Addressing Hallucination in Causal Q&A: The Efficacy of Fine-tuning over Prompting in LLMs

Georg Niess[1], Houssam Razouk[1], Stasa Mandic[1], Roman Kern[1,2]

[1]Graz University of Technology, [2]Know-Center GmbH

Correspondence: georg.niess@tugraz.at

Abstract

This paper presents our approach and findings for participating in the FinCausal 2025 competition (Moreno-Sandoval et al., 2025), which addresses causal question answering derived from financial documents, specifically English and Spanish annual reports. We investigate the effectiveness of generative models, such as Llama, in contrast to common extractive methods like BERT-based token classification. While prompt optimization and few-shot learning offer some improvements, they were insufficient for consistently outperforming extractive methods in FinCausal, suffering from hallucinations. In contrast, fine-tuning generative models was shown to be essential for minimizing hallucinations and achieving superior performance. Using our fine-tuned multilingual model for both tasks, we outperform our extractive and monolingual approaches, achieving top results for Spanish and second-best for English in the competition. Our findings indicate that fine-tuned large language models are well-suited for causal Q&A from complex financial narratives, offering robust multilingual capabilities and effectively mitigating hallucinations.

1 Introduction

Causality extraction from financial documents is vital for knowledge-driven decision-making (Gopalakrishnan et al., 2023). Financial analysts must identify the various factors that influence performance, including economic shifts, market trends, and regulatory policies. Detecting causality enables models to interpret cause-effect relationships in complex financial events, enhancing insights into financial risks, investment opportunities, and strategic decisions.

The FinCausal shared tasks have progressively advanced causality detection in finance, evolving from span-based detection in 2020 to addressing implicit causality in 2021 and multi-step reasoning in 2022. The focus of the 2025 task transitions to generative models for causality extraction, requiring models to answer open-ended questions about causes and effects through interpretative and abstractive methods. FinCausal 2025 aims for models to interpret both explicit and implicit causal relationships, moving beyond token-level accuracy to provide coherent, contextually relevant answers.

1.1 Task formulation of extractive Q&A

Objective. Given a natural language question and a corresponding passage of text, extract a contiguous span of text from the passage that directly answers the question.

Input. ***Question:*** A natural language question posed by a user, e.g., "What is the main reason why the Group trading continues to be subdued?"; ***Context Passage:*** A passage of text that contains the answer to the question, e.g., "Overall, Group trading continues to be subdued in large part due to legacy issues."

Output. ***Extractive Answer*****:** A contiguous span of text from the passage that directly answers the question, e.g., "legacy issues".

Evaluation Metrics. ***Exact Match (EM):*** The percentage of questions for which the extracted answer exactly matches the gold-standard answer. ***Semantic Answer Similarity (SAS):*** A measure of the semantic similarity between the extracted answer and the gold-standard answer, using a metric such as cosine similarity.

1.2 FinCausal 2025 Dataset

The dataset comprises English text segments from UK financial reports from 2017 and Spanish text segments from a corpus of Spanish financial annual reports from 2014 to 2018, structured for causal relationship extraction. Each entry includes an open-ended question to identify a cause or effect, a context passage, and an extractive answer.

Proceedings of the Joint Workshop of the 9th FinNLP, the 6th FNP, and the 1st LLMFinLegal, pages 253–258
January 19–20, 2025.

The dataset features diverse causal relationships, including explicit links with identifiable causal cues, implicit connections requiring contextual inference, and nested and enchained relations.

2 Related Work

Over the years, FinCausal tasks have progressed from extractive to generative approaches. In 2020 and 2021, models like BERT (Devlin, 2018) and RoBERTa (Liu, 2019) used token classification and BIO tagging to identify cause-effect spans, achieving high token-level accuracy (Mariko et al., 2020). In 2022, methods such as by Lyu et al. (2022) combined pre-trained models with post-processing heuristics, improving Exact Match (EM) and Semantic Answer Similarity (SAS) scores (Lyu et al., 2022). Ensemble techniques with models like SEC-BERT enhanced implicit causality detection but struggled with abstract responses.

