KDW at SemEval-2026 Task 12: Logic-Driven Distillation with Knowledge Graphs for Efficient Abductive Reasoning

Xinyan Xu[1,*] **Sihan Zhu**[2,*] **Hongjie Wu**[3,*]
[1]University of Tübingen, Germany
[2]Department of Computer Science, National University of Singapore, Singapore
[3]Independent Researcher, United Kingdom
xinyan.xu@student.uni-tuebingen.de,
e1504056@u.nus.edu, wu.hongjie@outlook.com

Abstract

Large language models (LLMs) such as GPT-4 and Gemini show strong reasoning ability but incur substantial computational cost in abductive reasoning settings. We present our system for "SemEval-2026 Task 12 — Abductive Event Reasoning: Towards Real-World Event Causal Inference for Large Language Models", which integrates knowledge graph (KG) evidence extraction with knowledge distillation to transfer structured reasoning from a large teacher model to a compact student model. Our approach ranks 8th in the shared task while achieving performance comparable to frontier LLMs at a fraction of the inference cost. The code is publicly available [1].

1 Introduction

LLMs have achieved strong performance on question answering and reasoning benchmarks that assume complete information and a single correct answer (Rajpurkar et al., 2016; Clark et al., 2018; Srivastava, Aarohi and Rastogi, Abhinav and Rao, Abhishek and others, 2023). Abductive event reasoning in SemEval-2026 Task 12 presents a more challenging setting, requiring models to infer plausible causes from partial, noisy, and indirectly related evidence. The task provides extensive factual documents related to each query, requiring models to identify specific evidence from these long event-based texts to support their reasoning. Models must then discern the direct causes of a target event from a set of candidates, or correctly identify if "None of the others" (NOTA) applies.

In contrast to traditional LLM inference strategies such as prompt engineering and Chain-of-Thought (CoT), we treat abductive reasoning as structured plausibility evaluation for each candidate option, requiring explicit evidential support and robust decision-making capabilities. Accordingly, we leverage Knowledge Graph (KG) techniques to perform option-specific evidence retrieval and extraction. Simultaneously, we employ knowledge distillation to empower a discriminative lightweight model with stable and potent reasoning abilities, while applying a conservative selection strategy with explicit handling of the NOTA logic. Our evaluation demonstrates that the system achieves high accuracy while maintaining significantly low inference costs.

2 Background and Related Work

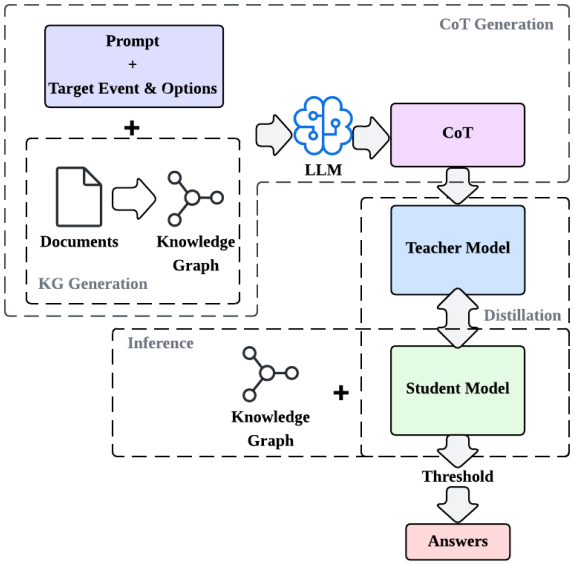
Abductive reasoning in NLP involves selecting the most plausible explanation for an observed outcome under incomplete information, in contrast to deductive or inductive inference.

Early lightweight systems for abductive reasoning predominantly adopted discriminative ranking strategies, as exemplified by benchmarks like αNLI (Bhagavatula et al., 2020). These methods treat abduction as a Multiple Choice Question Answering (MCQA) task, utilizing pre-trained models (e.g., DeBERTa) to calculate matching scores between an observation O and candidate hypotheses H_n. However, such approaches struggle with complex evidence extraction and the NOTA scenario. Alongside the rise of LLMs, generative explanation approaches that leverage prompt engineering to produce the most plausible hypothesis H from observation O have become one of the mainstream solutions (Wei et al., 2022), but the prohibitive computational demands of these large-scale models remain a significant barrier.

