OseiBrefo-Liang at SemEval-2026 Task 12: Hybrid Causal Knowledge Graphs and Neural-Symbolic Policy Optimisation for Abductive Event Reasoning

Emmanuel Osei-Brefo[1] **and Huizhi Liang**[1]
[1]School of Computing, Newcastle University, UK
emmanuel.osei-brefo@newcastle.ac.uk
[2]Newcastle University, UK
huizhi.liang@newcastle.ac.uk

Abstract

Abductive Event Reasoning (AER) requires selecting plausible causal explanations for observed events from incomplete and noisy textual evidence. Unlike deductive reasoning, abductive inference proceeds from effects to candidate causes and is highly sensitive to distractor information and implicit multi-hop relationships. We present a hybrid neural-symbolic framework that models abductive reasoning as structured causal validation rather than unconstrained generation. Our framework integrates hybrid retrieval, micro-level evidence grounding, concept-level causal abstraction, reinforcement learning-based decision calibration, and structured Theorem-of-Thought verification. Experiments on SemEval-2026 Task 12 show that LLM reasoning constrained by structured causal graphs achieves the strongest development performance of 0.5288 and a leaderboard score of 0.61 on the test set, substantially outperforming symbolic-only and policy-only variants. These findings indicate that explicit causal modelling improves robustness in document-grounded abduction tasks.

1 Introduction

Transformer-based large language models (Vaswani et al., 2017) trained at scale and adapted via parameter-efficient methods (Hu et al., 2021) have significantly advanced natural language reasoning. Structured prompting strategies such as Chain-of-Thought (CoT) (Wei et al., 2022) and Zero-Shot-CoT (Kojima et al., 2022), improve multi-step inference reliability. Tree-of-Thought reasoning (Yao et al., 2024) also extends reasoning into search-based exploration. While Theorem-of-Thought (ToTh) reasoning formalises problem solving as a structured, multi-agent theorem-proving process. It decomposes reasoning into abductive, deductive, and inductive inference agents, whose outputs are organized into a formal reasoning graph (Abdaljalil et al., 2025).

Despite these progress, most work emphasises deductive, mathematical, or commonsense reasoning (Sun et al., 2023). Abductive reasoning, inferring plausible causes for observed effects, remains comparatively underexplored. Philosophically, abduction is characterised by explanatory inference under uncertainty (Magnani, 2023). Unlike deduction, it lacks guaranteed logical validity and requires selecting the most plausible hypothesis among competing explanations.

Cognitive research shows that humans perform abductive reasoning through iterative hypothesis–verification cycles (Bruner, 2017) while suppressing irrelevant contextual information (Johnston and Dark, 1986). These findings suggest computational frameworks should explicitly model causal authenticity and incorporate mechanisms for distractor suppression.

Abductive commonsense reasoning datasets (Bhagavatula et al., 2020) highlight explanation selection challenges. Causal inference theory (Pearl, 2009) emphasises structured cause–effect modeling.

Neural-symbolic integration (Garcez et al., 2019) combines structured representations with neural generalization. Retrieval-augmented generation (Lewis et al., 2020) improves factual grounding but does not enforce causal constraints.

SemEval-2026 Task 12, formalises this challenge in Abductive Event Reasoning (AER). Systems must select causal explanations for real-world events given noisy multi-document evidence. Purely neural systems risk hallucinated causal alignment, while purely symbolic systems lack semantic flexibility. We propose a hybrid neural-symbolic architecture combining structured causal modeling with policy optimization and reasoning verification.

Our contributions are as follows:

- Dual-view causal modelling through micro-

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2268–2274

level and concept-level knowledge graphs.

- Hybrid retrieval for robust evidence grounding.
- Reinforcement learning aligned with partial-credit evaluation.
- Structured reasoning verification using Theorem-of-Thought.

2 Task Definition

Each instance of the dataset includes the following:

- Target event e
- Candidate options $O = \{o_A, o_B, o_C, o_D\}$
- Supporting documents $D = \{d_1, \ldots, d_n\}$

Where, A, B, C and D are the possible Answers for each question.