Causal information extraction. Some examples of causal information extraction can be reviewed by Saha et al. (2022). Specifically, the authors proposed a method for predicting whether a text span corresponds to cause and effect in a given text. Next, the authors classify whether these identified cause and effect spans are linked through a causal relation. Similarly, Khetan et al. (2020) employ an event-aware language model to predict causal relations by considering event information, sentence context, and masked event context. Another significant difficulty in extracting causality is the recognition of overlapping and nested entities. In response, Lee et al. (2022) tackle overlapping entities by employing the Text-to-Text Transformer (T5). In addition, Gärber (2022) has proposed a multistage sequence tagging (MST) approach to extract causal information from historical texts. The MST method extracts causal cues in the first stage and then uses this information to extract complete causal relations in subsequent stages. More recent work presented by Liu et al. (2023) proposes an implicit cause-effect interaction framework to improve the reasoning ability of the model, which tackles event causality extraction generatively using LLMs.

Extractive Q&A. Prasad et al. (2023) explore extractive Q&A on meeting transcripts, however, not testing generative models, finding that predictions do not stick to the sentences in the transcript and could include hallucinations. Mallick et al. (2023) propose to make a generative model generate the answer index instead of generating the complete answer to reduce hallucinations. Sengupta et al. (2024) test model pret training dependencies, i.e., in the FinCausal setting, if multi-language models can learn how to answer causal Q&A in Spanish from learning how to answer them in English.

3 Method

For this research, two types of extractive Q&A Methods have been investigated. First, token classification using BERT-based models detailed in Section 3.1, and second, generative models comparing a variety of pre-trained LLMs in a few-shot setting with fine-tuning Llama 3.1 with a multi-lingual dataset.

3.1 Encoder-based model token classification for extractive Q&A

Our proposed method, illustrated in Figure 1, utilizes text embedding models, such as BERT, for token classification. Similar techniques have been presented by Yoon et al. (2022). The method begins by tokenizing both the passage of text and the question, subsequently concatenating these tokenizations with a special token, [SEP], for our implementation. During the training phase, the training dataset answer is mapped to its first occurrence in the passage of text using IO annotation style. Next, we calculate the cross-entropy loss between the passage predicted class and the actual class derived from the training data. To refine loss calculation, a loss mask restricts loss calculation to only those tokens predicted from the passage, thereby excluding mispredictions related to the question or any special tokens, such as padding tokens.

3.2 Decoder-based models for extractive Q&A

The open-ended generation nature of LLMs makes them well suited to Q&A tasks. However, for extractive Q&A, the model must follow exact instructions and not hallucinate tokens that do not exist in the context passage. First, we used prompt optimization to reduce hallucinations by iterating over a small dataset and iteratively adding rules to the prompt. The final version of the prompt can be found in the Appendix A. Next, we used few-shot learning to show each model 5, 10, or 20 Q&A examples. Last, we took the optimized prompt and fine-tuned models on 2000 examples from the English, Spanish, or both datasets combined. Since

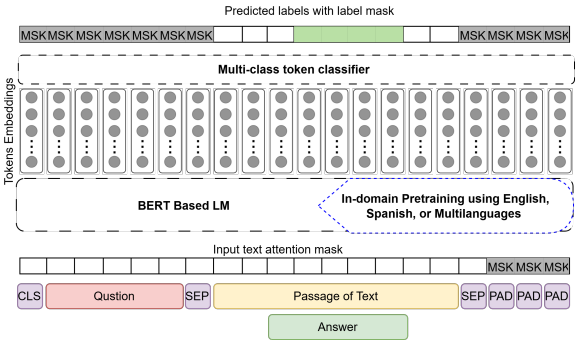

Figure 1: **BERT token classification for extractive Q&A.** The labels are inferred by mapping the answers to their first occurrence in the text. Cross-entropy loss is used to train the model. The loss is only calculated for tokens belonging to the text, excluding tokens from the question and special tokens.

large models require significant computational resources to fine-tune, we focused on training only one 70B model for both subtasks. This left us with several small monolingual models and one large multi-lingual model. We calculated the cosine similarity between the answers and used GPT-4o as a tiebreaker for the most differing answers to achieve our final results.

3.3 Model Selection

BERT was used to represent encoder-based models, while we used Llama 3.1, Mixtral, and Gemma 2 to represent generative models for prompt engineering and few-shot learning. For fine-tuning, we used Llama 3.1 8B and 70B. We also used Low-Rank Adaptation (LoRA) (Hu et al., 2021) to speed up fine-tuning. For the 8B model, a rank of 32 and an alpha of 16 were used, while for the 70B model, we used a rank of 8 and an alpha of 16 to fit memory constraints.

