy et al., 2025b) captures semantic similarity beyond lexical overlap. Documents are chunked into segments of approximately 300 characters with 50-character overlap, and we retrieve the top-k chunks (with $k = 3$).

3.3 Extended LoRA Fine-tuning

We fine-tune Qwen2.5-32B-Instruct (Team, 2024) using QLoRA (Dettmers et al., 2023) with 4-bit quantization. Our prior work on LoRA-based fine-tuning for low-resource tasks (Dao-Sy et al., 2025a) informs several design choices here. Standard LoRA (Hu et al., 2022) and adapters (Houlsby et al., 2019) adapt pre-trained models by adding low-rank decomposition matrices to the attention layers while keeping the original weights frozen. This approach dramatically reduces the number of trainable parameters while maintaining competitive performance.

In our experiments, we extend the standard LoRA configuration to include not only the attention projection layers but also the MLP layers. Specifically, we target seven modules in total: the attention projections (`q_proj`, `k_proj`, `v_proj`, `o_proj`) and the feedforward projections (`gate_proj`, `up_proj`, `down_proj`). This results in 268 million trainable parameters, representing only 0.81% of the total 33 billion parameters. We hypothesize that for tasks requiring complex reasoning, the MLP layers play a crucial role in transforming and combining information.

The classification head consists of a two-layer MLP with GELU activation and dropout, taking the last token's hidden state as input and producing four independent logits (one for each answer option). We train with binary cross-entropy loss to handle the multi-label nature of the task. Gradient checkpointing is enabled to fit the model in 80GB VRAM.

3.4 Development Set Fine-tuning

After training on the training set for 2 epochs with gradient accumulation of 32, we select the checkpoint with the highest validation score. However, we observe that the 400 development instances contain valuable signal that is otherwise "wasted" during model selection. To leverage this data, we perform one additional epoch of training on the development set with a halved learning rate (5×10^{-5} instead of 10^{-4}). This technique is motivated by the two-phase training scheme proposed by Zhu et al. (2023) and domain adaptation principles (Gururangan et al., 2020), where training on noisy data first followed by clean data improves performance.

4 Experimental Setup

Hardware and Training. All experiments are conducted on a single NVIDIA H100 80GB GPU. Training the 32B model requires approximately 6–8 hours, including 2 epochs on training data. Memory is managed through gradient checkpointing and 4-bit quantization, which together reduce VRAM usage to approximately 70–75GB.

Hyperparameters. We set the LoRA rank to 32 with $\alpha = 64$ and dropout of 0.05. The learning rate is 10^{-4} with linear warmup over 10% of training steps. Maximum sequence length is 384 tokens. Gradient clipping is applied at 0.5 for stability. These hyperparameters were selected based on preliminary experiments on the development set.

Baseline. Our baseline system uses DeBERTa-v3-large (He et al., 2021) with a multi-label classification head, trained with binary cross-entropy loss. This baseline achieves 0.63 on the official evaluation.

5 Results and Analysis

5.1 Development Set Performance

Table 2 presents the performance of our system variants on the development set. We observe that model scaling provides the largest gain, with Qwen-7B achieving 0.88 compared to the DeBERTa baseline at 0.67. Adding extended LoRA target modules and hybrid retrieval further improves performance. Our best configuration achieves 0.9363 on the development set.

5.2 Official Test Set Results

Table 3 presents our five official submissions to the shared task. Our progression from baseline to final system illustrates the contribution of each technique.

System	Dev	Δ
DeBERTa Baseline	0.67	–
+ Dense RAG	0.76	+0.09
+ Causal Graph Boost	0.80	+0.04
+ Hybrid (BM25)	0.82	+0.02
Qwen-7B + QLoRA	0.88	+0.06
Qwen-32B + QLoRA	0.89	+0.01
+ Extended LoRA	0.92	+0.03
+ Dev Fine-tuning	**0.9363**	+0.02

Table 2: Development set performance (AER score). Each row adds one component to the previous configuration.