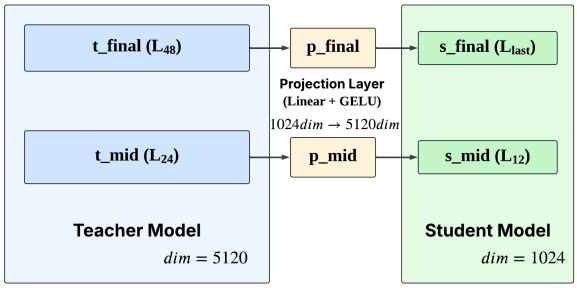
Alternatively, knowledge-augmented discriminative ranking which incorporates external knowledge graphs into the inference process (Lin et al., 2019), and knowledge distillation methods that transfer LLM reasoning capabilities to a sin-

[*]Equal contribution.
[1]https://github.com/KimmyHsu0909/SemEval2026_ARE

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 905–912



(a) The overall framework of our system.



(b) Overview of the multi-layer distillation framework.

Figure 1: System Architecture and Distillation Process.

gle (Hinton et al., 2015), deployable model, provide a novel perspective for addressing this task.

3 System Overview

Our framework operates in three primary stages: first, we extract structured evidence chains from source documents using Knowledge Graph techniques to provide factual grounding; second, we employ a knowledge-augmented distillation strategy, where a large-scale reasoning model generates CoT trajectories that serve as pedagogical signals to train a deployable student model; finally, the system performs multi-label causal inference by integrating the target event with KG-evidence, generating confidence scores for each option and selecting the most plausible causes based on a logic-consistent thresholding strategy. The overall architecture is illustrated in Figure a.

3.1 Structure-Aware Evidence Graph Construction and Reasoning

Following the paradigm of graph-based multi-hop reasoning (Ding et al., 2019), we construct an undirected evidence graph from the retrieved top-K documents. Instead of relying on a simple flat list of retrieved chunks, this graph topology allows us to trace multi-hop logical chains while explicitly modelling the semantic and structural relationships between textual fragments.

3.1.1 Graph Initialization

Unlike traditional Knowledge Graphs where nodes represent canonical entities and edges denote explicit relations, our evidence graph operates at the **sentence level**. It is an ad-hoc routing network designed to capture the logical flow between textual fragments.

Formally, we define the evidence graph as $G = (V, E)$. The vertex set V consists of two distinct types of nodes:

Endpoints: Two designated nodes representing the hypothesis option (h) and the observed target event (e).

Evidence Nodes: Sentences extracted from the top-K documents using spaCy (Honnibal et al., 2020), discarding fragments under 10 characters to preserve semantic density.

3.1.2 Hybrid Topology and Edge Weighting

To prevent overly dense graphs and semantic drift, we employ an asymmetric K-Nearest Neighbor (KNN) topology: endpoint nodes are connected to their top-6 neighbors ensuring reasoning entry points , while evidence nodes are connected to their top-4 neighbors to maintain sparsity. Edge weights combine three signals:

Semantic Similarity: Cosine similarity of all-MiniLM-L6-v2 (Reimers and Gurevych, 2019) embeddings.

Entity Co-occurrence: A symbolic reward for intersecting Named Entities (e.g., PERSON, ORG) extracted via spaCy, anchoring semantics in factual overlaps.

Structural Inductive Bias: An intra-document bonus of +0.6 for adjacent sentences and +0.1 for non-adjacent ones, preserving logical continuity.

These hyperparameters act as empirical heuristics based on natural discourse, circumventing the prohibitive cost of LLM-based grid search. To penalize weak connections during the pathfinding

stage, we map the final composite score S to a non-linear distance weight: $W = (1 - \min(1.0, S))^2$, heavily penalizing low-confidence edges during pathfinding.

3.1.3 Optimal Evidence Chain Retrieval

Following path-routing paradigms (Lin et al., 2019), we frame the reasoning process as a shortest-path problem, utilizing Dijkstra's algorithm to find the path from h to e minimizing the total weight $\sum W$.

To mitigate error accumulation and semantic drift inherent in long multi-hop reasoning (Yang et al., 2024), we introduce a **Dynamic Hop Penalty** as an internal filtering metric. Chains exceeding three hops incur a compounding penalty, suppressing convoluted paths prone to spurious correlations. A heavily penalized direct edge serves as a fail-safe for disconnected graphs.