4 Results and Discussion

The results demonstrate a clear advantage of fine-tuned generative models over fine-tuned extractive models for the open-ended causal extraction tasks in FinCausal 2025. Extractive models such as BERT performed moderately well in identifying explicit causal links where linguistic markers (e.g., "due to," "as a result of") were present. Table 1 summarizes the different variations of BERT models utilized in this experiment. Interestingly, BERT pre-trained on multiple languages can extend the question-answering ability acquired through fine-tuning the sub-task data between the sub-task test

Base Model	Train → Test	SAS	EM
BERT EN	EN → EN	0.9242	0.6152
	EN → ES	0.7145	0.0519
BERT ES	ES → ES	0.9516	0.5808
	ES → EN	0.4064	0.0942
BERT ML	EN → EN	**0.9251**	0.6032
	EN → ES	0.9395	0.4950
	ES → ES	0.9567	**0.7086**
	ES → EN	0.8262	0.3667
	EN+ES → EN	0.9210	**0.6733**
	EN+ES → ES	**0.9656**	0.6966

Table 1: Performance of BERT models trained on different datasets. **EN**: English, **ES**: Spanish, **ML**: Multilingual. **SAS**: Semantic Answer Similarity, **EM**: Exact Matching. Training datasets exclude practice data, which is used for validation. Test datasets are blinded.

datasets (English and Spanish) more effectively than BERT pre-trained on the English language or BERT pre-trained on the Spanish language. This aligns with the findings presented by Sengupta et al. (2024)

4.1 Impact of Few-Shot Learning and Prompt Optimization

While structured prompt optimization also contributed to performance improvements, especially for Llama, where the model demonstrated increased precision under the optimized prompt structure, models still hallucinated responses with extra explanations or several alternative answers. Few-shot learning proved essential to getting concise answers from generative models to help reduce these hallucinations. Interestingly, as seen in Figure 2, the configuration with the most shots did not consistently deliver the best results for all models. Nevertheless, even after strict prompting and few-shot learning, we had to rely on fine-tuning to reach the best performance.

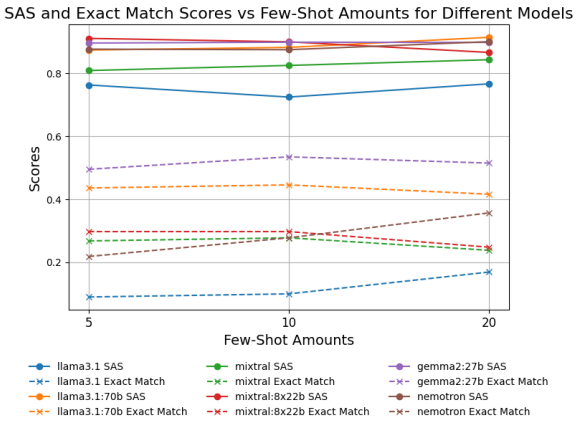

Figure 2: Few-Shot amounts for different LLMs.

Model	SAS	Exact Match
Llama 3.1 8B English	0.9649	0.8437
Llama 3.1 8B Spanish	0.9438	0.6934
Llama 3.1 8B Multilingual	0.9539	0.7415
Llama 3.1 70B Multilingual	0.9667	0.8437
GPT-4o Tiebreaker	**0.9732**	**0.8637**

Table 2: Performance of Various Models on SAS and Exact Match Metrics based on the blinded **English** evaluation set with 498 samples.

Model	SAS	Exact Match
Llama 3.1 8B English	0.9641	0.5848
Llama 3.1 8B Spanish	0.9807	0.8583
Llama 3.1 8B Multilingual	0.9775	0.8403
Llama 3.1 70B Multilingual	0.9802	0.8603
GPT-4o Tiebreaker	**0.9841**	**0.8703**

Table 3: Performance of Various Models on SAS and Exact Match Metrics based on the blinded **Spanish** evaluation set with 500 samples.

