# SemEval-2026 Task 12: Knowledge Graph with hyperbolic embedding in Abductive Event Reasoning

**Mingkai Wang**[1] **Varun Ojha**[1] **Huizhi Liang**[1]

[1]School of Computing, Newcastle University
Newcastle upon Tyne, UK
{c5025965, Varun.Ojha, huizhi.liang }@newcastle.ac.uk

**abstract: this task introduces Abductive Event Reasoning (AER), a novel shared task, to investigate the ability of Large Language Models(LLMs) to reason about the causality of real-world events. More specifically, a data set consisting of different topics and choices is introduced, and we need to enable the model to select the best options for the given event. Three methods are separately introduced to explore the question, including the traditional natural language processing(NLP) method (DeBERTa), the enhanced knowledge graph(KG), and the KG embedded in hyperbolic space. Our implementation could be downloaded: https://github.com/MingkaiW/SemEval2026-task12**

## 1 Introduction

Understanding why events happen is crucial for humans making sense of the world and for intelligent systems that aim to interpret it. While large language models (LLMs) have achieved strong performance on tasks such as event extraction, summarization, and even forecasting, they still struggle with abductive reasoning: inferring the most plausible cause of an observed outcome from incomplete, noisy, or distributed evidence.

SemEval-2026 Task 12: Abductive Event Reasoning (AER) is designed to investigate this ability. Given a real-world event and a set of retrieved documents, the systems must identify the most plausible and direct cause among multiple natural-language candidates. Compared to standard information extraction or summarization, AER demands structured, context-based causal reasoning that combines textual evidence with prior knowledge. Progress in this task has direct implications for transparency, explainability, and decision making, with potential applications in journalism, analysis, and public information systems.

In this paper, we study AER in a single-modality setting using three families of baselines: (i) a classical embedding-based multi-label classifier built on frozen representations and Logistic Regression, (ii) a strong discriminative multiple-choice model based on DeBERTa-v3, and (iii) knowledge-graph–augmented LLMs that leverage a COMET-enhanced causal KG in either Euclidean (RotatE) or hyperbolic (RotH) geometry. In the development set, these systems obtain scores of approximately 0.73 (DeBERTa-v3), 0.60 (Euclidean KG–LLM) and 0.61 (hyperbolic KG–LLM). The results show that the discriminative DeBERTa-v3 baseline remains the strongest overall, COMET-based KG–LLM methods substantially narrow the gap, and moving from Euclidean to hyperbolic KG embeddings yields only marginal improvements under our current training regime.

## 2 Related work

The foundational benchmarks for abductive reasoning were first introduced in 2020 by the alphaNLI dataset[Bhagavatula et al., 2020]. However, despite the fact that LLMs have demonstrated impressive capabilities in event extraction, summarization, and future prediction, they still fall short in abductive reasoning—inferring the most likely cause of a given outcome from incomplete or distributed information. This could be viewed in many works such as [Kiciman et al., 2024] [Chi et al., 2024] [Miliani et al., 2025] [Liu et al., 2025b]. In this task of SemEval-2026, it is necessary to reason about the documents retrieved about real-world news events [Cao et al., 2026]. Traditional NLP methods and knowledge graph methods are suitable for this task. Furthermore, we refer to a method of enhanced KG[Liu et al., 2025a] to improve its score. Another key point is that, We consider the hyperbolic embedding model and try to integrate it with the knowledge graph model to obtain a knowledge graph model with hyperbolic embedding and



*Proceedings of the The 20th International Workshop on Semantic Evaluation (2026)*, pages 821–825

verify its effect.[Sun et al., 2020] [Chami et al., 2020]

## 3 Dataset

The team that proposed the task has built a high-quality dataset that spans politics, finance, technology, public emergencies, etc., where causal options are carefully constructed and validated through both LLMs and human annotators. Given the input of an event (e.g., "Cryptocurrency Market Prices Soar") and a set of retrieved documents, the target request that the model identify the most plausible and direct cause, such as "Government announces national cryptocurrency reserve". One or more options may be correct, and the result of each instance is measured by the Evaluation metric, where the full match gets 1 point; the partial match gets 0.5 points; and the wrong match gets 0 points.

The whole data set is divided into train data, development data and test data, the instances in the first two include answers (the gold labels) for training and tuning models, and in the test data set the answer field is removed for evaluating the final result.

Each data set above all split consists of two files:

- questions.jsonl: event descriptions and multiple-choice options, each line in it is a JSON object representing one multiple-choice reasoning instance;
- docs.jsonl: retrieved contextual documents for each event, each record in the docs file contains background context for an event.

Each instance consists of:

- Event: A short description of an observed real-world event;
- Context: A set of retrieved documents related to the event (also have some distractor docs);
- Options (A–D): Four candidate explanations for the event, written as natural language sentences.

