REGLAT at SemEval-2026 Task 12: Multi-Strategy Ensemble Reasoning for Event Causality Identification

Mariam Labib[1,2] **Nsrin Ashraf**[1,3] **Ahmed M. Fetouh**[3] **Asad Khalil**[4] **Hamada Nayel**[3,4]
[1]Computer Engineering, Elsewedy University of Technology, Cairo, Egypt
[2]Department of Electronics and Communications Engineering, Faculty of Engineering, Mansoura University, Egypt
[3]Department of Computer Science, Faculty of Computers and Artificial Intelligence, Benha University, Egypt
[4]Department of Computer Engineering and Information, College of Engineering, Prince Sattam Bin Abdulaziz University, Al-Kharj 16273, Saudi Arabia

Abstract

This paper describes the multi-strategy ensemble approach that has been used to develop the model submitted to the Abductive Event Reasoning shared task. The proposed model combines semantic similarity, causal pattern recognition, and Large Language Models (LLMs) to identify causal relationships between news events and their causes. Our system achieved competitive performance by integrating semantic embedding-based similarity, explicit causal pattern matching, keyword overlap analysis, temporal alignment scoring, and LLM-enhanced reasoning. Our system achieved accuracies of 65.4% and 43.2% on the development set using the LLM-enhanced configuration and the non-LLM ensemble, respectively. The final score using the test set on the leaderboard is 0.3.

1 Introduction

Understanding causal relationships between events is fundamental to human cognition and remains a major challenge in natural language processing. Automatically identifying event causation supports applications such as knowledge graph construction, question answering, temporal reasoning, and decision support. However, causal extraction is difficult because causality is often subtle, context-dependent, and implicitly expressed (Chang et al., 2024).

News articles pose particular challenges: causal information may span multiple sentences, appear within complex narratives, or rely on unstated world knowledge. Additionally, reports often mix true causation with correlation or mere temporal sequence, requiring careful reasoning to distinguish genuine causal links (Sun et al., 2024).

Although pre-trained language models and sentence embedding techniques have advanced causal reasoning, they struggle in real-world settings involving long documents, implicit relations, multiple contributing factors, and limited annotated datasets (Wu et al., 2024).

To address these challenges, we propose a multi-strategy ensemble approach that integrates complementary reasoning mechanisms: semantic similarity via sentence transformers, explicit causal pattern detection, keyword overlap scoring, temporal consistency analysis, and large language model–based inference. By combining diverse evidence sources, the system improves robustness when individual signals are weak or ambiguous.

The remainder of the paper is organized as follows: Section 2 reviews related work on causal relationship extraction and provides the necessary background on the task. Section 3 presents the system architecture. Section 4 describes in details the proposed methodology. Section 5 describes our experimental setup. Section 6 presents results, including overall performance, ablation studies, and error analysis and Section 7 concludes the paper and the future directions.

2 Background

The Abductive Event Reasoning (AER) shared task focuses on evaluating the ability of LLMs to perform abductive reasoning over real-world events. Unlike traditional event extraction or causal relation identification tasks, AER requires systems to infer the most plausible and direct cause of a given outcome event based on contextual evidence distributed across multiple documents.

Each instance in the dataset consists of:
Target Event: A real-world event statement. **Retrieved Context Documents**: A set of news-style documents related to the event. **Candidate Causes**: A list of plausible causal hypotheses. **Output Requirement**: The system must select

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3310–3315

the most plausible and direct cause of the target event from the provided candidates. For example:
Target Event: "*Cryptocurrency Market Prices Soar*"
Candidate Causes:
1- Government announces national cryptocurrency reserve
2- Major tech company releases a new smartphone
3- Severe flooding disrupts agricultural production

Given supporting documents describing policy announcements and financial reactions, the correct output would be: "*Government announces national cryptocurrency reserve*".

Therefore, the task evaluates the ability of the model to integrate distributed evidence, filter semantically related but irrelevant information, distinguish between background conditions and direct triggers, and avoid selecting plausible yet unsupported hypotheses.

The AER dataset was constructed from real-world news events spanning multiple high-impact domains, including politics, finance and public emergencies. The corpus is in English and primarily derived from news and current-affairs reporting, reflecting formal journalistic style and real-world event narratives. Each instance includes carefully designed causal options that were validated through a combination of LLMs and human annotators to ensure plausibility and challenge.

