

Causal Connections: Leveraging Multilingual Fine-Tuning for Financial QA@FinCausal 2026

Akash Kumar Gautam, Serhii Hamotskyi, Christian Hänig

Anhalt University of Applied Sciences

{akash-kumar.gautam, serhii.hamotskyi, christian.haenig}@hs-anhalt.de

Abstract

This paper describes team HSA_CORAL's submission to the FinCausal 2026 shared task on extracting cause–effect relations from financial narratives via extractive question answering in English and Spanish. We compare three modeling families: (i) encoder-only token tagging with multilingual BERT, (ii) encoder–decoder generation with multilingual BART, and (iii) decoder-only LLMs (Llama 3.1 and GPT variants) using prompt refinement, few-shot demonstrations, and supervised fine-tuning. Across settings, prompting and few-shot examples yield competitive performance, while supervised fine-tuning provides the largest gains. Our best system, GPT-4.1 Mini fine-tuned on combined English and Spanish training data, achieves a tied highest score on the English subtask (score 4.8140) and ranks third on Spanish (score 4.7753) under the shared task's LLM-as-a-judge metric. Overall, the results highlight the value of task-specific adaptation and multilingual fine-tuning for cross-lingual transfer in financial causality QA.

Keywords: Large Language Models, Financial NLP, Financial Question Answering.

1. Introduction

Understanding cause–effect relationships in financial texts is essential for informed decision-making. They reveal drivers of stock prices, economic changes, market behavior, and regulatory decisions across different countries. For analysts and investors, identifying causal connections provides valuable insights into financial risks, potential investments, and strategic planning (Gopalakrishnan et al., 2023).

The FinCausal shared task has steadily advanced how we detect causality in finance. Earlier editions focused on identifying causal phrases directly in text. Later versions introduced more complex challenges, such as recognizing implicit causality and multi-step reasoning (Mariko et al., 2022, 2020; Zavitsanos et al., 2023). The most recent task required generative models to answer open-ended questions about causes and effects, using Exact Match (EM) and Semantic Answer Similarity (SAS) (Risch et al., 2021) as evaluation metrics (Moreno-Sandoval et al., 2025).

The 2026 edition introduces several new elements (Moreno-Sandoval et al., 2026). The dataset includes fragments with new complex causal relationships, rephrased questions that demand more sophisticated reasoning, and randomly divided train-test splits. A novel challenge is the use of an LLM-as-a-judge for evaluation, which rates responses on a 1–5 adequacy scale. The task pushes models beyond simple text matching toward understanding both explicit and implicit causal relationships in detailed financial narratives.

We explored three approaches to extractive question-answering: (i) token classification with encoder-only models like BERT, (ii) sequence-to-sequence generation with encoder-decoder models like BART, and (iii) few-shot prompting with decoder-only large language models.

Across our experiments, we find that fine-tuned generative models consistently outperform encoder-only and encoder-decoder architectures on the extractive QA task. Notably, GPT-4.1 Mini fine-tuned on a multilingual corpus achieves the best overall performance, securing the top position on the English subtask and third place on Spanish. Adding 20 few-shot examples proves critical, enabling even a smaller model to surpass its larger counterparts in zero-shot settings. These results demonstrate that multilingual fine-tuning and few-shot learning together provide a reliable approach for causal relation extraction in financial documents.

2. Related Work

Causal Information Extraction Early approaches to causal relationship detection in financial texts relied heavily on rule-based systems and traditional machine learning methods such as Support Vector Machines (SVMs) and decision trees (Ghosh and Naskar, 2022; Baranes et al., 2019). While these models demonstrated some success in identifying patterns in financial reports and news articles, they required extensive feature engineering and often struggled to capture the complexity and temporal dynamics of causal relations in domain-specific documents.

The introduction of BERT (Devlin et al., 2019) and its multilingual variants marked a significant shift in the field (Yang et al., 2019; Wan and Li, 2022). These pre-trained language models enabled more nuanced understanding of contextual information with minimal feature extraction, substantially improving performance on tasks involving document-level comprehension (Zhang and Jankowski, 2022).

