X-NLP at SemEval-2026 Task 12: Prompting LLMs for Abductive Event Reasoning

Caelen Mattie and **Patrick Bowen** and **Milton King**

St. Francis Xavier University
Antigonish, NS, Canada

Abstract

In this work, we applied two different systems to the SemEval 2026 Shared Task 12 (SemEval-2026 Task 12 Organizers, 2026), which explores abductive event reasoning. Specifically, this task involves determining the cause of an event from a list of candidate causes. Instances are accompanied with documents that can provide applicable knowledge for the target event. Both of our systems involve prompting LLMs and our best performing system leverages retrieval-augmented generation. Our best performing system achieved a score of 84% and ranked 40th out of 221 total submissions.

1 Introduction

In recent years, Large Language Models (LLMs) have achieved remarkable performance in descriptive and predictive tasks, such as event extraction (Ye et al., 2024) and future-state forecasting (Ye et al., 2024). However, state-of-the-art LLMs have struggled to excel in the domain of abductive event reasoning (AER) (Bhagavatula et al., 2019), as they are highly susceptible to semantic distraction (Shi et al., 2023).

To help address these limitations, we participated in the AER shared task 12 at SemEval 2026 (SemEval-2026 Task 12 Organizers, 2026). This task challenges models to move beyond simple summarization, and explicit reasoning as to why a particular event occurred. Provided with a specific event, such as "Cryptocurrency Market Prices Soar", along with a set of available documents, a model is tasked with identifying the most plausible and direct cause(s), such as "Government announcement of a national cryptocurrency reserve." Here we present our systems that include LLMs, prompt engineering, and retrieval-augmented generation (RAG) to accomplish this task. Our best performing system ranked 40th out of 221 submissions[1].

[1]10 submissions received a score of 0.

2 Related work

Abductive reasoning in NLP involves inferring the most plausible explanation or cause for an observed event or outcome, often from incomplete evidence (Bhagavatula et al., 2019).

This form of reasoning, distinct from deductive or inductive approaches, has been explored in commonsense inference tasks, such as the abductive natural language Inference (αNLI) dataset. This is used in cases where models must select the best hypothesis to bridge two observations (Bhagavatula et al., 2019).

Talmor et al. (2019) introduced CommonsenseQA, a multiple-choice question-answering dataset requiring abductive reasoning over implicit commonsense knowledge derived from ConceptNet, where models must infer the most plausible answer among semantically related distractors.

Our work builds on these; extending abductive reasoning to more complex, real-world domains as opposed to fictional commonsense scenarios.

RAG has emerged as a key technique to enhance LLMs' abilities in reasoning tasks by incorporating external context (Lewis et al., 2020). Liu et al. (2024) identify issues such as the "lost in the middle" phenomenon, where models perform poorly with excessive context documents; informing our choice of limiting context document retrieval to k = 1 in System 1.

Recent studies have integrated abductive reasoning with RAG to enhance causal inference in real-world textual data, such as news articles. Lin (2025) proposes a framework that generates and validates missing premises using RAG, improving LLM robustness in event-based causal tasks like those in AER. This differs from our work in that it generates missing premises directly, instead of leaving it to the LLM to infer.

Qin et al. (2025) propose a framework utilising knowledge graph driven RAG for false premise

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2281–2289

detection and hallucination mitigation in commonsense reasoning. This addresses the issue of LLMs generating inferences from false premises. The study primarily targets single-turn query verification rather than document-based abductive reasoning over real-world news sources.

Embedding models like Qwen3 (Zhang et al., 2025) have been used for efficient similarity-based retrieval in other zero-shot settings (Reimers and Gurevych, 2019). Both of our approaches advance this line of work by applying LLMs with zero-shot RAG to AER, without fine-tuning.

The inclusion of *persona prompting*, where an LLM is instructed to act as a certain role as part of a prompt, is recommended by official documentations of many LLMs [2345] However, this technique has seen limited research with regards to its effect on performance (Basil et al., 2025). Incorporating persona prompting has been shown to not reliably increase accuracy when answering factual questions in some cases (Zheng et al., 2024), yet it also outperforms some zero-shot baselines across twelve widely-used benchmarks (Kong et al., 2024) and produced statistically significant improvements in prediction accuracy for subjective NLP tasks (Hu and Collier, 2024). It has been shown that output tone is sensitive to persona prompting (Lutz et al., 2025), suggesting that it could indirectly encourage analytical depth and Chain-of-Thought (CoT) style thinking.