Through pilot experiments, it has been proved that even state-of-the-art LLMs struggle with this task due to semantic distraction and a tendency to plausible yet incorrect answers. These findings reveal an important limitation in the current reasoning capacity of LLM.

## 4 Baselines Overview

### 4.1 LLMs

Baseline 1 uses fixed text (and optionally image) embeddings with a shallow multi-label classifier. For each AER instance, we compute frozen encoder embeddings and its associated documents, aggregate them into a single feature vector, and train a multi-output Logistic Regression model to predict the set of correct options.

### 4.2 Natural language process (DeBERTa)

Baseline 2 shares a single text-preprocessing stage and then splits into two branches: a zero-shot UnifiedQA generator and a fine-tuned DeBERTa-v3 multiple-choice classifier. Both branches operate on the same question–context serialization and output a set of option labels that are scored with the official SemEval metric.

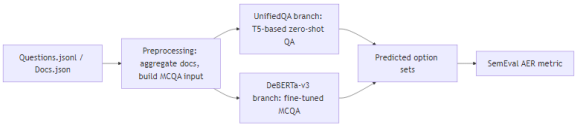


Figure 1: Pipeline for Baseline 2 (UnifiedQA + DeBERTa-v3). A shared preprocessing step produces a unified MCQA representation, which is consumed by a zero-shot UnifiedQA branch and a fine-tuned DeBERTa-v3 branch.

Each instance is first converted into a UnifiedQA-style text-to-text input and a SWAG-style MCQA record. UnifiedQA directly decodes the option strings (e.g. "A,B") without task-specific training, while 'microsoft/deberta-v3-base' is fine-tuned in the MCQA format. Then both predictions are equally evaluated.

### 4.3 Enhanced Knowledge graph

Relative to a standard KG pipeline (data → triples → KG → embeddings → downstream task), our Baseline 3 introduces several changes:

- Task-specific node set: entities anchored to SemEval AER questions, candidate options, and salient spans from retrieved documents;
- COMET-based causal edges: edges generated mainly by COMET-ATOMIC 2020 causal hypotheses with task-focused relations (Causes, isBefore, xEffect, oEffect, isAfter) instead of generic factual predicates;

- KG as knowledge layer: the KG is used primarily for retrieval and knowledge compression, while final abductive reasoning is performed by the LLM (DeepSeek-chat);
- Textualized KG context: the retrieved triples are verbalized into short natural-language snippets and fused with the text of the document and the MCQA prompt;

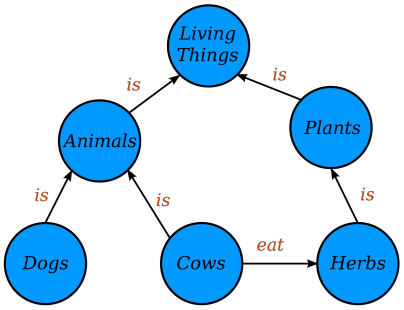


Figure 2: Figure of a standard knowledge graph

### 4.4 Knowledge graph embedded on the hyperbolic space

From a geometric perspective, Baseline 3 and Baseline 4 share the same COMET-enhanced causal KG, the same triples, and the same ranking objective; the only change is the embedding space.

In Baseline 3, we used a Euclidean RotatE embedding in a complex space. Each entity $e$ is represented as a complex vector $e \in \mathbb{C}^d$ and each relation $r$ as a unit module complex vector $r \in \mathbb{C}^d$ with $|r_k| = 1$. Given a triple $(h, r, t)$, RotatE applies a relation-specific rotation and scores

$$f_{\text{RotatE}}(h, r, t) = \gamma - \big\|h \circ r - t\big\|_2, \qquad (1)$$

where $h \circ r$ is the multiplication of complexes in elemental order, $\|\cdot\|_2$ is the Euclidean norm, and $\gamma$ is a fixed margin. Training minimizes a margin-based ranking loss with negative triples $\mathcal{N}(h, r, t)$:

$$\mathcal{L}_{\text{RotatE}} = \sum_{(h,r,t)} \Big[ -\log \sigma\big(f_{\text{RotatE}}(h, r, t)\big) - \sum_{(h',r,t') \in \mathcal{N}(h,r,t)} \log \sigma\big(-f_{\text{RotatE}}(h', r, t')\big) \Big].$$