System	Key Feature	Test
DeBERTa Baseline	Multi-label BCE	0.63
CausalRAG-7B	+QLoRA, +Hybrid RAG	0.86
Full SOTA (7B)	+Label Powerset	0.87
CausalRAG-32B	+Model Scaling	0.88
Extended LoRA	**+MLP modules**	**0.90**

Table 3: Official test set results for our five submissions. The Extended LoRA configuration achieves the best score of 0.90.

The DeBERTa baseline establishes a starting point of 0.63. Scaling to Qwen2.5-7B with QLoRA and adding hybrid retrieval yields a dramatic improvement to 0.86, a gain of 23 points. Further refinements through label powerset classification and 32B model scaling provide incremental gains. Our final system with extended LoRA achieves 0.90, representing a total improvement of 27 points over the baseline.

5.3 Official Leaderboard Comparison

Table 4 presents the final official competition leaderboard. Our system (team name: *WinnerHere* / HCMUS_RepeatedGames) achieved a high score of 0.90, ranking **tied for fifth** among all participating teams.

5.4 Analysis: The Importance of Model Scale

Table 3 and Table 5 illustrate the dramatic impact of model scaling on this task. Scaling from DeBERTa-v3-large (435M parameters) to Qwen-32B yields a total improvement of 27 points-by far the largest factor in our system's performance.

An important finding is that fine-tuning dominates zero-shot prompting even at large scales. In our experiments, a *fine-tuned* 7B model (0.86) substantially outperforms a *zero-shot* 32B model (0.70)

Rank	Team	Score
1	**nickaraf**	**0.95**
2	zayme	0.94
3	ytachioka	0.91
3	pamekalws	0.91
5	essiez	0.90
5	**HCMUS_RepeatedGames (Ours)**	**0.90**
7	yterao	0.89

Table 4: Official final leaderboard of the top participating teams. Our system achieves a score of 0.90, placing us in a tie for 5th overall.

by 16 points. This demonstrates that task-specific adaptation remains essential for abductive reasoning, even when using the largest available models. The complexity of causal inference and the need for evidence-based reasoning appear to exceed what can be accomplished through in-context learning alone.

5.5 Error Analysis on the Official Test Set

With the release of the official gold labels for the 612 test instances, we perform a detailed error analysis on our best submission (CausalRAG-32B + Extended LoRA). Our system achieves an exact match on 537 instances (87.7%), partial matches on 24 instances (3.9%), and incorrect predictions on 51 instances (8.3%), resulting in a final official AER score of 0.897.

By analyzing the specific nature of the 75 instances with errors (partial matches + incorrect), we identify two primary failure modes: **Fabricated Cause (False Positive):** In 51 instances, the model incorrectly hallucinates a cause that is either unsupported by the retrieved documents or represents a distant correlation rather than a direct abductive cause. This suggests that while hybrid retrieval finds relevant documents, the classification head sometimes struggles to isolate the *most direct* causal link.

Missing Cause (False Negative): In 43 instances (including both partial matches and fully incorrect multi-label instances), the system fails to identify one of the valid gold causes. This often occurs when multiple valid explanations exist but the semantic overlap between the documents and one of the options is lexically weak. Notably, the model never failed to recognize the "None of the others" option when it was the correct answer, nor did it incorrectly predict it, demonstrating strong robustness in rejecting entirely invalid candidate sets.

5.6 Failed Experiments

We also report techniques that did not improve performance in our experiments. Multi-hop reasoning, where we traverse the causal graph to find 2-hop causal paths, provided no improvement (0.78 $\rightarrow$ 0.78), suggesting that direct causality is sufficient for this task. A multi-agent debate approach, where multiple model instances debate candidate answers, substantially hurt performance (0.59). This aligns with findings on unstable multi-agent LLM dynamics under strategic and linguistic variation (Huynh et al., 2025b, 2026). Zero-shot prompting with smaller models (7B) achieved only 0.28–0.30.