3 Dataset

This work utilizes the official dataset from SemEval-2026 Task 12: AER (SemEval-2026 Task 12 Organizers, 2026). It contains instances of real world events. Events are sourced from various online English-language news outlets. The multiple choice answer structure is designed to test abductive event reasoning. The dataset is split into sample, train, dev, and test sets. Each set has corresponding *questions* and *docs* files. The *questions* files include 200, 1819, 400, and 612 non-overlapping instances respectively.

In *questions*, each instance includes:

1. an identifier (topic id) that uniquely identifies the topic
2. an identifier (id) that uniquely identifies the question
3. a short target event description
4. four candidate explanations as natural language options (A, B, C, D), where one or more may be correct. In the test set, one option is always “None of the others are correct causes"
5. the gold answer(s) validated by human annotators

The documents here are intended to be used as context by the model; providing information about plausible causes for an event. In *docs*, each instance includes:

1. an id corresponding to the topic
2. a short text description of the topic (distinct from the target event)
3. a subsection containing 20+ related documents for each topic, each with:
   (a) an article title
   (b) an online link to the original source
   (c) a short excerpt from the text
   (d) the media outlet which released the article
   (e) the entire article content
   (f) a compressed JPEG image in base64 format

The evaluation measure for the task is classification accuracy. This is measured through a points system in which a partial match (one or more correctly selected candidate explanations for an event) yields 0.5 points, and an exact match yields 1 point. Points are added together then divided by the number of questions to get an accuracy score.

4 System overview

In this section we describe our two systems (three submissions[6]) that were for Task 12 (SemEval-2026 Task 12 Organizers, 2026). Two submissions were from two very similarly structured systems, therefore, we discuss them together as one system with two different variations (2a and 2b).

[2]https://developers.openai.com/api/docs/guides/prompt-engineering

[3]https://docs.cloud.google.com/vertex-ai/generative-ai/docs/learn/prompts/prompt-design-strategies

[4]https://docs.cloud.google.com/vertex-ai/generative-ai/docs/learn/prompts/system-instruction-introduction

[5]https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/system-prompts

[6]Our team made four submissions, but one submission resulted in accuracy of 0 due to a submission error.

Both systems use a combination of a remote LLM, prompt engineering, and RAG to maximise the number of correct guesses on the abductive reasoning task. Importantly, all input data used by our systems was sourced from the documents provided for the task. The LLMs did not have internet search access at inference time.

4.1 System 1

The system works by first passing the query to an embedding model. The query's embedding is compared against the embeddings of document titles in the matching topic. The document with the highest query-title similarity (as measured by cosine similarity) is then retrieved. The query, selected document, and answer options are then passed to the LLM for processing.

This system utilizes Grok-4.1-Fast as its underlying Large Language Model (LLM). This specific variant was selected because it demonstrated superior performance and efficiency during our testing phase. Furthermore, Grok-4.1 models rank highly on the LM Arena Text and Search leaderboards[7] (Chiang et al., 2024), a public, blind-evaluation platform where users vote for the most effective model responses.

To ensure consistent output structure and facilitate easy parsing, the LLM was specifically instructed to format its final answer as a single line, with the response enclosed in quotation marks and separated from the rest of the output by a blank line. The LLM was instructed that it would get 1 full point for an exact match and 0.5 points for a partial match. If the LLM response cannot be parsed (e.g. in case of an instruction following error), the system defaults to selecting all answers.

The hyperparameters for this system include whether or not to use CoT prompting, the number of relevant documents to provide to the LLM for context, and the size of the model used for embedding questions and document titles.

Qwen3-Embedding-0.6B and Qwen3-Embedding-8B (Zhang et al., 2025) were tested for the local title and question embedding. Initial results showed that the larger model had a moderate positive impact on the quality of the LLM's final answers, likely due to enhanced embedding quality and semantic understanding. Qwen3-Embedding-8B was preferred despite requiring more processing time than the smaller embedding model.

