AI4PC-Howard University at SemEval-2026 Task 12: Evidence-Guided Abductive Scoring with Option-Conditioned Retrieval and Constrained LLM Evaluation

Ifeoluwakiitan Ayandosu and Saurav K. Aryal*
AI4PC Lab
Howard University
ayandosu10@gmail.com
saurav.aryal@howard.edu

Abstract

Abductive event reasoning in the wild requires selecting plausible explanations for an event from noisy, partially relevant multi-document context. We present an evidence-guided abductive scoring pipeline for SemEval-2026 Task 12 that separates evidence selection from explanation scoring.For each topic, we chunk documents and retrieve option-conditioned evidence using dense embeddings, then apply a cross-encoder reranker to form compact evidence packs per option. A constrained large language model scorer evaluates each option using only its evidence pack and outputs structured signals capturing evidence support, explanatory directness, and contradiction. We then apply deterministic decision rules to produce single or multi-label predictions, including robust handling of "none of the above" style options through lexical-cue detection rather than reliance on option position. This modular design reduces distraction from irrelevant documents, improves comparability across options, and enables controlled calibration for multi-answer outputs. Our approach demonstrates that retrieval-focused evidence compression combined with disciplined, signal-based scoring can effectively support abductive reasoning without explicit knowledge graphs or end-to-end prompting over full document context.

1 Introduction

Abductive Event Reasoning (AER) tests whether a system can identify the most plausible explanation of a real-world event using noisy multi-document evidence. SemEval-2026 Task 12 poses this as a multiple-choice, multi-label problem (Task Organizers, 2026): given a target event, retrieved web documents, and four candidate explanations A–D, select all supported labels. Gold answers can include one to four labels, and partial credit is awarded for overlapping sets.

The dataset presents several challenges beyond causal ambiguity. Documents include topical distractors; the "insufficient information" option is not reliably tied to label D (sometimes D is a normal explanation or a duplicate of another option); and identical option texts can appear under different labels with both included in the gold answer. These properties make label-based heuristics brittle and motivate text-based processing.

Our system follows a two-stage pipeline: (1) option-conditioned evidence retrieval with cross-encoder reranking, and (2) constrained plausibility scoring with deterministic decision rules. Our key findings are:

- Prompt framing has a larger impact than retrieval parameters: changing the scorer from "cause" to "accounts for" framing improved mean development exact-match (EM) by $\sim$10 points.
- Multi-label calibration is the dominant error source; most errors involve predicting one correct label but missing others.
- Topic isolation is critical: sharing a retrieval index across topics collapsed 291/400 predictions to a single label.
- Lexical-cue detection of insufficient-information options outperforms label-based heuristics.

2 Background

Abductive NLI (Bhagavatula et al., 2020) formalized explanation selection over narrative contexts. Retrieval-augmented pipelines (Lewis et al., 2020) with cross-encoder reranking (Nogueira and Cho, 2019) are standard for knowledge-intensive tasks (Petroni et al., 2021; Rijal and Aryal, 2025; Aryal and Akomoize, 2025); dense passage retrieval (Karpukhin et al., 2020) provides high recall but often surfaces topically related but non-causal evidence, making reranking critical in causal settings (Sapkota et al., 2023; Tiwari et al., 2025). Knowledge-graph approaches to event reasoning