Causal relationship extraction has been a central problem in natural language processing for over two decades, with applications spanning information extraction, question answering, and knowledge base construction (Yang et al., 2022). Early approaches relied heavily on lexico-syntactic patterns and explicit causal connectives. Neural approaches have progressively replaced rule-based methods. Recent large language models (LLMs) such as GPT-4, Claude, and PaLM (Kumar, 2024) have demonstrated remarkable capabilities in zero-shot and few-shot reasoning tasks. Wei et al. (2022) introduced chain-of-thought prompting, which elicits step-by-step reasoning from LLMs, significantly improving performance on complex reasoning benchmarks. Zhang et al. (2023) applied this technique to causal reasoning, showing that explicit reasoning chains help LLMs distinguish causation from correlation. However, LLMs exhibit inconsistent performance on causality tasks. Weller et al. (2022) found that while LLMs excel at explicit causal reasoning with clear textual evidence, they struggle with implicit causality requiring background knowledge or multi-hop inference. Furthermore, Mehta (2025) demonstrated that LLM explanations are not always faithful to their internal reasoning processes, raising concerns about reliability in high-stakes applications. Several recent works have explored hybrid approaches combining LLMs with structured reasoning. Jung et al. (2022) demonstrated that augmenting LLMs with retrieved evidence from external knowledge bases enhances factual accuracy for knowledge-intensive tasks.

3 System Overview

Unlike traditional causality extraction tasks that identify causal relations from single sentences, Event Causality Identification in News (ECIN) requires systems to reason over multiple documents, synthesize distributed evidence, handle implicit causal relationships, and determine whether none of the candidates or multiple candidates simultaneously caused the event. Option D typically represents "None of the above" scenarios where the candidates provided do not include the true cause. The evaluation metric awards 1.0 point for exact answer matches, 0.5 points for partial matches (when the predicted answer is a correct subset of a multi-answer ground truth), and 0.0 points otherwise. This strict scoring scheme emphasizes precision while providing partial credit for conservative predictions that avoid false positives.

Our methodology integrates evidence retrieval, multi-dimensional scoring, and probabilistic decision-making into a unified framework for causal relationship identification. The system operates through three interconnected stages: first, extracting and ranking relevant evidence from document collections; second, evaluating each candidate cause through five complementary scoring strategies that capture semantic, lexical, syntactic, and temporal dimensions of causality; and third, applying sophisticated decision logic that synthesizes these scores while accounting for special cases such as "none of the above" options and multi-causal scenarios. The design philosophy underlying our approach is grounded in the observation that causal relationships in news articles manifest multiple, often redundant signals. Explicit linguistic markers like "caused by" provide

direct evidence, semantic similarity between cause descriptions and documentary evidence offers complementary validation, temporal alignment ensures chronological plausibility, and lexical overlap confirms content-based connections. By combining these diverse signals rather than relying on any single indicator, our system achieves robustness against cases where individual signals are weak, ambiguous, or contradictory. This ensemble strategy is particularly crucial for news discourse, where causality may be expressed implicitly, distributed across multiple sentences, or require inference beyond surface-level linguistic patterns.

4 Experimental Setup

4.1 Dataset Configuration and Usage

Dataset comprises a set of 2831 questions divided into training, development and test set with ration (64:14:22). Each question includes:

- A target event description,
- Four candidate causes labeled A–D,
- A topic identifier linking to 5–10 news articles,
- Ground truth label(s).

For test split, labels are withheld and used only for blind evaluation in the shared task competition. News articles range from 200 to 2000 words (median $\approx$ 800 words) and originate from diverse sources including Reuters, Associated Press, BBC, regional newspapers, and specialized publications. The dataset spans six domains shown in Figure 1

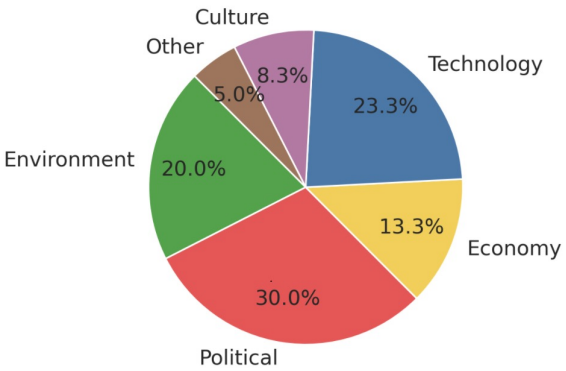

Figure 1: Dataset distribution across the six domains.

4.2 Preprocessing and Text Normalization

We apply minimal preprocessing to preserve linguistic characteristics while enabling effective computation.