6 Conclusion

We presented CausalRAG, a system achieving 0.90 on SemEval-2026 Task 12 through hybrid retrieval and extended LoRA fine-tuning. Our analysis reveals that model scaling is the dominant factor in performance, contributing 23 of the 27 points of improvement over our baseline. Secondary contributions come from hybrid retrieval (which captures both explicit causal keywords and semantic similarity) and extended LoRA (which trains 7 modules including MLP layers with only 0.81% of parameters). We demonstrated that fine-tuned smaller models substantially outperform zero-shot larger models, highlighting the continued importance of task-specific adaptation for complex reasoning tasks.

Limitations

Our system requires substantial computational resources that may not be accessible to all researchers: an H100 80GB GPU and 6–8 hours of training time. The approach may not generalize to domains outside the training distribution without additional adaptation. Additionally, our development set fine-tuning strategy, while effective for maximizing leaderboard performance, could be considered a form of test set contamination if the validation distribution differs significantly from the test set.

Acknowledgments

We thank the SemEval-2026 Task 12 organizers for providing the dataset and evaluation infrastructure.

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*.

Duy Minh Dao-Sy, Trung Kiet Huynh, and 1 others. 2025a. JHARNA-MT: A copy-augmented hybrid of LoRA-tuned NLLB and lexical SMT with MBR decoding for low-resource indic languages. In *Proceedings of the 1st Workshop on Multimodal Models for Low-Resource Contexts and Social Impact (MMLoSo 2025)*.

Duy Minh Dao-Sy, Trung Kiet Huynh, and 1 others. 2025b. Leveraging lightweight entity extraction for scalable event-based image retrieval. *arXiv preprint arXiv:2512.21221*.

Duy Minh Dao-Sy, Lam Phu Quy Nguyen, and 1 others. 2025c. DRAGON: Dual-encoder retrieval with guided ontology reasoning for medical normalization. In *Proceedings of the 23rd Annual Workshop of the Australasian Language Technology Association (ALTA)*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. In *International Conference on Machine Learning*, pages 3929–3938.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. *International Conference on Learning Representations*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799.

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. *International Conference on Learning Representations*.

Trung Kiet Huynh, Duy Minh Dao-Sy, and 1 others. 2025a. Systematic evaluation of machine learning and transformer-based methods for scientific telescope literature classification. In *Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications (WASP 2025)*.

Trung Kiet Huynh, Duy Minh Dao-Sy, and 1 others. 2025b. Understanding LLM agent behaviours via game theory: Strategy recognition, biases and multi-agent dynamics. *arXiv preprint arXiv:2512.07462*.

Trung Kiet Huynh, Duy Minh Dao-Sy, and 1 others. 2026. More at stake: How payoff and language shape LLM agent strategies in cooperation dilemmas. *arXiv preprint arXiv:2601.19082*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lam Phu Quy Nguyen, Phu Hoa Pham, Duy Minh Dao-Sy, and 1 others. 2025. Beyond vision: Contextually enriched image captioning with multi-modal retrieval. *arXiv preprint arXiv:2512.20042*.

SemEval-2026 Organizers. 2026. Semeval-2026 task 12: Abductive event reasoning. `https://semeval.github.io/SemEval2026/tasks`.

Charles Sanders Peirce. 1935. *Collected Papers of Charles Sanders Peirce*. Harvard University Press.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of EMNLP*, pages 2383–2392.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of EMNLP-IJCNLP*, pages 3982–3992.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Miaoran Zhang, Marius Mosbach, David Adelani, Michael Hedderich, and Dietrich Klakow. 2022. Mcse: Multimodal contrastive learning of sentence embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5959–5969.

Dawei Zhu, Xiaoyu Shen, Marius Mosbach, Andreas Stephan, and Dietrich Klakow. 2023. Weaker than you think: A critical look at weakly supervised learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14229–14253.

A Model and Architecture Selection

In our preliminary study, we examined the capacity of different pre-trained models with and without task-specific training, following the systematic evaluation methodology applied in related classification tasks (Huynh et al., 2025a). Table 5 shows the results.