Subsequent work has focused on fine-tuning these models on domain-specific financial datasets, achieving state-of-the-art results in tasks such as sentiment analysis and event extraction (Mariko et al., 2020). Fine-tuned pre-trained language models have consistently outperformed traditional machine learning approaches, particularly when working with large-scale financial reports (Jin et al., 2023; Huang et al., 2023; Sarmah et al., 2023). Liu et al. (2023) propose an implicit cause-effect interaction framework to improve event causality extraction.

Pretrained Language Models More recently, proprietary large language models such as GPT-4 have been explored for question-answering tasks in the financial domain (Zhang et al., 2023; Kalpakchi and Boye, 2023). These models have demonstrated strong few-shot learning capabilities, enabling them to generalize from limited examples without task-specific fine-tuning (Xiao et al., 2022; Guo et al., 2023). This line of work underscores the growing potential of generative models for specialized NLP tasks, including the extraction of causal relationships in finance (Nayak et al., 2022; Kim et al., 2023).

3. Methodology

We compare three approaches to extractive QA for FinCausal 2026: (1) token classification using BERT-based models, (2) extractive QA using encoder-decoder models, and (3) generative models. These approaches are compared across a variety of pretrained large language models in few-shot settings using a multilingual dataset.

3.1. Encoder-Based Extractive QA

This approach utilizes text embedding models such as BERT for token classification, following a similar methodology to Yoon et al. (2022). The process begins by tokenizing both the context and the question, which are then concatenated with a special [SEP] token. During training, each sample is annotated using an IO tagging scheme, where

the answer span is mapped to its first occurrence within the passage.

We compute the cross-entropy loss between the predicted token classes and the ground truth labels derived from the training data. To refine the loss calculation, a loss mask is applied to restrict the computation exclusively to tokens originating from the passage. This masking strategy excludes tokens from the question and special symbols (e.g., [SEP], padding). This focuses learning on the passage tokens.

3.2. Encoder-Decoder Extractive QA

Our second approach treats extractive QA as a sequence-to-sequence generation task using BART (Lewis et al., 2020). The input consists of the question and context concatenated with a delimiter, and the model is trained to generate the answer token-by-token.

We fine-tune BART by minimizing the negative log-likelihood of the target answer sequence. During inference, beam search is employed to decode the most likely answer. This formulation is effective for extractive tasks as it leverages BART’s pre-trained language understanding while flexibly handling answer spans. We evaluate multilingual BART variants on both the English and Spanish subtasks.

3.3. Decoder-Based Extractive QA

The flexibility of large language models makes them well-suited for extractive QA tasks. However, it is important to ensure that models follow instructions and avoid hallucinations beyond the provided context (Xu et al., 2024). To address this, we implement a multi-step strategy combining prompt optimization, few-shot selection, and fine-tuning.

First, we perform prompt optimization through several iterative refinements on a small subset of the training dataset. The final version of the prompt used is described in Appendix A. This process helps identify instruction formats and phrasings that consistently yield extractive, context-grounded answers. We incorporate few-shot examples within the prompt, selecting relevant QA demonstrations using cosine similarity. Specifically, given a test context and question, we retrieve the most similar QA pairs from the training set based on multilingual embedding similarity¹, and include them as examples in the prompt to guide the model’s output.

¹<https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2>

Model	Fine-tuned	Corpus	English	Spanish
BERT Base Multilingual	yes	Multilingual (en+es)	3.9800	3.9810
BART Facebook Base	yes	Multilingual (en+es)	4.1200	4.0300
Llama-3.1 8B	yes	Multilingual (en+es)	4.0200	3.9100
GPT-3.5 turbo	no	-	4.7040	4.7060
GPT-4.1 mini	yes	Monolingual (en)	4.7560	4.7141
GPT-4.1 mini	yes	Monolingual (es)	4.7210	4.7674
GPT-4.1 mini	yes	Multilingual (en+es)	4.8140	4.7753
GPT-5.2	no	-	4.7600	4.7350

Table 1: Evaluation results on blinded English and Spanish test sets, with scores provided by an external LLM judge. The *Corpus* column indicates whether models were fine-tuned on monolingual or multilingual data. Bold entries denote the highest score in each language column. Decoder-only models were tested with varying numbers of few-shot examples; results, obtained with 20 examples, are reported here. Best results using our approach are presented in **bold**.



Figure 1: External LLM-judge score (English subtask) as a function of the number of few-shot demonstrations in the prompt for the best-performing decoder-only configuration.