2.1 Evaluation Metric

System performance is evaluated at the instance level using an exact and partial matching scheme over the predicted answer options.

Let G denote the set of gold-standard correct options for a given instance, and let P denote the set of options predicted by the system. Each instance receives a score according to the following criteria:

- **1.0 (Full Match):** if $P = G$.
- **0.5 (Partial Match):** if P is a non-empty proper subset of G, i.e., the prediction contains at least one correct option and does not include any incorrect options.
- **0.0 (Incorrect):** otherwise, including cases where P contains any incorrect option or is empty.

The objective is to predict $\hat{Y} \subseteq O$. This can be summarised as:

$$Score(\hat{Y}, Y) = \begin{cases} 1.0 & \hat{Y} = Y \\ 0.5 & \hat{Y} \subset Y \\ 0.0 & \text{otherwise} \end{cases} \quad (1)$$

The final system score is computed as the average score across all evaluation instances:

$$\text{Final Score} = \frac{1}{N} \sum_{i=1}^{N} \text{Score}(P_i, G_i), \quad (2)$$

where N is the total number of evaluation instances.

3 Methodology

3.1 System Overview

Our approach builds on previous advances in prompt-based reasoning, structured search, and neural-symbolic modelling while addressing limitations specific to abductive reasoning. In contrast to forward reasoning frameworks, we explicitly model reverse inference from observed events to candidate causes. Inspired by hypothesis–verification patterns in human cognition (Bruner, 2017), we transform abductive inference into structured forward validation over causal graphs.

Additionally, it integrates hybrid retrieval, micro-level evidence graphs, concept-level causal graphs, reinforcement learning–based decision optimisation, and multi-path reasoning verification. This unified architecture enables explicit modelling of cause authenticity while mitigating interference from irrelevant evidence.

By combining structured causal modelling with multi-path reasoning and policy optimisation, our framework advances abductive event reasoning beyond purely neural or purely symbolic approaches.

Our system consists of five major components, which are; Hybrid retrieval module, micro-level evidence knowledge graph, concept-level causal knowledge graph, reinforcement learning policy optimisation, and theorem-based reasoning verification.

Figure 1 shows the architecture of our proposed approach.

The hybrid retrieval component computes a composite relevance score by linearly combining multiple retrieval signals, including TF-IDF, BM25, and lexical matching:

$$S_{retr} = \alpha S_{tfidf} + \beta S_{bm25} + \gamma S_{lex}. \quad (3)$$

Retrieved evidence is then used to construct a micro-level knowledge graph $G_m = (V_m, E_m)$, where nodes represent evidence fragments and candidate options, and edges encode local causal relationships extracted from text.

For each candidate option o_k, a micro-level support score is computed by aggregating supporting and contradictory evidence weights:

$$S_{micro}(o_k) = \sum_{support} w_e - \sum_{contradict} w_e. \quad (4)$$

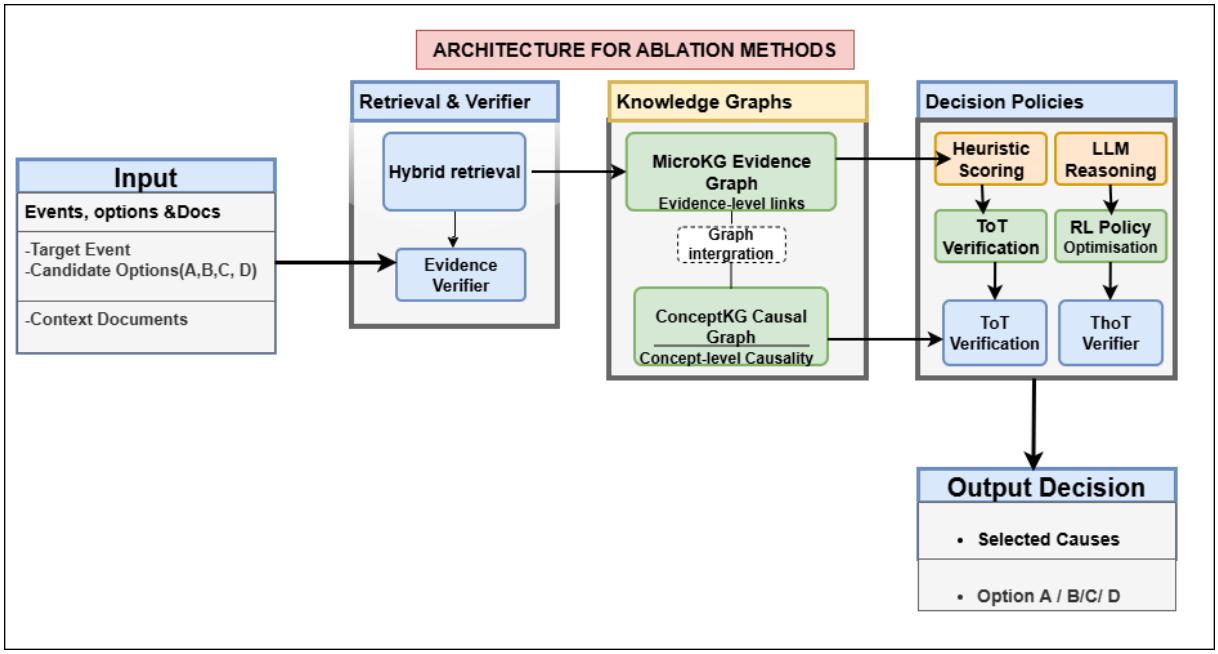


Figure 1: System architecture for the proposed framework to evaluate the Abductive Event Reasoning

In parallel, a concept-level causal knowledge graph $G_c = (V_c, E_c)$ models higher-level semantic relationships between abstract concepts.

The concept-level score for each candidate option is derived by aggregating causal connections between concepts appearing in the option and those associated with the target event:

$$S_{concept}(o_k) = \sum_{(c_i, c_j)} \mathbf{1}(c_i \in o_k, c_j \in e). \quad (5)$$

Decision calibration is performed using a reinforcement learning policy $\pi_\theta(a \mid s)$:

$$\pi_\theta(a \mid s). \quad (6)$$

The reward function is aligned with the official evaluation metric. Finally, structured reasoning validation is conducted through a Theorem-of-Thought verification mechanism, where confidence scores are aggregated and regularised using an entropy term.

Refer to Appendix A for the implementation details used.

4 Experiments

The dataset used had 1819 instances as a training set, 400 as a development set, and a test set of 612 instances.

Figure 2(a) shows the distribution of causality signals in the training set, indicating the relative frequency of causal patterns. Figure 2(b) illustrates the distribution of supporting documents per instance, highlighting the variability in evidence availability across examples.

5 Results and Analysis

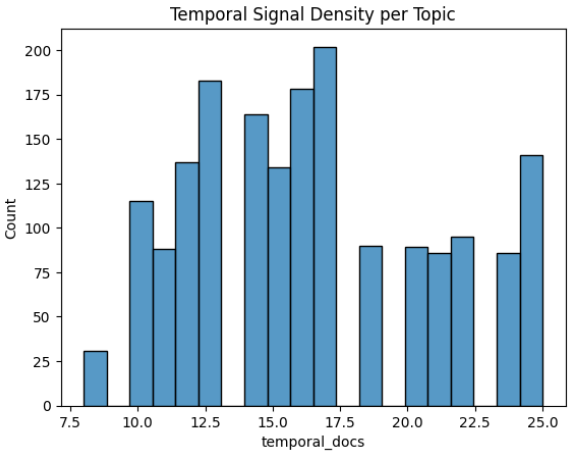
Our proposed system achieves a test score of **0.61** on the leaderboard and a development score of **0.5288** when micro-level and concept-level knowledge graphs are combined with LLM reasoning, shown as S3 in Table 1). This variant outperformed all symbolic-only and policy-only versions.