3.1.4 Semantic-Driven Prompt Injection

The retrieved shortest path constitutes our optimal evidence chain. Although numerical metrics (scores and penalties) govern internal path retrieval, empirical tests show that exposing such metadata distracts the LLM. Consequently, we adopt a semantic-only representation, injecting the sentences as a cohesive logical narrative. The full prompt template used is detailed in Appendix A.

3.2 CoT Generation

We employ DeepSeek-R1 (Guo, Daya and Yang, Dejian and Zhang, Haowei and Song, Junxiao and Wang, Peiyi and Zhu, Qihao and others, 2025) to generate high-quality CoT reasoning trajectories, grounded by the KG evidence chains extracted in the preceding stage. By treating these CoT sequences as high-quality pedagogical rationales rather than simple augmented data, we shift the distillation objective from terminal answer prediction to logic-level alignment. This methodology ensures that the student model internalizes the complex logical induction processes of the Source Teacher, effectively bridging the gap between factual knowledge and causal reasoning.

For each option, the model is prompted to generate a four-step reasoning chain based on following diagnostic protocol:

Event Identification: Extracting and clarifying the core event from the option.

Temporal Alignment: Verifying that the candidate cause chronologically precedes the target effect.

Causal Sufficiency: Assessing whether the event provides a logically sufficient explanation without latent intermediate gaps.

Directness Conclusion: Determining the final plausibility of the direct causal link based on the preceding synthesis.

The detailed prompt is presented in Appendix B

3.3 Knowledge Distillation via Multi-Layer Feature Alignment

To bridge the gap between heavy generative logic and efficient discriminative inference, we design a multi-layer distillation framework (see Figure b). We adopt Qwen2.5-14B (Yang, An and Yang, Baosong and Zhang, Beichen and Hui, Binyuan and Zheng, Bo and Yu, Bowen and others, 2024) as the teacher (T) to provide high-order causal insights, and DeBERTa-v3-Large (He et al., 2023) as the student (S) for its superior efficiency in discriminative tasks. This setup allows the student to achieve LLM-level reasoning depth while maintaining high inference throughput.

3.3.1 Teacher Feature Extraction

We feed the previously derived CoT sequences C into the teacher model, prepended with an evaluation-oriented instruction. This process elicits the teacher's internal representations by treating the CoT as a logical prompt. We then extract hidden states from two strategic depths: the 24th layer (L_{24}) to capture structural logic and the 48th layer (L_{48}) for high-level semantic synthesis. The last-token hidden states are utilized as compact soft targets:

$$\mathbf{h}_T^{mid}, \mathbf{h}_T^{fin} = \text{Teacher}(C)$$

where $\mathbf{h}_T \in R^{5120}$ represents the distillation targets for student alignment.

3.3.2 Feature Alignment with Projection Layers

The student model ($d_s = 1024$) is equipped with two MLP-based projection heads, ϕ_{mid} and ϕ_{fin}, to bridge the dimensionality gap. Each head consists of a two-layer MLP with GELU activation, mapping student states to the teacher's 5120-dimensional space:

$$\mathbf{p}_{mid} = \phi_{mid}(\mathbf{h}_S^{L_{12}}), \quad \mathbf{p}_{fin} = \phi_{fin}(\mathbf{h}_S^{L_{last}})$$

We employ Cosine Similarity to supervise the alignment, focusing on directional semantic consistency

to enhance generalization:

$$\mathcal{L}_{distill} = \sum_{i \in \{mid, fin\}} (1 - \cos(\mathbf{p}_i, \mathbf{h}_T^i))$$

3.4 Entity-Aware Consistency Augmentation

To improve robustness and reduce reliance on superficial lexical cues, we implement an entity-aware masking strategy. Using spaCy (Honnibal et al., 2020), we replace named entities with type-specific placeholders (e.g., PERSON, ORG, GPE, LOC, EVENT) for a randomly sampled 50% of the dataset. For each instance, we construct a dual-view input consisting of the original sequence x and its masked counterpart x_{mask}. Following the R-Drop principle (Liang et al., 2021), we enforce representation invariance under these perturbations:

$$\mathcal{L}_{consist} = 1 - \cos(\mathbf{h}_S(x), \mathbf{h}_S(x_{mask}))$$

This regularization encourages the model to prioritize abstract causal structures over specific entity mentions.