As a further enhancement, we fine-tune the model on up to 2,000 samples, experimenting with three configurations: training on English data only, Spanish data only, and a combined bilingual dataset. Fine-tuning reinforces task-specific behavior and significantly reduces hallucinations, leading to more reliable extractive answers.

This hybrid approach—combining prompt engineering, dynamic few-shot selection, and targeted fine-tuning—allows us to leverage the generative strength of LLMs while maintaining the extractive precision required for the task.

4. Experiments

4.1. Dataset

The dataset comprises financial narratives in English and Spanish designed for an extractive question answering task focused on causal relationships [Moreno-Sandoval et al. \(2026\)](#). Given a context and a question, the task requires identifying a text span that expresses a causal relation within the financial narrative ([Moreno-Sandoval et al., 2025](#)).

Questions are formulated abstractly, targeting either the cause or the effect described in the text, with causes defined as agents or facts that can be extracted verbatim from the provided context.

For complex causal structures—such as causal chains or non-linear relationships—up to two questions are included per context to ensure comprehensive coverage. The English dataset is sourced from financial annual reports from 2017, drawn from the UCREL corpus and the English portion of the 2018 FinT-esp corpus, while the Spanish dataset is extracted from a corpus of Spanish financial annual reports spanning 2014 to 2018. The training set for both subtasks consists of 2,000 samples, with test sets of 500 samples for English and 503 samples for Spanish, respectively.

4.2. Model Selection

BERT² was used as an encoder model, multilingual BART³ for encoder-decoder model and we evaluated Llama-3.1 and several GPT⁴ variants for

²<https://huggingface.co/google-bert/bert-base-multilingual-cased>

³<https://huggingface.co/facebook/bart-base>

⁴<https://developers.openai.com/api/docs/models>

decoder-only setups. For fine-tuning of Llama3.1 we used Low-Rank Adaptation (LoRA) to speed the process (Hu et al., 2022).

4.3. Results

Table 1 reports the best-performing configuration for each model family. Scores correspond to the shared task’s external LLM-as-a-judge evaluation on the blind test set, rated on a 1–5 adequacy scale for both English and Spanish.

Encoder and Encoder-Decoder Setups. Extractive models such as BERT perform reasonably well at identifying exact spans corresponding to causal relationships, particularly when the answer is explicitly stated in the context. However, BART consistently outperforms BERT across both languages, benefiting from its sequence-to-sequence formulation which allows for more flexible and accurate span generation. Interestingly, the multilingual variant of BERT fine-tuned on combined English and Spanish data yields better performance on the blinded test set than its monolingual counterparts trained on individual languages. For brevity, we report only the best-performing configurations.

Few-Shot Learning with Decoder-Only Models

In our decoder-only experiments, we evaluated Llama 3.1 and several variants of GPT. For fine-tuning Llama 3.1, we employed Low-Rank Adaptation (LoRA) to reduce computational overhead while maintaining task performance. Our findings indicate that while well-structured prompts with clear instructions are effective at reducing hallucinations, the inclusion of few-shot examples substantially improves the quality and precision of the generated answer spans. Notably, adding examples from the training set resulted in substantial improvements for GPT-4.1 Mini compared to zero-shot settings—even enabling it to outperform a much larger model (GPT-5.2⁵) on both Spanish and English subtasks.

Figure 1 plots the LLM-as-a-judge scores against the increasing number of few-shot examples included in the prompt. The trend shows a clear improvement in output quality as more examples are added, with scores increasing correspondingly on the leaderboard. However, increasing the number of examples beyond a certain point did not yield further gains; in some cases, it even led to the generation of hallucinated content.

Given the context window constraints of each model, we systematically varied the number of few-

shot examples. The best results across all decoder-only models were achieved with 20 few-shot examples, which we adopt as our final configuration.

Fine-Tuning Our experiments reveal a clear advantage of fine-tuned generative models over both encoder-only and encoder-decoder architectures. Multilingual fine-tuning consistently outperforms monolingual fine-tuning on both English and Spanish subtasks, demonstrating strong cross-lingual transfer for causal relation extraction in financial narratives. Among all configurations, GPT-4.1 Mini with multilingual fine-tuning achieved the best results across both languages.