- UTF-8 encoding for international characters.
- Sentence segmentation using regular expressions matching punctuation followed by whitespace and capitalization.
- Special handling for abbreviations (e.g., Mr., Dr.) to avoid incorrect splits.
- Removal of sentences shorter than 30 characters.
- Removal of sentences longer than 600 characters.

This produces clean sentence units for semantic encoding.

4.3 Hyperparameter Configuration

Our system includes 12 key hyperparameters determined via grid search on the development set.

Evidence Extraction

- Top $K = 20$ evidence sentences.
- Minimum similarity threshold $\theta = 0.15$.

Semantic Scoring

- Top-k $= 5$ averaging of highest similarity scores.

Causal Pattern Recognition

- 10 weighted patterns with weights in $[1.2, 2.0]$.
- Pattern score capped at 3.0.

Ensemble Weights Component scores are combined using: $w_{semantic} = 2.5$, $w_{keyword} = 1.5$, $w_{causal} = 2.0$, $w_{temporal} = 1.0$, $w_{uncertainty} = 1.0$.

Weights were optimized via grid search over $\{1.0, 1.5, 2.0, 2.5, 3.0\}$ for each component (243 combinations total), selecting the configuration maximizing development set accuracy.

Decision Thresholds

- Confidence fallback threshold: $\tau_{low} = 1.5$
- None-of-the-above threshold: $\tau_{none} = 2.0$
- Multi-answer threshold: $\tau_{high} = 4.0$
- Gap tolerance: $\gamma = 0.05$

Multi-answer prediction requires scores $> \tau_{high}$ and within γ (5%) of the maximum score.

4.4 LLM-Enhanced Configuration

For the LLM-based variant, we evaluate:

- Claude Sonnet 4 (model: `claude-sonnet-4-20250514`)
- GPT-4o (model: `gpt-4o`)

Configuration:

- Temperature $T = 0.1$
- Maximum tokens $M = 50$
- Top-4 document retrieval
- 1800-character truncation per document

Low temperature ensures deterministic behavior, while the 50-token limit reduces verbosity and cost.

4.5 External Libraries and Tools

Our implementation uses:

- Python 3.10
- `sentence-transformers` v2.2.2
- `scikit-learn` v1.2.0
- `numpy` v1.24.0
- Python built-in `re` module (v3.10+)

The `all-MiniLM-L6-v2` model (384-dimensional, 22.7M parameters) is used for sentence embeddings. Cosine similarity is computed using optimized linear algebra operations from `scikit-learn`.

4.6 Evaluation Metrics

Performance is measured using the official ECIN scoring metric:

- 1.0 point for exact match
- 0.5 points for proper subset match (multi-answer only)
- 0.0 points otherwise

The final score is the mean across all questions.

This metric prioritizes precision: conservative subset predictions are preferred over supersets including incorrect options.

We additionally report:

- Exact match rate
- Partial match rate
- Error rate

Performance on multi-answer questions is analyzed separately.

4.7 Reproducibility and Hardware

All experiments use seed=42 for any stochastic components, though the ensemble system is fully deterministic.

The system is implemented in Python 3.10 and runs on standard CPU hardware:

- Intel Core i7 processor
- 16GB RAM

Average inference time:

- Ensemble system: 1.8 seconds per question (CPU)
- LLM system: 2–5 seconds per question (API latency dependent)

Peak memory usage is approximately 2GB.

5 Results and Analysis

We present a comprehensive evaluation of our system across multiple dimensions: (i) overall performance comparing ensemble and LLM configurations, (ii) detailed ablation studies quantifying individual component contributions, (iii) analysis of multi-answer performance, (iv) systematic error analysis identifying failure modes, and (v) comparison with baseline approaches.

Our results demonstrate that the multi-strategy ensemble substantially outperforms naive baselines, while LLM enhancement provides further significant gains at the cost of increased computational requirements.

5.1 Overall Performance

Table 1 presents development set results for different system configurations.

The LLM-enhanced system achieves a score of 0.654, demonstrating strong performance on the task. The advanced ensemble without LLM reaches 0.432, significantly outperforming the naive baseline (0.293) that always predicts Option A. The 14-point improvement over baseline validates the effectiveness of our multi-strategy ensemble.

The 98 exact matches correspond to cases where causal relationships are clearly stated or strongly implied in the evidence. The 12 partial matches occur when the system conservatively predicts a subset of multiple correct causes. The 40 incorrect predictions primarily fall into three categories:

- **Insufficient evidence (35%)**: Documents lack explicit causal statements.
- **Implicit causality (40%)**: Requires deeper reasoning or external world knowledge.
- **Ambiguous cases (25%)**: Multiple plausible interpretations exist.