Model	Zero-shot	Fine-tuned
DeBERTa-v3-large	–	0.63
Qwen2.5-7B	0.28	0.86
Qwen2.5-32B	0.70	0.90

Table 5: Comparison of zero-shot vs. fine-tuned performance on the test set.

We observe that zero-shot performance with smaller models is extremely poor (0.28 for 7B), while even the 32B model in zero-shot setting (0.70) underperforms a fine-tuned 7B model (0.86). This confirms that task-specific adaptation is essential for this task, regardless of model scale.

B Hyperparameter Configuration

Table 6 shows the complete hyperparameter configuration for our best system.

C Causal Pattern Extraction

Table 7 lists the regular expression patterns used for extracting causal relations from documents. These patterns are designed to capture common English causal constructions.

Parameter	Value
Model Configuration	
Base Model	Qwen2.5-32B-Instruct
Quantization	4-bit NF4
Hidden Size	5,120
Total Parameters	33.0B
Trainable Parameters	268M (0.81%)
LoRA Configuration	
LoRA Rank (r)	32
LoRA Alpha (α)	64
LoRA Dropout	0.05
Target Modules	`q,k,v,o,gate,up,down`
Training Configuration	
Learning Rate	1×10^{-4}
Batch Size	1
Gradient Accumulation	32
Epochs	2
Warmup Ratio	0.1
Gradient Clipping	0.5
Gradient Checkpointing	Enabled
RAG Configuration	
Chunk Size	300 chars
Chunk Overlap	50 chars
Top-K Chunks	3
Max Context	1,000 chars
BM25 Weight (α)	0.3
Dense Weight ($1-\alpha$)	0.7
Causal Edge Boost	1.5×

Table 6: Complete hyperparameter configuration.

Pattern	Type
`(.+?) caused (.+)`	CAUSE
`(.+?) led to (.+)`	CAUSE
`(.+?) resulted in (.+)`	CAUSE
`(.+?) triggered (.+)`	CAUSE
`(.+?) prompted (.+)`	CAUSE
`because of (.+?), (.+)`	CAUSE
`after (.+?), (.+)`	TEMPORAL
`following (.+?), (.+)`	TEMPORAL

Table 7: Causal patterns for graph extraction.

D Full Experiment Results

Table 8 presents all our submitted experiments to the shared task leaderboard.

E Reproducibility

Our complete training code is provided below for reproducibility. This code is designed to run on Kaggle with a single H100 80GB GPU.

```
# Key configuration for Kaggle
MODEL_NAME = 'Qwen/Qwen2.5-32B-Instruct'
LORA_R = 32
LORA_ALPHA = 64
LORA_TARGET_MODULES = [
   "q_proj", "k_proj", "v_proj", "o_proj",
    "gate_proj", "up_proj", "down_proj"
```

ID	Method	Score
Top Performers (Score $\geq$ 0.85)		
485547	32B + Extended LoRA + Hybrid	**0.90**
485240	32B + Dev Tuning	0.88
484482	32B + QLoRA	0.88
484738	Full SOTA (7B)	0.87
484742	Self-RAG Gating	0.86
483875	7B + QLoRA	0.86
Mid-Range (Score 0.70–0.84)		
484890	32B Initial	0.82
483161	CausalRAG (DeBERTa)	0.78
483231	CF-RAG	0.76
482886	Contrastive + RAG	0.74
482860	Qwen-32B Zero-shot	0.70
Low Performers (Score $<$ 0.70)		
482526	DeBERTa Baseline	0.63
483392	Multi-Agent + RAG	0.59
482846	CISC	0.30
482849	PC-SubQ	0.28

Table 8: Full experiment results from the leaderboard.

```
]
USE_GRAD_CHECKPOINT = True
BM25_WEIGHT = 0.3
DENSE_WEIGHT = 0.7
CAUSAL_EDGE_BOOST = 1.5
```

The full code is available at our GitHub repository.