Notably, this fine-tuned model outperforms a much larger model, GPT-5.2, in zero-shot settings. This suggests that the test dataset’s contexts and questions contain lexical characteristics and causal relationships (for both Spanish and English) that can be reliably identified only when models have access to task-specific annotation guidelines through fine-tuning. The observation points to an interesting conclusion: fine-tuned pre-trained language models of moderate size may offer better performance for specialized tasks than much larger models used in inference-only mode, highlighting the importance of fine-tuning over parameter scaling for domain-specific applications.

Error Analysis and Limitations While our models achieved strong overall results, error analysis reveals persistent errors in handling nested causal structures and contexts containing multiple potential causes. In these challenging cases, models often selected the wrong causal pair, with errors more frequent in the Spanish subtask—suggesting that cross-lingual generalization remains challenging for complex causal structures.

The lack of detailed annotation guidelines for identifying relevant spans further compounds this issue, as prompt optimization alone cannot fully resolve such ambiguities. Looking ahead, alignment techniques such as Direct Preference Optimization (Rafailov et al., 2023) and Reinforcement Learning from Human Feedback (Bai et al., 2022) offer promising ways for learning the implicit patterns underlying human annotations for structurally complex examples.

5. Conclusion

This work demonstrates that generative models outperform encoder and encoder-decoder architectures on causal question-answering tasks for

⁵OpenAI currently supports supervised fine-tuning for GPT models up to GPT-4.1-2025-04-14.

financial narrative documents, provided that hallucinations are mitigated through careful prompt engineering and few-shot examples. Multilingual fine-tuning with a large number of training samples from both Spanish and English, significantly boosts performance, highlighting the effectiveness of cross-lingual transfer in this domain. Future work may explore LLM-based evaluation as a means to further enhance output quality, as well as more sophisticated strategies for handling nested and ambiguous causal structures.

Acknowledgments

This work has been completed as part of project CORAL (Constrained Retrieval-Augmented Language Models), funded by the German Federal Ministry of Research, Technology, and Space (BMFTR) under grant number 16IS24077C.

6. References

- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.
- Amos Baranes, Rimona Palas, et al. 2019. Earning movement prediction using machine learning-support vector machines (svm). *Journal of Management Information and Decision Sciences*, 22(2):36–53.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186.
- Sohom Ghosh and Sudip Kumar Naskar. 2022. Lipi at fincausal 2022: Mining causes and effects from financial texts. In *Proceedings of the 4th financial narrative processing workshop@ LREC2022*, pages 121–123.
- Seethalakshmi Gopalakrishnan, Victor Zitian Chen, Wenwen Dou, Gus Hahn-Powell, Sreekar Neddunuri, and Wlodek Zadrozny. 2023. Text to causal knowledge graph: A framework to synthesize knowledge from unstructured business texts into causal graphs. *Information*, 14(7):367.
- Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.
- Allen H Huang, Hui Wang, and Yi Yang. 2023. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841.
- Yiqiao Jin, Xiting Wang, Yaru Hao, Yizhou Sun, and Xing Xie. 2023. Prototypical fine-tuning: Towards robust performance under varying data sizes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12968–12976.
- Dmytro Kalpakchi and Johan Boye. 2023. Quasi: a synthetic question-answering dataset in swedish using gpt-3 and zero-shot learning. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 477–491.
- Yuheun Kim, Lu Guo, Bei Yu, and Yingya Li. 2023. Can chatgpt understand causal language in science claims? In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 379–389.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7871–7880.
- Jintao Liu, Zequn Zhang, Kaiwen Wei, Zhi Guo, Xian Sun, Li Jin, and Xiaoyu Li. 2023. Event causality extraction via implicit cause-effect interactions. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 6792–6804.
- Dominique Mariko, Hanna Abi-Akl, Estelle Labidurie, Stephane Durfort, Hugues De Mazancourt, and Mahmoud El-Haj. 2020. The financial document causality detection shared task (fincausal 2020). In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 23–32.
- Dominique Mariko, Hanna Abi-Akl, Kim Trottier, and Mahmoud El-Haj. 2022. The financial causality