5.1 Ablation Study

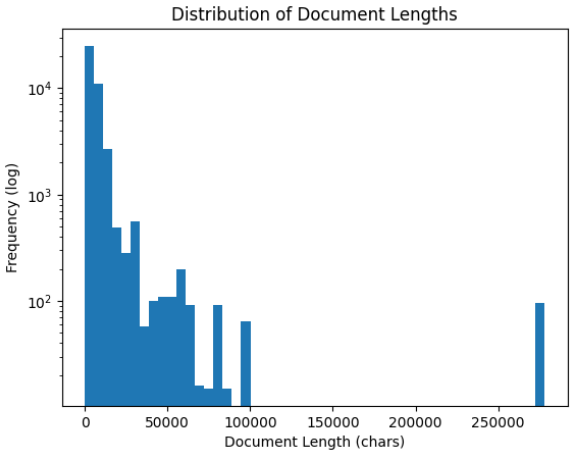
To better understand the contribution of each architectural component, we conducted a comprehensive ablation study on the full development set which had 400 instances. Table 1 reports the evaluation score on the development data, average prediction set size, and invalid prediction rate.

5.2 Component-Level Analysis

Heuristic Baseline. The base verifier heuristic, which is just the LLM with ID, S3 performed poorly and achieved only 0.0838 with an average prediction size of 2.685. The large prediction set size indicates uncontrolled over-selection, confirming that abductive reasoning requires structured modelling rather than naive scoring. This served as the baseline used.



(a) Causality signal in the training set



(b) Document distribution

Figure 2: Dataset statistics: (a) distribution of causality signals in the training set; (b) distribution of supporting documents per instance.

ID	System Variant	Dev Score	Avg. Pred Size	Invalid Rate
S1	Base Verifier Heuristic	0.0838	2.685	0.00
S2	microKG Only	0.3675	1.005	0.00
S3	LLM Reasoning on (micro+concept) KG View	**0.5288**	1.115	0.00
S4	RL Verifier Only	0.4538	1.175	0.00
S5	RL Verifier + microKG	0.4538	1.175	0.00
S6	conceptKG Only	0.3225	1.000	0.00
S7	hybridKG (micro+concept) No LLM	0.3200	1.000	0.00
S8	Tree-of-Thought (Beam)	0.3200	1.000	0.00
S9	Graph-of-Thought Decision	0.3200	1.000	0.00
S10	RL Verifier + microKG + conceptKG	0.3863	1.315	0.00
S11	Hybrid Verifier-Level + RL(micro+concept)	0.4088	1.2625	0.00
S12	ThoT Verifier-Level + RL(micro+concept)	0.4588	1.3125	0.00

Table 1: Ablation results on the SemEval-2026 Task 12 development set. The base LLM model used was openAI's GPT-4o mini model

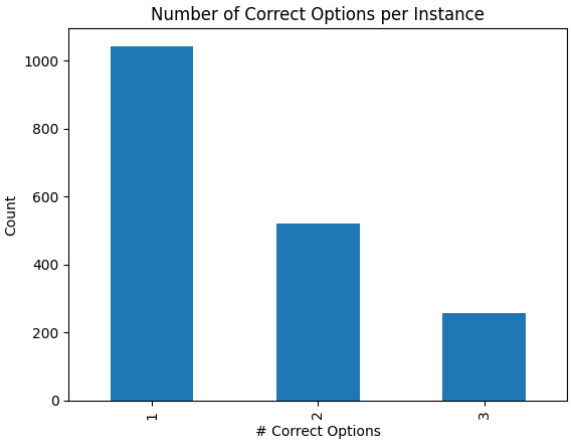


Figure 3: Correct options in answers

Impact of Micro-Level Evidence Modeling. The micro-level knowledge graph alone with ID S2 improved performance substantially to 0.3675 while maintaining a near-single-option prediction size of 1.005. This demonstrates that explicit modelling of local evidence–option relationships provides strong grounding and reduces distractor influence.

Concept-Level Causal Abstraction. The concept-level knowledge graph alone with ID S6, achieved a score of 0.3225, slightly below microKG. This suggests that corpus-level causal abstraction captures useful general patterns but is less discriminative than direct evidence grounding when used independently.