In Baseline 4, we keep the same triples and loss shape, but embed entities in a Lorentz-model hyperbolic space instead of a Euclidean space. Entities live on the Lorentzian hyperboloid

$$\mathbb{H}^d_L = \big\{ x \in \mathbb{R}^{d+1} : \langle x, x \rangle_L = -1,\ x_0 > 0 \big\},$$

where the Lorentzian inner product is defined as

$$\langle u, v \rangle_L = -u_0 v_0 + \sum_{i=1}^{d} u_i v_i,$$

and the hyperbolic distance between two embeddings of the entity $u, v \in \mathbb{H}^d_L$ is defined as

$$d_L(u, v) = \operatorname{arcosh}\big(-\langle u, v \rangle_L\big).$$

The relations are modeled as transformations $\phi_r$ that act on hyperbolic points, and the triple score becomes

$$f_{\text{RotH}}(h, r, t) = -\, d_L\big(\phi_r(h),\ t\big).$$

Then the loss is analogous to RotatE but uses the hyperbolic score:

$$\mathcal{L}_{\text{RotH}} = \sum_{(h,r,t)} \Big[ -\log \sigma\big(f_{\text{RotH}}(h, r, t)\big) - \sum_{(h',r,t') \in \mathcal{N}(h,r,t)} \log \sigma\big(-f_{\text{RotH}}(h', r, t')\big) \Big].$$

## 5 Experiment setup

### 5.1 Data Splits and Usage

We follow the official SemEval-2026 Task 12 data splits: sample (200 instances, 10 topics), train (1,819 instances, 36 topics), dev (400 instances, 36 topics) and test (612 instances, 20 topics). All baselines consume the JSONL/JSON files provided by the organizer ('questions.jsonl' and 'docs.json') without changing labels or topics.

For Baseline 2, UnifiedQA is applied in zero-shot to the preprocessed dev and test splits. The DeBERTa-v3 branch trains on the processed train split and uses dev for model selection; the selected checkpoint is then run once in test to produce the final submission. For Baseline 3 and Baseline 4, we use the full train split to build the causal KG (events and options as entities) and run the KG–LLM QA pipeline in dev and test.

### 5.2 Preprocessing and Hyperparameters

Table 1 summarizes the main preprocessing and hyperparameter settings for all single-modality baselines. We list only the most influential numerical choices; additional implementation details are described in the corresponding baseline subsections.

Beyond the settings in Table 1, the Baseline 2 additionally uses a SWAG-style multiple-choice encoding with full multi-label field `labels_all`.

| Baseline | Input / representation | Model / geometry | Dim | Max length / tokens | Batch | Epochs | LR |
|---|---|---|---|---|---|---|---|
| B1: Classical | Qwen text emb; SigLIP img (opt.); concat + StandardScaler | Logistic Regression (multi-output, Euclidean) | – | – | – | – | – |
| B2: UQA + DeBERTa | Q + 4 options + $\leq$ 5 docs; title+body $\rightarrow$ 2048 chars | UnifiedQA (T5); DeBERTa-v3 MCQA | 768 | UQA: 512–768; DeBERTa: 256 | 4–8 | 3 | 2e-5 |
| B3: KG + RotatE | Same text as B2 + COMET triples $\rightarrow$ Causal KG | RotatE KGE (Euclidean) + DeepSeek-chat | 256 | KG text $\approx$ 400–500 | 256 | 100 | 1e-3 |
| B4: KG + RotH | Same as B3 (COMET-enhanced KG) | RotH KGE (hyperbolic, Lorentz) + DeepSeek-chat | 32 | Same as B3 | 256 | $\leq 60$ | 5e-4 |

Table 1: Preprocessing and key hyperparameters for the baseline 1-4.

Baselines 3 and 4 share the same COMET-enhanced causal KG and prompt-fusion strategy, differing only in the geometry of the KG encoder (Euclidean RotatE vs. hyperbolic RotH).

### 5.3 Libraries and Tooling

The experiments are run in a Conda environment based on Python 3.10 on Windows 11, with the packages PyTorch 2.10.0 and Transformers 4.36.2. UnifiedQA and DeBERTa-v3 are loaded from the Hugging Face Model Hub (https://huggingface.co), and COMET-ATOMIC 2020 from 'mismayil/comet-bart-ai2'. Hyperbolic embeddings are based on Geoopt 0.5.1 (https://github.com/geoopt/geoopt). LLM calls use the official Python SDKs for OpenAI (openai 2.15.0) and Anthropic (anthropic 0.76.0), plus huggingface-hub 0.36.0 when needed.

## 6 Results

### 6.1 Comparisons between NLP and KG

| System | Backbone / LLM | Score | Exact Match | Partial Match | Error Rate |
|---|---|---|---|---|---|
| Baseline 1 (LLM) | claude-3-haiku | 0.3925 | 31.5% | 15.5% | 53.0% |
| **Baseline 2 (MCQA)** | DeBERTa v3 base | **0.7250** | **49.5%** | **46.0%** | **4.5%** |
| Baseline 3 (KG + Haiku, no COMET) | claude-3-haiku | 0.4625 | 33.25% | 26.0% | 40.75% |
| Baseline 3 (KG + DeepSeek + COMET) | DeepSeek-chat | 0.6088 | 46.25% | 29.25% | 24.50% |

Table 2: Dev_set (dev_data, $N = 400$) comparison among Baseline 1 to Baseline 3

The DeBERTa v3 base classifier (Baseline 2) achieves the best dev-set score, with high exact and partial match rates and a very low error rate. Among KG-LLM systems, switching from Haiku to DeepSeek-chat and enabling COMET-based KG expansion yields a strong gain over the Haiku + KG configuration.