3.5 Fine-grained Structural Contrastive Learning

To capture nuanced diagnostic logic, we introduce an intra-sample structural contrastive loss ($\mathcal{L}_{CL}$). Unlike standard methods, we align student features $\mathbf{P} = \{\mathbf{p}_1, \ldots, \mathbf{p}_N\}$ with their teacher counterparts $\mathbf{H}_T = \{\mathbf{h}_{T,1}, \ldots, \mathbf{h}_{T,N}\}$ via a position-wise similarity matrix $\mathbf{M} \in R^{N \times N}$, where each element represents the cosine similarity scaled by a temperature hyperparameter τ:

$$\mathbf{M}_{jk} = \frac{\cos(\mathbf{p}_j, \mathbf{h}_{T,k})}{\tau}$$

The objective is to maximize the alignment of corresponding candidate pairs while penalizing cross-candidate similarities within the same question:

$$\mathcal{L}_{CL} = -\frac{1}{N} \sum_{j=1}^{N} \log \frac{\exp(\mathbf{M}_{jj})}{\sum_{k=1}^{N} \exp(\mathbf{M}_{jk})}$$

This structures the student's latent space to mirror the teacher's inter-candidate reasoning boundaries.

3.6 Dynamic Multi-Objective Optimization

The total loss $\mathcal{L}_{total}$ is a curriculum-weighted combination of classification ($\mathcal{L}_{BCE}$) with label smoothing, distillation ($\mathcal{L}_{distill}, \mathcal{L}_{CL}$), consistency ($\mathcal{L}_{consist}$), and an auxiliary answer-count prediction loss ($\mathcal{L}_{count}$)

Let ρ denote the training progress ratio. The total objective is formulated as:

$$\mathcal{L}_{total} = \omega_1 \mathcal{L}_{BCE} + \omega_2(\mathcal{L}_{distill} + \gamma \mathcal{L}_{CL}) + \omega_3 \mathcal{L}_{consist} + \lambda \mathcal{L}_{count}$$

We employ a dynamic scheduling strategy where ω_1 increases with training progress ρ, while ω_2 and γ decay. This transition shifts the focus from high-order reasoning emulation to task-specific specialization. ω_3 remains constant to ensure feature robustness (details in Appendix C).

3.7 Inference and Multi-Label Causal Prediction

During inference, the model integrates event, options, and KG evidence to generate confidence scores $P(O_i)$ via Sigmoid activation. We apply a dual-threshold protocol: options exceeding threshold τ (the specific value of which is detailed in Appendix D) are selected; if none pass, the highest-scoring option is chosen. To maintain logical consistency, a "None of the others" (NOTA) exclusion rule is implemented, where NOTA takes precedence if multi-label conflicts occur.

4 Experimental Setup

4.1 Datasets & Implementation

We use the official datasets (Train/Dev/Test) provided by the SemEval-2026 Task 12 organizers for all stages of our experiments. All data are publicly accessible via the official repository.[2] Training was conducted on an NVIDIA A100, while inference requires only a single T4 GPU, highlighting resource efficiency. Specific hyperparameters are in Appendix E.

4.2 Baselines

We compare our system against the following three baselines to evaluate its performance:

DeepSeek-V3.2(Zero-shot) (DeepSeek-AI and others, 2025): A LLM used to provide a performance benchmark for large-scale reasoning without task-specific tuning.

Vanilla DeBERTa (He et al., 2021): A standard DeBERTa model fine-tuned directly on training labels, serving as a baseline for compact architectures without external enhancement.

BGE-Distilled DeBERTa: A distillation baseline where the student model aligns with features

[2] https://github.com/sooo66/semeval2026-task12-dataset

extracted by the BGE embedding model (Chen et al., 2024) from CoT texts.

4.3 Evaluation Metrics

The system performance is evaluated using the official scoring protocol provided by the organizers. Let $\mathbf{Y}$ be the set of golden labels and $\hat{\mathbf{Y}}$ be the predicted labels. The score S for each instance is defined as:

$$S = \begin{cases} 1.0, & \text{if } \hat{\mathbf{Y}} = \mathbf{Y} \text{ (Full Match)} \\ 0.5, & \text{if } \hat{\mathbf{Y}} \cap \mathbf{Y} \neq \emptyset \text{ and } \hat{\mathbf{Y}} \neq \mathbf{Y} \text{ (Partial Match)} \\ 0.0, & \text{otherwise} \end{cases}$$

The overall ranking is based on the mean score across the entire evaluation dataset.

5 Results

5.1 Main Results

Table 1: Main results on the SemEval-2026 Task 12 test set.