The Qwen3-Embedding-8B local embedding model (using Ollama[8] on localhost) together with a remote LLM (Grok-4.1-Fast via NanoGPT[9]) was used as the final configuration when processing the text data present in each subset tested.

A diagram of System 1 is displayed in Figure 1 of the Appendix.

4.2 System 2a and 2b

Systems 2a and 2b use simple persona prompting and provide the whole topic document set to LLMs in the Gemini-2.5 family (Comanici et al., 2025). System 2a uses Gemini-2.5-Flash and System 2b uses Gemini-2.5-Pro.

The system prompt consists of four sections. It begins with the statement that the LLM is an expert analyst, and is instructed to analyze the following documents and answer the question. Secondly, all documents that matched the topic of the given question were appended. Thirdly, the question itself and the available choices for causes were presented. Finally, it included a direction to output strictly either a single letter, or letters separated by commas, that best explain the cause.

5 Experimental setup

In both of our team's systems, some split of the SemEval Task 12 dataset was used for validation and testing purposes. The official final scores were computed by the task organizers based on model predictions submitted for the held-out test set instances.

Fine-tuning was not performed directly on the weights of either the embedding model or Grok-4.1-Fast for System 1, nor Gemini-2.5-Flash or Gemini-2.5-Pro for Systems 2a and 2b. The approaches operate in a zero-shot manner and rely entirely on RAG and the LLMs' inherent reasoning abilities.

Since exhaustively testing all models with all possible hyperparameters would not be feasible, experiments on System 1 followed a coarse-to-fine approach, where a broad set of LLMs are tested initially and only the most promising candidates are experimented with further. To determine the best configuration for the system, evaluations were done on CoT prompting, the number of relevant

[7] Leaderboard available at https://arena.ai/leaderboard

[8] https://ollama.com

[9] https://nano-gpt.com/api

documents provided (including zero), the embedding model size, and providing a text description of the evaluation procedure.

In all cases, the sample subset was used for initial system testing, and the dev subset was used to validate the final system configuration prior to formal evaluation (test set).

The following language models were evaluated on sample, but not implemented in the final version of System 1 due to poor or noncompetitive performance:

- Mistral-Nemo-12B-Instruct-2407
- Meta-Llama-3-1-8B-Instruct-FP8
- gemma-3-27b-it
- qwen3-coder
- grok-4-fast
- gpt-oss-20b
- llama-3.2-3b-instruct
- gemini-2.5-flash-lite-preview-09-2025
- gpt-5-nano
- gemini-2.5-flash-lite-preview-09-2025-thinking
- llama-3.1-8b-instruct
- gemma-3-4b-it
- azure-gpt-4o-mini

Table 1 shows an ablation study comparing our baseline model against various configurations (dev set), where each variant represents the removal or alteration of one specific component. Accuracy is calculated using our evaluation script based on the SemEval-2026 Task 12 description [10]. These findings were used to inform our choices for each component included.

It was found that providing only the most relevant document ($k = 1$) yielded the best predictive accuracy on dev, indicating the LLM may get overwhelmed with data when provided with multiple semantically similar documents, which is consistent with prior findings (Liu et al., 2024). This may also imply the method of sourcing a reference document by comparing the similarity of the title to the original question is effective when there are not many documents in the set that have high semantic similarity (where a document is considered only to be the title concatenated with the article content).

Table 1: Accuracy (Acc) comparison across different LLM configurations in the ablation study.

Component Modified	**Acc (%)**
Submitted Configuration	74.9
Chain-of-Thought Enabled	74.1
Remove Evaluation Description	53.0
Small Embedding Model	65.1
RAG Disabled ($k = 0$)	63.3
2 Context Documents Provided	63.6
4 Context Documents Provided	60.2

In designing the retrieval component, we utilise a single-stage bi-encoder architecture rather than a multi-stage pipeline. This decision was driven by a few targeted methodological objectives. We first aimed to minimize semantic noise. Increasing the context volume from $k = 1$ to $k = 2$ or $k = 4$ resulted in repeated performance degradation observed in the experiments. Titles were chosen as the main object of retrieval because they serve as concise, high-level semantic summaries of document content (Li et al., 2010), making them efficient proxy targets for similarity matching against the given queries. This architecture also seeks to prioritise computational efficiency; by leveraging title-based embeddings, we ensure a scalable, low-latency pipeline that demonstrates strong reasoning without the overhead required for full multi-document processing.