*Corresponding author

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2630–2634

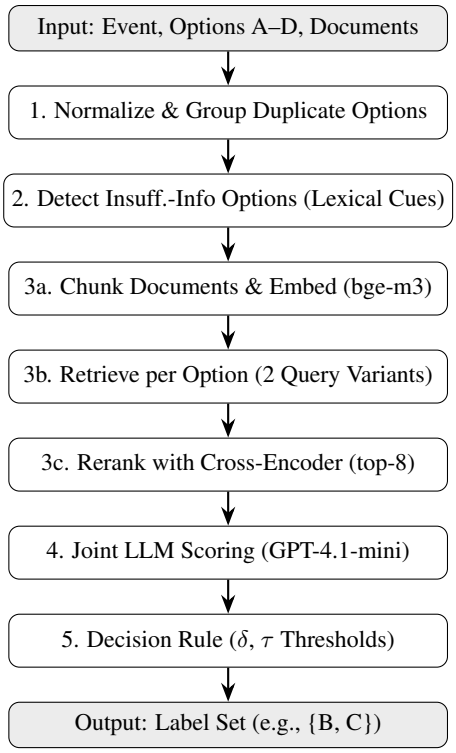

Figure 1: System pipeline. Each topic is processed independently. Steps 3a–3c are repeated per option to produce option-specific evidence packs.

(Sap et al., 2019) offer structured causal representations but introduce extraction errors; we avoid this by operating directly over text. Our work conditions retrieval separately on each candidate explanation and uses structured scoring signals rather than a single plausibility number, enabling fine-grained calibration of multi-label decisions (Kadavath et al., 2022).

3 System Overview

The pipeline (Figure 1) has five steps: (1) normalize and group duplicate options, (2) detect insufficient-information options by lexical cues, (3) retrieve and rerank option-conditioned evidence, (4) score with constrained plausibility signals, and (5) select labels via calibrated decision rules. Full configuration details are in Appendix A.

3.1 Option Normalization

We normalize options by lowercasing, stripping punctuation, and collapsing whitespace. Identical normalized texts are grouped into a single candidate; at prediction time, the group expands back to all labels. This handles cases where B and D map to the same text and both appear in the gold answer.

3.2 Insufficient-Information Detection

We detect insufficient-information options via lexical-cue matching against keywords such as "insufficient information," "none of the above," "none of the others," and "none are correct." Matched options are flagged for downstream decision logic but still scored normally. This approach is limited to surface-level cues and will miss paraphrased expressions of insufficiency; embedding-based detection was considered but not implemented.

3.3 Evidence Retrieval and Reranking

Topic documents are chunked ($\sim$250 tokens, 40-token overlap) and embedded with BAAI/bge-m3 (Chen et al., 2024) via the Sentence-Transformers library (Reimers and Gurevych, 2019) into a per-topic FAISS index (Johnson et al., 2021). For each option, we issue two query variants—a direct query (event + option) and a causal query ("what caused [event]" + option)—retrieve the top-60 chunks across the union, and rerank with BAAI/bge-reranker-v2-m3 to keep the top-8 as the evidence pack (Aryal and Pant, 2025; Aryal and Akomoize, 2025; Aryal and Pant, 2025).

3.4 Constrained Plausibility Scoring

All four options and their evidence packs are scored jointly in a single call to GPT-4.1-mini (OpenAI, 2025) (API string `gpt-4.1-mini`, temperature 0). The prompt (Appendix B) returns three signals per option: support strength $s \in [0, 1]$, explanatory directness $d \in [0, 1]$, and contradiction $c \in \{0, 1\}$. The final score is:

$$\text{score}(o_i) = s_i \cdot d_i \cdot (1 - \lambda \cdot c_i) \qquad (1)$$

with $\lambda = 0.8$. Missing fields default to 0 (support/directness) or false (contradiction).

The contradiction indicator is a binary flag from GPT-4.1-mini with no independent NLI verification, making it brittle and model-dependent.

Prompt framing. Our initial prompt asked whether an option was the "cause" of the event, which penalized valid gold labels that are contributing factors or immediate circumstances rather than strict causes. Changing to "how well does the explanation account for the event" yielded our largest single improvement (Section 4).