System	Task Score(Test set)
`our system`	**0.88**
`DeepSeek-V3.2 (Zero-shot)`	0.72
`Vanilla DeBERTa`	0.36
`BGE-Distilled DeBERTa`	0.71

Table 1 presents the performance of our system compared to various baselines on the official test set.The results lead to several key observations:

Superiority over Zero-shot LLMs: Our system outperforms the DeepSeek-V3.2 (Zero-shot) baseline by 0.16 points. This demonstrates that our approach delivers comparable or superior reasoning accuracy to LLMs within targeted domains, but with much higher efficiency and lower resource consumption.

Advantage of LLM-based Distillation: Our system shows a significant gain over the BGE-Distilled DeBERTa (0.88 vs. 0.71). Although both utilize the same CoT data, our approach of distilling internal hidden states from a generative teacher provides richer pedagogical signals for logical reasoning than general-purpose embedding vectors from BGE.

Significant Improvement from Vanilla Baseline: Compared to Vanilla DeBERTa, our system improves the score by 0.52 points. This underscores the massive impact of integrating external Knowledge Graphs and cross-architecture distillation in transforming a compact model into a powerful reasoner.

5.2 Ablation Study

Table 2: Ablation results on the SemEval-2026 Task 12 test set.

Model Variant	Score
`full system`	**0.88**
`w/o KG-derived Evidence`	0.79
`w/o CoT-based Distillation`	0.67
`w/o Entity-aware Masking`	0.83
`w/o Contrastive Learning`	0.81
`w/o Intermediate Layer`	0.73

We conducted an ablation study to evaluate the contribution of each component to our system's performance. As shown in Table 2, the following conclusions can be drawn:

Crucial Role of KG-Evidence: Removing the KG-derived evidence causes significant performance drop, with the score plummeting from 0.88 to 0.79. This demonstrates that structured factual evidence is a critical factor for accurate abductive reasoning.

Effectiveness of CoT Distillation: The exclusion of CoT-based distillation leads to a substantial decrease to 0.67, confirming that transferring reasoning trajectories from the LLM significantly improves the student's logical inference ability.

Impact of Multi-level Alignment: Without intermediate layer alignment, the score drops to 0.73, proving that aligning hidden states across multiple levels is more effective than using only the final layer.

Benefits of Masking and Contrastive Learning: Removing entity-aware masking and contrastive learning results in scores of 0.83 and 0.81, respectively. These results indicate that both components further enhance the model's robustness and discriminative power.

Overall, the full system achieves the best results, validating that all proposed modules contribute positively to the final performance.

6 Conclusion

In this paper, we presented our system for SemEval-2026 Task 12. Our framework integrates KG evidence extraction with a Logic-Driven Distillation strategy. Experimental findings demonstrate that

our system can achieve reasoning performance comparable to LLMs while significantly reducing computational costs and resource overhead. Future work will explore the integration of more diverse knowledge sources and iterative reasoning paths to further enhance the robustness of abductive inference.

7 Limitation

Despite the promising results, our system has several limitations. First, the performance is heavily reliant on the comprehensiveness of the underlying Knowledge Graph. In scenarios involving emerging events or long-tail entities, the evidence extraction module may fail to retrieve sufficient factual grounding. Second, while logic-driven distillation significantly boosts the efficiency of the student model, there remains a reasoning gap between the compact DeBERTa and the massive teacher LLM when handling multi-step, counterfactual abductive scenarios. Lastly, our current framework does not explicitly account for temporal dynamics in causal inference, which is crucial for real-world event reasoning where the validity of a cause may change over time.

References

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

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multilinguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14002–14020.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

DeepSeek-AI and others. 2025. DeepSeek-V3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Ming Ding, Chang Zhou, Qibin Chen, Hongxia Yang, and Jie Tang. 2019. Cognitive graph for multi-hop reading comprehension at scale. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2694–2703.

Guo, Daya and Yang, Dejian and Zhang, Haowei and Song, Junxiao and Wang, Peiyi and Zhu, Qihao and others. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations (ICLR)*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations (ICLR)*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in python. *Zenodo*.

Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. 2021. R-Drop: Regularized Dropout for Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 10871–10882.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-aware graph networks for commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2829–2839.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Srivastava, Aarohi and Rastogi, Abhinav and Rao, Abhishek and others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research (TMLR)*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-Thought prompting

elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 24824–24837.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024. Do large language models latently perform multi-hop reasoning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10210–10229.