### 6.2 Comparisons between KG and Hyperbolic embedded KG

We next evaluate Baseline 4, which keeps the COMET-enhanced causal KG and DeepSeek-chat LLM but replaces the Euclidean RotatE embeddings of Baseline 3 with hyperbolic RotH embeddings trained on the same graph.

| System | Backbone / LLM | Score | Exact Match | Partial Match | Error Rate |
|---|---|---|---|---|---|
| Baseline 3 (KG + DeepSeek + COMET) | DeepSeek-chat | 0.6088 | 46.25% | 29.25% | 24.50% |
| **Baseline 4(Hyperbolic KG + DeepSeek)** | **DeepSeek-chat** | **0.6100** | **46.75%** | **28.50%** | **24.75%** |

Table 3: Dev set (dev_data, $N = 400$) comparison among Baseline 3 and Baseline 4

Baseline 4 slightly improves the overall dev score over the best Baseline 3 configuration (0.6100 vs. 0.6088), with a marginally higher exact-match rate and a very similar error rate. This suggests that once a strong COMET-enhanced prompt is in place, switching the KG representation from Euclidean RotatE to hyperbolic RotH has a tiny effect on the accuracy of the end-task.

## 7 Conclusion

We investigate abductive event reasoning in a single-modality setting using three progressively stronger baselines. A classical embedding-based classifier (Baseline 1) provides a lightweight refer-

ence point, but is clearly outperformed by a fine-tuned DeBERTa-v3 multiple-choice model (Baseline 2), which achieves the best overall dev-set score. Building a COMET-enhanced causal knowledge graph and prompting LLMs with structured causal context (Baseline 3) substantially improve the score, demonstrating the value of external common sense knowledge. Extending KG with hyperbolic RotH embeddings (Baseline 4) does not improve much compared to Baseline 3, and retrieval-only KGE ablations perform far below COMET-based prompts. In general, our results suggest that high-quality textualized causal knowledge and strong discriminative MCQA backbones remain the primary drivers of AER performance, while more complicated KG geometries offer limited gains under current training budgets.

## References


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

Pengfei Cao, Yubo Chen, Mingxuan Yang, Chenlong Zhang, Mingxuan Liu, Kang Liu, and Jun Zhao. 2026. Semeval-2026 task 12: Abductive event reasoning: Towards real-world event causal inference for large language models.

Ines Chami, Adva Wolf, Da-Cheng Juan, Frederic Sala, Sujith Ravi, and Christopher Ré. 2020. Low-dimensional hyperbolic knowledge graph embeddings. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 6901–6914.

Haoang Chi, He Li, Wenjing Yang, Feng Liu, Long Lan, Xiaoguang Ren, Tongliang Liu, and Bo Han. 2024. Unveiling causal reasoning in large language models: Reality or mirage? In *Advances in Neural Information Processing Systems*.

Emre Kiciman, Robert Osazuwa Ness, Amit Sharma, and Chenhao Tan. 2024. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*.

Junming Liu, Siyuan Meng, Yanting Gao, Song Mao, Pinlong Cai, Guohang Yan, Yirong Chen, Zilin Bian, Ding Wang, and Botian Shi. 2025a. Aligning vision to language: Annotation-free multimodal knowledge graph construction for enhanced llms reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 981–992.

Xiaoyu Liu, Paiheng Xu, Junda Wu, Jiaxin Yuan, Yifan Yang, Yuhang Zhou, Fuxiao Liu, Tianrui Guan, Haoliang Wang, Tong Yu, Julian McAuley, Wei Ai, and Furong Huang. 2025b. Large language models and causal inference in collaboration: A comprehensive survey. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6406–6443, Albuquerque, New Mexico. Association for Computational Linguistics.

Martina Miliani, Serena Auriemma, Alessandro Bondielli, Emmanuele Chersoni, Lucia Passaro, Irene Sucameli, and Alessandro Lenci. 2025. Explica: Evaluating explicit causal reasoning in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17335–17355, Vienna, Austria. Association for Computational Linguistics.

Zequn Sun, Muhao Chen, Wei Hu, Chengming Wang, Jian Dai, and Wei Zhang. 2020. Knowledge association with hyperbolic knowledge graph embeddings. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 5704–5716.