3.5 Decision Rule

Let o^* be the highest-scoring option. An additional option o_i is included if: $\text{score}(o_i) \geq \text{score}(o^*) - \delta$,

Configuration	T1 *28q*	T4 *15q*	T3 *45q*	Mean
Cause framing	.571	.800	.311	.561
+ Insuff. detect.	.607	.933	—	—
Explan. framing	—	—	.444	—
Best per-topic	.607	.933	.444	.661
Full dev (400q)		.490		
Test (612q)		**.530**		

Table 1: Exact-match accuracy. T1/T3/T4 are topic IDs. Dashes indicate the configuration was not re-evaluated on that topic. "Best per-topic" aggregates the best result per topic across configurations.

$\text{score}(o_i) \geq \tau_{\text{add}}$, $s_i \geq 0.65$, $d_i \geq 0.40$, and $c_i = 0$. If an insufficient-information option exists and all non-insufficient scores fall below τ_{best}, the insufficient option is selected; it is never included alongside non-insufficient options. Thresholds (δ=0.03, τ_{add}=0.60, τ_{best}=0.55) were hand-tuned on the development set without systematic sensitivity analysis.

4 Results

Table 1 reports exact-match (EM) accuracy. On the official test set (612 questions), our system achieves 0.53 EM, placing 142nd on the leaderboard. We report only EM because per-instance predictions were not preserved, preventing computation of partial-credit metrics (Jaccard similarity, F1).

Prompt framing dominates. Across the three topics evaluated under both framings, mean EM improved from 0.561 to 0.661—a 10-point gain exceeding any retrieval or threshold change. On Topic 3 alone, EM rose from 0.311 to 0.444.

Insufficient-information detection. Adding lexical-cue detection improved Topic 4 from 0.800 to 0.933 by correctly handling "None of the others are correct causes."

Topic isolation. Sharing one FAISS index across all topics collapsed 291/400 predictions to label A (EM 0.318); per-topic indexing restored EM to 0.490.

Scope of experiments. Due to computational constraints (∼1–2 hours per full evaluation), we did not run controlled ablations for the reranker, query variants, or joint vs. independent scoring. Results reflect iterative development rather than factorial experiments.

5 Error Analysis

We analyze errors from per-topic development evaluations.

Under-prediction of multi-label answers (∼60% of errors). The most common pattern is predicting one correct label but missing others. For instance, on question q-78 (gold {A,C,D}), the system predicted only {A}; options C and D had moderate support but fell below the inclusion threshold. The multiplicative score (Eq. 1) exacerbates this: if either s or d is 0.6, the product is 0.36, which fails the τ_{add}=0.60 gate.

Incorrect single-label winner (∼25%). The system sometimes picks a label entirely absent from the gold set. On q-40 (gold {A,B}), the system predicted {C}—evidence for C was topically related but temporally distinct from the target event.

Insufficient-information misclassification (∼15%). On q-162 (gold {A}), the system predicted {D} because all non-insufficient scores were marginal and fell below τ_{best}. Conversely, on q-176 (gold {B,C}), the system selected D because the lexical-cue detector flagged it and the non-insufficient scores were borderline.

6 Discussion and Limitations

"Explanation" is broader than "cause." Our most actionable finding is that gold labels often include contributing factors, immediate circumstances, and directly linked consequences—not only strict causes. Prompts enforcing causal language systematically under-score valid labels.

Decision rules are the bottleneck. The scorer usually identifies the top option correctly; the dominant error is in the multi-label gate. This is a calibration problem that may benefit from learning the threshold from data rather than hand-tuning.

Limitations. Our system has several concrete limitations: (1) all decision thresholds (δ, τ_{add}, τ_{best}) are hand-tuned without systematic sensitivity analysis; (2) the contradiction signal is a binary flag from GPT-4.1-mini with no independent NLI verifier; (3) the scorer depends on a closed commercial API that may change; (4) we did not perform controlled ablations due to compute constraints; (5) insufficient-information detection uses keyword matching only; (6) the multiplicative score overpenalizes options with one moderately low signal.

7 Ethical Considerations

The task relies on web documents that may contain political or social biases inherited through retrieval and LLM scoring. Incorrect causal attributions can reinforce misleading narratives in politically sensitive topics. We release code and prompts to support transparency.