Yang, An and Yang, Baosong and Zhang, Beichen and Hui, Binyuan and Zheng, Bo and Yu, Bowen and others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

A Appendix

The exact prompt template used to query the LLM is presented below. We explicitly strip all numerical metadata (such as confidence scores and hop penalties) to provide a semantic-only logical narrative.

LLM Prompt Template for Evidence Reasoning

Task: Identify if the Hypothesis is the direct cause of the Observed Event.

Event: *[event_text]*
Hypothesis: *[opt_text]*

Evidence Graph Analysis (Option Specific)
- Chain:
evidence_lines **Question:** Based on the chain, is the Hypothesis valid?

B Appendix

LLM Prompt Template for CoT generating

You are an expert in Abductive Reasoning. Your task is to analyze a "Target Event" and evaluate its potential explanatory options based on provided evidence.

Target Event: *[event_text]*
Evidence Context: *[formatted_options_block]*

Reference Golden Answer
The correct option(s) for this question is: [gold_answer]
Task:
For each option (A, B, C, D), generate a short, explicit reasoning chain to evaluate whether the option constitutes the most plausible direct cause of the Target Event.

A "direct cause" must satisfy all of the following:
- It can be justified through a single-step causal reasoning process.
- It is directly supported by evidence in the provided documents.
- It is logically and commonsensically sufficient to trigger the Target Event.

For each option:
1. State the key event described in the option.
2. State whether this event occurs before the Target Event.
3. State whether this event alone is sufficient to trigger the Target Event, without requiring additional intermediate events.
4. Conclude whether the option constitutes a direct cause of the Target Event.

Use short, factual statements.
Avoid comparative, evaluative, or meta-level language.
Do not mention correctness, plausibility rankings, or reference answers.

Example:
"A": "reasoning": "1. Option A describes U.S. airstrikes on Kataib Hezbollah facilities on December 29 in Iraq and Syria.2. These airstrikes occurred after earlier militia attacks but before the ballistic missile attacks.3. The airstrikes targeted an Iran-aligned militia, not Iranian state leadership or territory directly.4. Retaliation for such strikes would plausibly involve militia or proxy responses rather than direct Iranian ballistic missile launches.5. Therefore, this event alone is not sufficient to directly trigger Iran's ballistic missile attacks on Al Asad and Erbil air bases."

Output Format (Strict JSON):

"A": "reasoning": "...",
"B": "reasoning": "...",
"C": "reasoning": "...",
"D": "reasoning": "..."

C Appendix

The progress ratio is calculated as $\rho = \min(1.0, \frac{step}{max\_steps})$.

The specific linear scheduling functions and constant weights used in our experiments are defined as: $\omega_1 = 0.5 + 0.5\rho$, $\omega_2 = 1.0 - 0.7\rho$, $\gamma = 1.0 - 0.2\rho$, $\omega_3 = 5.0$, $\lambda = 0.1$

This configuration was empirically determined to balance the magnitude of gradients from heterogeneous sources.

D Appendix

To identify the optimal decision threshold τ, we performed a grid search over the range $[0.90, 0.99]$ with a step size of 0.01. Figure 2 presents the performance across various candidate thresholds. Empirically, we find that $\tau = 0.96$ yields the optimal score of 0.88, which was subsequently adopted for all downstream experiments.

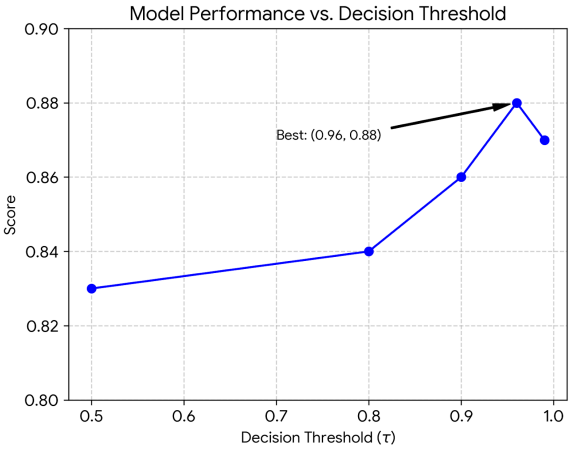


Figure 2: Sensitivity analysis of the decision threshold

E Appendix

Table 3: Main hyper-parameters tuned in our system.

Hyperparameter	Range/Value
Epoch	10
Batch Size	16
Weight Decay	0.01
Learning Rate	5e-6
Warm-up Rate	0.1