System 1's final configuration does not include CoT prompting, includes one relevant document, uses the large embedding model (Qwen3-8b), and provides the LLM a text description of the evaluation procedure. This combination yielded the best accuracies on the dev subset.

Experiments on System 2a were conducted, where various formats of persona prompts were evaluated on the dev subset. The best performing prompt was used for both System 2a and System 2b submissions. System 2b differs from System 2a only in its use of Gemini-2.5-Pro rather than Gemini-2.5-Flash as the underlying LLM, and was

[10] Due to a technical oversight, the evaluation metric used to tune the models assigns partial points if the LLM's response includes any of the correct answers, which differs from the metric used on the test set. The scores recorded for the test set were calculated by the official metric from the SemEval organizers.

submitted after System 2a had been evaluated on the test set.

Both systems rely on a subset of the libraries and tools in the following list, which includes their corresponding usage:

- **Ollama** [11] — local inference server running Qwen3-8B for embedding generation
- **scikit-learn** (Pedregosa et al., 2011) — cosine similarity computation between embeddings
- **NLTK** (Bird, 2006) — basic text tokenization and whitespace normalization
- **NanoGPT API**[12] — remote inference with Grok-4.1-Fast
- **Google Gemini API** (Comanici et al., 2025) — remote inference with Gemini-2.5-Flash and Gemini-2.5-Pro

6 Results

6.1 Performance

System 1 achieved the strongest performance of the two systems our team submitted to the shared evaluation as determined by the point-based accuracy metric. On the unseen test set, System 1 received a score of 84%, System 2a scored 67%, and System 2b scored 71%. By comparison, the highest scoring submission in the competition achieved 95% accuracy. System 1 ranked 40th out of 221 total submissions, placing it in the second decile. Half of the submissions achieved an accuracy greater than 64%, with relatively few (approximately 10%) scoring between 20 and 50 percent.

6.2 Quantitative analysis

Tables 2 and 3 show summary statistics of the final submission results from all teams to the shared task. We observe a strong negative skew in the data, with clustering around higher accuracy scores. The mean is observed to be lower than the median, pulled down by many failures and near-failures. Given this data, we believe the task shows room for improvement in robustness and reliability.

6.3 Error analysis

Each instance requires selecting any of four unique natural language answer options. Given the evaluation focuses on overall selection accuracy, class confusions would not be informative. Instead, we show instances from the dev set where our systems predict incorrectly.

Table 2: Summary statistics of all submission scores on test.

Mean	SD	Median	IQR	Min	Max	Skew
0.540	0.312	0.640	0.510	0.000	0.950	−0.563

Table 3: Summary statistics of all submission scores excluding zero results (10 observations) on test.

Mean	SD	Median	IQR	Min	Max	Skew
0.566	0.295	0.650	0.480	0.01	0.950	−0.663

6.4 Example 1, Model 1

Prediction: D | **Answer:** A,B,C

Event: On November 24, 2018, "yellow vest" protestors blocked the Champs Elysées in Paris, leading to confrontations.

A. The Macron government announced an increase in gas taxes effective January 1, 2019.

B. The Macron government announced an increase in gas taxes effective January 1, 2019.

C. The Macron government announced an increase in gas taxes effective January 1, 2019.

D. 227 people were injured in the protests, with six severely.

We believe that this question was answered incorrectly because the LLM is associating the intense wording of the event with that of option D and it does not have the necessary world knowledge to connect the protests to the true cause which was tax increases. Furthermore, the document selected by the embedding model was short (188 characters total) and similar to the target event but not relevant to it, further encouraging the system to pick the more strongly worded option.

6.5 Example 2, Model 1

Prediction: B | **Answer:** C

Event: South Korea banned seafood imports from the Fukushima area.