8 Conclusion

We presented an evidence-guided abductive scoring pipeline for SemEval-2026 Task 12 achieving 0.53 exact-match on the official test set (rank 142). Our main finding is that prompt framing—whether the scorer evaluates "causation" or "explanatory support"—yields a ∼10-point EM improvement, exceeding all retrieval and threshold changes. Multi-label calibration remains the dominant bottleneck. Future work includes learning the decision gate from data, adding embedding-based insufficient-information detection, running controlled ablations, and incorporating an independent NLI verifier (Bowman et al., 2015).

Acknowledgments

We thank the SemEval-2026 Task 12 organizers for designing and maintaining the shared task.

References

Saurav Aryal and Mildness Akomoize. 2025. Howard university-ai4pc at semeval-2025 task 3: Logit-based supervised token classification for multilingual hallucination span identification using xgbod. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1790–1794.

Saurav Aryal and Kritika Pant. 2025. Howard university-ai4pc at semeval-2025 task 9: Using open-weight bart-mnli for zero shot classification of food recall documents. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1919–1923.

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 *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 632–642.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

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

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*.

OpenAI. 2025. GPT-4.1 family of models. Accessed: February 2026.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yaber, Nicola De Cao, Yacine Jernite, Luca Singh, Dmitry Lagun, Timothée Lacroix, and 1 others. 2021. KILT: A benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 2523–2544.

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 (EMNLP)*, pages 3982–3992.

Suprabhat Rijal and Saurav Aryal. 2025. Howard university-ai4pc at semeval-2025 task 7: Crosslingual fact-checked claim retrieval-combining zero-shot claim extraction and knn-based classification for multilingual claim matching. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1777–1782.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019.

ATOMIC: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.

Hrishav Sapkota, Saurav Keshari Aryal, and Howard Prioleau. 2023. Zero-shot classification reveals potential positive sentiment bias in african languages translations.

Task Organizers. 2026. SemEval-2026 Task 12: Abductive event reasoning. Accessed: January 2026.

Saharsha Tiwari, Saurav K Aryal, and Legand Burge. 2025. Enhancing geospatial reasoning in large language models: An optimized retriever approach using r-tree-based point-in-polygon and nearest neighbor search. In *International Conference on Information and Communication Technology for Intelligent Systems*, pages 509–523. Springer Nature Singapore Singapore.

A System Configuration

Component	Setting
Embedding model	BAAI/bge-m3
Reranker	BAAI/bge-reranker-v2-m3
LLM scorer	GPT-4.1-mini (`gpt-4.1-mini`)
LLM temperature	0
LLM max tokens	1000
API access period	Jan–Feb 2026
Chunk size	250 tokens
Chunk overlap	40 tokens
Retrieved chunks (N)	60
Reranked chunks (K)	8
Contradiction penalty λ	0.8
Multi-label margin δ	0.03
Multi-label threshold τ_{add}	0.60
Insuff. threshold τ_{best}	0.55
Hardware	Google Colab, 1×T4 GPU

Table 2: Full system configuration.

As GPT-4.1-mini is a closed commercial model, exact reproducibility may be affected by provider updates.

B Scoring Prompt

The following prompt is used for joint scoring. Event, options, and evidence packs are substituted at runtime.

```
You are evaluating how well each
explanation accounts for an event
using only the evidence snippets.

Event: {event}

Explanations:
A: {option_A}  B: {option_B}
C: {option_C}  D: {option_D}

Evidence for A: {evidence_A}
Evidence for B: {evidence_B}
Evidence for C: {evidence_C}
Evidence for D: {evidence_D}

For each option, output:
support_strength: 0 to 1
causal_directness: 0 to 1
  (treat as explanatory directness)
contradiction: true or false
justification: one short sentence
  citing snippet ids like [A1] [B2]

Return JSON only. Do not wrap in
backticks. Do not add extra text.
```