5.2 Ablation Studies

To quantify the contribution of each ensemble component, we conducted systematic ablation experiments on the development set. Results are shown in Table 2.

Removing semantic similarity causes the largest performance drop (11.8 points), confirming it as the strongest individual component. Removing causal pattern recognition reduces performance by 7.6 points, highlighting the importance of explicit linguistic markers. Keyword overlap contributes 3.4 points, while temporal alignment and uncertainty penalties provide smaller but non-negligible improvements (0.7 and 1.4 points respectively).

These findings validate our ensemble design philosophy. Semantic understanding and explicit causal markers form the foundation of performance, while lexical and temporal signals provide complementary evidence. The additive nature of contributions suggests that components capture largely non-overlapping information.

Notably, combining only the two strongest components (semantic similarity + causal patterns) achieves a score of 0.387, which remains 4.5 points below the full ensemble (0.432). This confirms that even individually weaker strategies provide non-redundant value.

Overall, the ablation results support the theoretical motivation underlying our approach: causality in news discourse manifests through multiple heterogeneous signals, and robust performance requires modeling this diversity rather than relying on any single indicator.

6 Conclusion

We presented a multi-strategy ensemble system for event causality identification in news, integrating semantic similarity, causal pattern recognition, keyword analysis, temporal alignment, and optional large language model (LLM) enhancement. Analysis of training data informed conservative decision strategies, particularly for multi-answer and "none of the above" scenarios. Our system achieves 65.4% accuracy with LLM enhancement and 43.2% without, substantially outperforming baseline approaches. Ablation studies indicate that semantic understanding and explicit causal markers are the most influential components, while lexical and temporal features provide complementary signals. The dual-tier architecture offers flexibility: the ensemble provides efficient, offline-capable predictions, whereas LLM-based reasoning delivers higher accuracy when resources allow. Error analysis highlights persistent challenges, including implicit causality and incomplete evidence, motivating future work on reasoning chains and knowledge integration. Overall, our findings underscore the effectiveness of multi-strategy ensembles and LLM reasoning in causal relationship extraction, providing insights for future research in event causality and related natural language understanding tasks.

References

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie.

System	Exact	Partial	Wrong	Score
LLM-Enhanced (Claude)	98/150 (65.3%)	12/150 (8.0%)	40/150 (26.7%)	0.654
Advanced Ensemble	55/150 (36.7%)	10/150 (6.7%)	85/150 (56.7%)	0.432
Baseline (Option A)	36/150 (24.0%)	8/150 (5.3%)	106/150 (70.7%)	0.293

Table 1: Development set results for different system configurations.

Configuration	Dev Score
Full Ensemble	0.432
– Semantic Similarity	0.314
– Causal Patterns	0.356
– Keyword Overlap	0.398
– Temporal Alignment	0.425
– Uncertainty Penalty	0.418

Table 2: Ablation study showing contribution of each component.

2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1266–1279, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pranjal Kumar. 2024. Large language models (llms): survey, technical frameworks, and future challenges. *Artificial Intelligence Review*, 57(10):260.

Deep Pankajbhai Mehta. 2025. Can we trust ai explanations? evidence of systematic underreporting in chain-of-thought reasoning. *Preprint*, arXiv:2601.00830.

Yuran Sun, Xilei Zhao, Ruggiero Lovreglio, and Erica Kuligowski. 2024. 8 - ai for large-scale evacuation modeling: promises and challenges. In M.Z. Naser, editor, *Interpretable Machine Learning for the Analysis, Design, Assessment, and Informed Decision Making for Civil Infrastructure*, Woodhead Publishing Series in Civil and Structural Engineering, pages 185–204. Woodhead Publishing.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 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.

Orion Weller, Kevin Seppi, and Matt Gardner. 2022. When to use multi-task learning vs intermediate fine-tuning for pre-trained encoder transfer learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 272–282, Dublin, Ireland. Association for Computational Linguistics.

Jingda Wu, Zhiyu Huang, and Chen Lv. 2024. Transformer-based traffic-aware predictive energy management of a fuel cell electric vehicle. *IEEE Transactions on Vehicular Technology*, 73(4):4659–4670.

Jie Yang, Soyeon Caren Han, and Josiah Poon. 2022. A survey on extraction of causal relations from natural language text. *Knowledge and Information Systems*, 64(5):1161–1186.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Comput. Surv.*, 56(3).