- extraction shared task (fincausal 2022). In *Proceedings of the 4th Financial Narrative Processing Workshop@ LREC2022*, pages 105–107.
- Antonio Moreno-Sandoval, Blanca Carbajo-Coronado, Jordi Porta Zamorano, Yanco Amor Torterolo Orta, and Doaa Samy. 2025. The financial document causality detection shared task (fincausal 2025). In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 214–221.
- Antonio Moreno-Sandoval, Jordi Porta, Yanco Torterolo, Alexia Stanescu, Melina Chatzi, and Sofía Roseti. 2026. The Financial Document Causality Detection Shared Task (FinCausal 2026). In *Proceedings of the 7th Financial Narrative Processing Workshop (FNP 2026) at LREC 2026*, Palma de Mallorca, Spain. ELRA.
- Tapas Nayak, Soumya Sharma, Yash Butala, Koustuv Dasgupta, Pawan Goyal, and Niloy Ganguly. 2022. A generative approach for financial causality extraction. In *Companion Proceedings of the Web Conference 2022*, pages 576–578.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.
- Julian Risch, Timo Möller, Julian Gutsch, and Malte Pietsch. 2021. Semantic answer similarity for evaluating question answering models. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 149–157.
- Bhaskarjit Sarmah, Dhagash Mehta, Stefano Pasquali, and Tianjie Zhu. 2023. Towards reducing hallucination in extracting information from financial reports using large language models. In *Proceedings of the Third International Conference on AI-ML Systems*, pages 1–5.
- Chang-Xuan Wan and Bo Li. 2022. Financial causal sentence recognition based on bert-cnn text classification. *The Journal of Supercomputing*, 78(5):6503–6527.
- Jiayu Xiao, Liang Li, Chaofei Wang, Zheng-Jun Zha, and Qingming Huang. 2022. Few shot generative model adaption via relaxed structural alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11204–11213.
- Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.
- Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with bertserini. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (Demonstrations)*, pages 72–77.
- Wonjin Yoon, Richard Jackson, Aron Lagerberg, and Jaewoo Kang. 2022. Sequence tagging for biomedical extractive question answering. *Bioinformatics*, 38(15):3794–3801.
- Elias Zavitsanos, Aris Kosmopoulos, George Giannakopoulos, Marina Litvak, Blanca Carbajo-Coronado, Antonio Moreno-Sandoval, and Mo El-Haj. 2023. The financial narrative summarisation shared task (fns 2023). In *2023 IEEE International Conference on Big Data, BigData 2023*, pages 2890–2896. Institute of Electrical and Electronics Engineers.
- Le Zhang, Yihong Wu, Fengran Mo, Jian-Yun Nie, and Aishwarya Agrawal. 2023. Moqagpt: Zero-shot multi-modal open-domain question answering with large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1195–1210.
- Ning Zhang and Maciej Jankowski. 2022. Hierarchical bert for medical document understanding. *arXiv preprint arXiv:2204.09600*.

7. Language Resource References

- Antonio Moreno-Sandoval, Yanco Amor Torterolo Orta, Maria Alexia Stanescu, and Melina Chatzi. 2026. [The Financial Document Causality Detection Shared Task \(FinCausal 2026\): Dataset](#).

A. Prompt

Figure 2 describes the prompt used by decoder only language models for generating the final response as part of the extractive question-answering task.

Task Description

Given a financial context and a question, extract an exact answer from the context about cause or effect that addresses the question. The answer will be either the cause or the effect to a specific event mentioned in the context.

Instructions

- 1. Read the context carefully:**
 - Understand the events and relationship described in the context.
- 2. Understand the question:**
 - Determine if the question is asking for cause or effect.
 - Identify the specific events or statement the question refers to.
- 3. Extract the answer verbatim:**
 - Locate the exact sentence or phrase in the context that answers the question.
 - **Do not paraphrase, summarise, or add any external information. The answer must be copied word-for-word from the context.**
- 4. Provide only the answer:**
 - **Do not include any instructions, explanations or formatting.**
 - **Output only the extracted answer and nothing else.**
- 5. Answer should be in the language in which question is asked and the context is mentioned:**
 - **Understand the context very well.**

Examples

```
{ { formatted_examples } }
```

Your Task

Context:

```
{ { context } }
```

Question:

```
{ { question } }
```

Answer:

[Provide only the exact answer from the context]

Remember

- **Output only the answer. Do not include any additional text. Do not include “Answer” in your output.**
- **The answer must exactly match a portion of the context.**
- **Do not add instructions, explanations, or any extra information.**

Figure 2: Prompt used for extractive QA by decoder models used in the FinCausal 2026 shared task.