A. The Chinese government criticized Japan's plan to release wastewater.

B. A huge earthquake and tsunami in 2011 caused a meltdown at the Fukushima plant.

[11] https://ollama.com
[12] https://nano-gpt.com/api

C. Japan announced its plan to release treated Fukushima wastewater into the sea.

D. Fishermen, including Haruo Ono, sued the Japanese government to stop the water release.

We believe that the system predicted this incorrectly because it considered the most famous, ultimate root cause (the 2011 disaster) instead of a more recent event which was the true cause. The LLM may have focused on a salient, factually true background fact while missing the precise trigger the question was looking for. The document retrieved in this case is about China banning imports, not South Korea which is not relevant to the event in question. The context document includes the sentence, "Japan's devastating 2011 earthquake and tsunami caused water within the Fukushima nuclear plant to be contaminated with highly radioactive material." further reinforcing option B as the selection as they contain similar content.

7 Conclusion

Our systems performed well relative to other submissions on the shared task of AER. Our best performing system achieved a score of 84% as measured by accuracy on the test set. System 1 uses a combination of Grok-4.1-Fast, prompt engineering and embedding-based RAG for accurate explanation prediction, while System 2a and System 2b utilize Gemini-2.5-Flash and Gemini-2.5-Pro alongside simple persona prompting.

We hypothesize that LLMs may rely on more widely known background facts as explanations for certain events as opposed to predicting more relevant event triggers.

Future work could include improving RAG effectiveness in System 1 with the assistance of query rewriting and document re-ranking. This may provide more relevant context to the LLM. It could also consider the approach of Lin by attempting to generate missing premises explicitly. System 2 could experiment further with different personas such as professional or domain-specific roles (e.g. historian, investigative journalist, temporal logic expert). Document summaries could also be provided instead of directly using an article's content.

References

Savir Basil, Ina Shapiro, Dan Shapiro, Ethan Mollick, Lilach Mollick, and Lennart Meincke. 2025. Prompting science report 4: Playing pretend: Expert personas don't improve factual accuracy. *arXiv e-prints*, pages arXiv–2512.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. *arXiv preprint arXiv:1908.05739*.

Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 interactive presentation sessions*, pages 69–72.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv e-prints*, pages arXiv–2507.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in llm simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. Better zero-shot reasoning with role-play prompting. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4099–4113.

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*, volume 33, pages 9459–9474. Curran Associates, Inc.

Decong Li, Sujian Li, Wenjie Li, Wei Wang, and Weiguang Qu. 2010. A semi-supervised key phrase extraction approach: Learning from title phrases through a document semantic network. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 296–300, Uppsala, Sweden. Association for Computational Linguistics.

Shiyin Lin. 2025. Abductive inference in retrieval-augmented language models: Generating and validating missing premises. In *2025 5th International Conference on Network Communication and Information Security (ICNCIS)*, pages 62–66. IEEE.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173.

Marlene Lutz, Indira Sen, Georg Ahnert, Elisa Rogers, and Markus Strohmaier. 2025. The prompt makes the person (a): A systematic evaluation of sociodemographic persona prompting for large language models. *arXiv preprint arXiv:2507.16076*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Yuehan Qin, Shawn Li, Yi Nian, Xinyan Velocity Yu, Yue Zhao, and Xuezhe Ma. 2025. Don't let it hallucinate: Premise verification via retrieval-augmented logical reasoning. *arXiv preprint arXiv:2504.06438*.

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.

SemEval-2026 Task 12 Organizers. 2026. Semeval-2026 task 12: Abductive event reasoning.

Freda Shi, Xinyun Chen, Ishan Misra, Nathan Wang, Fu-Hao Huang, Luke Zettlemoyer, Wen-tau Yih, and Jiajun Pan. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158.

Zelin Ye, Mingyu Derek Ma, Yu Wang, Xinyu Zhang, and Nanyun Peng. 2024. Mirai: Evaluating llm agents for event forecasting. *arXiv preprint arXiv:2407.01234*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. 2024. When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models.