Symbolic Hybrid Without LLM Reasoning. The hybridKG variant without LLM reasoning with ID S7 achieved 0.3200, and demonstrates that the simple combination of symbolic graph signals does not automatically yield performance gains. This indicates that neural reasoning is necessary to effectively integrate multi-view graph information.

LLM Reasoning Over Structured Graphs. The strongest result was achieved by ID S3 with a score of 0.5288 on the development set. This was achieved when the LLM was made to reason over both micro-level and concept-level graphs. This confirms that structured graph constraints enhance neural reasoning and help mitigate unsupported causal inferences. The improvement over ID S2 and ID S6 highlights the complementary roles of local grounding and abstract causal structure.

Reinforcement Learning Calibration. The RL-based verifier alone with ID S4 achieved a score of 0.4538, which outperformed symbolic-only graph variants. This indicates that policy learning improves calibration under the partial-credit evaluation metric. However, adding microKG explicitly to RL (S5) did not further improve performance, suggesting that the learned policy already captures much of the micro-level signal.

Verifier-Level Structured Reasoning. Structured reasoning at the verifier level improves robustness when combined with RL. Hybrid Verifier-Level + RL with ID S11 achieved a score of 0.4088, while Theorem-of-Thought Verifier-Level + RL with ID S12 improved further to 0.4588. This demonstrates that structured reasoning chains can enhance stability when integrated with learned decision policies.

Pure Tree-of-Thought and Graph-of-Thought Strategies. Tree-of-Thought with ID, S8 and Graph-of-Thought with ID, S9 decision strategies without strong evidence grounding plateau at 0.3200. This suggests that reasoning search mechanisms alone cannot compensate for weak evidence filtering or insufficient causal modelling.

5.3 Key Findings

The ablation study reveals several important insights, such as:

- Local evidence grounding (microKG) is more informative than generalized concept abstraction when used independently.
- LLM reasoning constrained by structured knowledge graphs provides the largest performance gain.
- Reinforcement learning improves decision calibration but does not replace structured reasoning.
- Structured reasoning (ThoT) enhances verifier-level robustness when combined with policy learning.
- Pure reasoning frameworks without strong evidence modelling underperform.

Overall, the results support the central hypothesis of this work, which is that abductive event reasoning benefits from explicit causal structure combined with calibrated neural reasoning rather than unconstrained generation.

6 Error Analysis

To better understand system limitations, we manually inspected incorrect predictions on the development set and identified recurring error patterns. We do not introduce new quantitative claims beyond the reported ablation results but instead analyse qualitative failure modes.

6.1 Implicit Multi-Hop Causality

Several errors occur when the causal relationship between a candidate explanation and the target event requires multi-step reasoning that is not explicitly captured in either the micro-level or concept-level graphs. In such cases, partial evidence is present, but intermediate causal links are missing, leading to under-selection.

6.2 Lexical Distractor Influence

Although hybrid retrieval reduces noise, some distractor options share strong lexical overlap with supporting documents without representing genuine causal relations. These cases reveal that lexical similarity can still propagate into graph construction, especially at the micro-level.

6.3 Concept Over-Generality

The concept-level knowledge graph abstracts semantic relationships across documents. However, in some instances, abstraction introduces overly broad causal connections between semantically related but non-causal concepts. This occasionally leads to over-selection.

6.4 Subset Calibration Errors

Because the evaluation metric rewards exact subset prediction, some errors arise from threshold calibration rather than incorrect ranking. In several cases, the correct option receives the highest score, but additional lower-confidence options are included, resulting in partial or zero credit.

Overall, these error categories suggest that improving edge precision in concept graphs and refining subset calibration strategies are promising directions for future work.

7 Conclusion

We presented a hybrid neural-symbolic framework for Abductive Event Reasoning in SemEval-2026 Task 12. By modelling abduction as structured causal validation over dual-view knowledge graphs, our approach integrates hybrid retrieval, micro-level evidence grounding, concept-level abstraction, reinforcement learning–based calibration, and Theorem-of-Thought verification.

Ablation results demonstrate that LLM reasoning constrained by structured causal graphs achieves the strongest development performance of 0.5288 and a leader board score of 0.61 on the test set, substantially outperforming symbolic-only and policy-only variants. These findings indicate that explicit causal modelling improves robustness in document-grounded abductive reasoning tasks.

Future work will look at how to improve multi-hop causal edge construction, refine subset calibration under partial-credit metrics, and explore tighter neural-symbolic integration mechanisms. It will also look at including more capable LLM models such as GPT-5.2 as the base model in addition to other fine-tuned open source models.

References

Samir Abdaljalil, Hasan Kurban, Khalid Qaraqe, and Erchin Serpedin. 2025. Theorem-of-thought: A multi-agent framework for abductive, deductive, and inductive reasoning in language models. In *Proceedings of the 3rd Workshop on Towards Knowledgeable Foundation Models (KnowFM)*, pages 111–119, Vienna, Austria. Association for Computational Linguistics.

Chandra Bhagavatula, Doug Downey, Antoine Bosselut, and Yejin Choi. 2020. Abductive commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3435–3448. Association for Computational Linguistics.

Jerome S. Bruner. 2017. *A Study of Thinking*, 2nd edition. Routledge, New York, NY, USA.

Artur S. d'Avila Garcez, Marco Gori, Luís C. Lamb, Luciano Serafini, Michael Spranger, and Son N. Tran. 2019. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. *Journal of Applied Logics*, 6(4):611–631.

Edward J. Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

William A. Johnston and Veronica J. Dark. 1986. Selective attention. *Annual Review of Psychology*, 37(1):43–75.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Patrick Lewis, Ethan Oguz, Ruty Rinott, Sebastian Riedel, Pontus Stenetorp, Dani Yogatama, and Tim Rocktäschel. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Lorenzo Magnani, editor. 2023. *Handbook of Abductive Cognition*, 1 edition. Springer, Cham, Switzerland.

Judea Pearl. 2009. *Causality: Models, Reasoning, and Inference*, 2nd edition. Cambridge University Press, Cambridge, UK.

Hongda Sun, Weikai Xu, Wei Liu, Jian Luan, Bin Wang, Shuo Shang, Ji-Rong Wen, and Rui Yan. 2023. From indeterminacy to determinacy: Augmenting logical reasoning capabilities with large language models. *arXiv preprint arXiv:2310.18659*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H. Chi, Quoc V. Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36. Curran Associates, Inc.

A Implementation Details

Hybrid Retrieval: For each instance, a query is constructed through the concatenation target event and candidate option. TF-IDF and BM25 scores are computed using standard sparse retrieval over the document corpus. Lexical matching is implemented using token overlap between query and document chunks. The final retrieval score is computed as a weighted sum of normalised scores, and the top-k evidence chunks are selected.

Edge Extraction: Edges in the micro-level graph are constructed using rule-based causal pattern matching over retrieved evidence. Specifically, we identify causal indicators such as "because", "due to", "leads to", and "results in". If a sentence contains both a candidate option and a target event (or related concept) connected by such indicators, a causal edge is created.

Support and Contradiction Weights: Supporting evidence weights are assigned when the extracted relation is consistent with the candidate causing the event. Contradictory weights are assigned when negation or inverse causal patterns are detected. All weights are normalised based on retrieval confidence.

Concept Graph Construction: Concept-level graphs are constructed through the extraction of key noun phrases and entities from retrieved documents. Co-occurrence and causal patterns are used to form directed edges between concepts.

Reinforcement Learning Setup: We model decision selection as a classification problem over candidate subsets. The policy is implemented using gradient boosted decision trees trained on development data. The reward function directly follows the official evaluation metric, assigning 1.0, 0.5, or 0.0 based on prediction correctness.

LLM Reasoning: The LLM receives structured representations of both micro-level and concept-level graphs as input and generates reasoning traces used for final scoring.