8 Appendix

8.1 System 1 Prompt

The following prompt template was used to query the language model for System 1. The fields of the prompt were filled programmatically on a per-query basis.

```
You are an expert
cause-effect analyst.
Analyze the following
documents and answer the
question.

---------------------------

Cause

---------------------------

Question:  [target_event]

---------------------------

Effect

---------------------------

Options:
A: [option_A]
B: [option_B]
C: [option_C]
D: [option_D]

---------------------------

Relevant Documents:
[retrieved_documents]

---------------------------

Note:  The documents
do not necessarily
contain the answer to
the question; they are
just potentially relevant
context.

Evaluation:  Based on
the information provided,
select ANY OF (A, B, C, or
D) that best explains the
cause.  You will receive
1 point for an exactly
correct guess (e.g., guess
A,C = answer A,C), 0.5
points for a partially
correct guess (one or more
matching letters, e.g.,
guess A but correct was A,
B, C), and 0 points for no
match.

---------------------------
```

Requirement: After completing your reasoning, you must output the final answer on the very last line, prefixed with "FINAL ANSWER:".

When Chain-Of-Thought (CoT) prompting was enabled, the following was inserted immediately after the retrieved documents section.

Instructions:

1. Restate the target event in your own words in one sentence.

2. For each option (A, B, C, D), do:

- Briefly explain how this option could cause the target event.
- Cite at least two specific facts or sentences from the Relevant Documents that support this option, if any.
- Point out any contradictions or missing links with the Relevant Documents.
- Give this option a plausibility score from 0 to 1.

3. Compare the four options and:

- Identify which option(s) provide the most direct, well-supported explanation with the fewest extra assumptions.
- If multiple options are plausible and not mutually exclusive, you may select more than one.

4. Think again:

- For the best option(s), briefly check if there is a strong reason they might be wrong given the documents.
- If you find a serious problem, adjust your choice.

8.2 System 2a and 2b Prompt

The following prompt template was used to query the language model for System 2a and 2b. The fields of the prompt were filled programmatically on a per-query basis.

You are an expert analyst. Analyze the following documents and answer the question.

Contextual Documents:

[topic_documents]

Question: *[target_event]*

Options:
A: *[option_A]*
B: *[option_B]*
C: *[option_C]*
D: *[option_D]*

Based on the documents, select the letter(s) (A, B, C, or D) that best explains the cause. Output strictly either the single letter, or letters separated by commas (e.g. "A", "A,C,D"). Do not output any other text.

8.3 Visualization of System 1 Pipeline

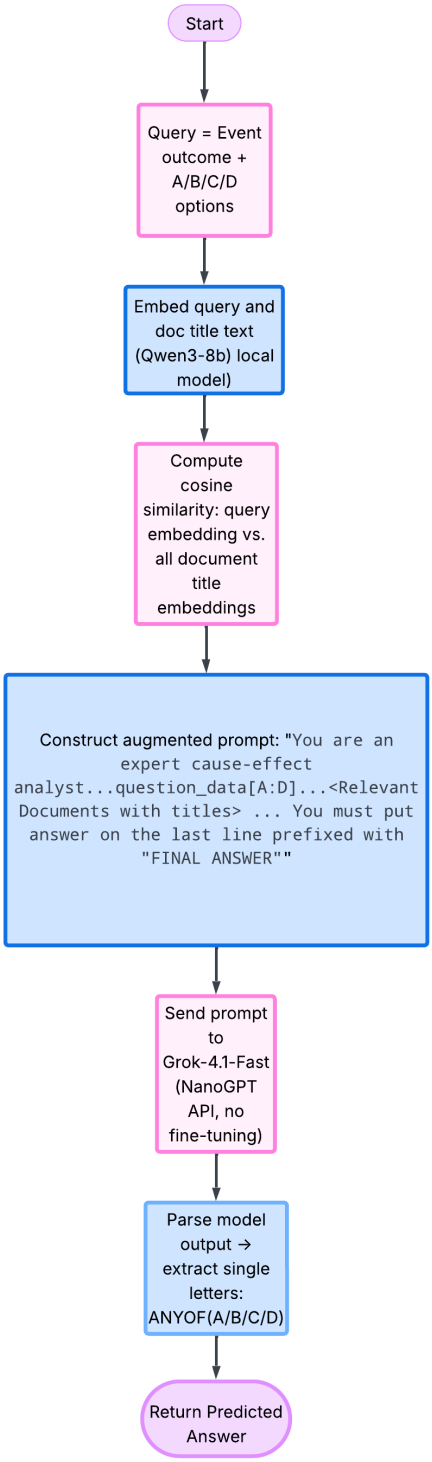

Figure 1: Pipeline of question data to predicted answer