4.2 Fine-tuning

Since Llama consistently improves with more few-shot examples, we chose this model family for our fine-tuning experiments as seen in Table 2 and 3. For the first results, the smaller Llama 8B was chosen. Interestingly, the model learned to perform well even in subtasks in languages other than the training data, leading us to focus on multilingual fine-tuning. Llama 3.1 70B fine-tuned on both English and Spanish demonstrated a marked improvement, achieving a SAS score of 0.9667 and EM of 0.8437 for English, and SAS 0.9802 EM 0.8603 for Spanish. This model's generative capabilities allowed it to move beyond simple span extraction, generating responses that reflected a more comprehensive understanding of causal relationships. Both Llama 8B and 70B could interpret some implicit causal links due to their capacity for abstractive summarization. Since we had responses from several models of similar quality, we calculated the cosine similarity between the answers using GPT-4o as a tiebreaker for the most differing answers.

In summary, the generative models, particularly Llama, demonstrated clear advantages in adapting to open-ended causal tasks by generating responses that better captured the causal structure. Llama 3.1 70B emerged as the top-performing model, achieving the highest SAS and EM scores and excelling in both explicit and implicit causal detection.

4.3 Error Analysis

Both extractive models and generative models struggled at times to extract the correct answer in implicit causal relationships, where explicit causal markers (e.g., "because," "due to") were absent. They also occasionally generated responses that relied on surface-level cues within the context rather than accurately inferring the cause-and-effect relationship. Another frequent challenge was passages that nested causality. For example, in cases where multiple potential causes were mentioned, Llama 3.1 sometimes failed to identify the most relevant one, instead providing a response that included all possible causes without clear prioritization. Without annotation guidelines, it is unclear if this is due to model limitations or guideline ambiguity.

4.4 Future Directions

We encountered uncertainty in error analysis due to the absence of annotation guidelines for extracting causal answers. Extending causal information extraction guidelines, such as the ones outlined by Razouk et al. (2024a), is a promising future direction. Further, while fine-tuning Large Language Models reduced hallucinations in extractive Q&A tasks, exploring logit manipulation techniques (Niess and Kern, 2024b,a) could further enhance performance by directly changing the output probabilities of specific tokens. Lastly, the extracted causal information does not fully align with causal modeling guidelines, suggesting the need to develop evaluation methods that better integrate these standards, as discussed by Razouk et al. (2024b).

5 Conclusion

Generative methods can outperform common extractive methods in extractive Q&A tasks, provided that hallucinations are minimized. However, prompt engineering alone is not sufficient to achieve this. While few-shot learning represents an improvement, it also falls short of consistently achieving better results than extractive methods. In contrast, fine-tuning provides the necessary control to remove nearly all hallucinations in these tasks. Moreover, fine-tuned LLMs demonstrate remarkable adaptability to tasks in a language not encountered during fine-tuning, offering excellent multilingual capabilities. Using an additional model as a tiebreaker further enhances performance and suggests promising potential for a future mixture of expert solutions tailored to extractive Q&A tasks.

References

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Seethalakshmi Gopalakrishnan, Victor Zitian Chen, Wenwen Dou, Gus Hahn-Powell, Sreekar Nedunuri, 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.

Daniel Gärber. 2022. Causal Relationship Extraction from Historical Texts using BERT. Master's thesis, Graz University of Technology.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Vivek Khetan, Roshni Ramnani, Mayuresh Anand, Shubhashis Sengupta, and Andrew E Fano. 2020. Causal bert: Language models for causality detection between events expressed in text. *arXiv preprint arXiv:2012.05453*.

Jooyeon Lee, Luan Huy Pham, and Ozlem Uzuner. 2022. Mnlp at fincausal2022: Nested ner with a generative model. In *Proceedings of the 4th Financial Narrative Processing Workshop@ LREC2022*, pages 135–138.

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.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Chenyang Lyu, Tianbo Ji, Quanwei Sun, and Liting Zhou. 2022. DCU-lorcan at FinCausal 2022: Span-based causality extraction from financial documents using pre-trained language models. In *Proceedings of the 4th Financial Narrative Processing Workshop @LREC2022*, pages 116–120, Marseille, France. European Language Resources Association.

Prabir Mallick, Tapas Nayak, and Indrajit Bhattacharya. 2023. Adapting pre-trained generative models for extractive question answering. *arXiv preprint arXiv:2311.02961*.

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, Barcelona, Spain (Online). COLING.

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) - at COLING-2025*.

Georg Niess and Roman Kern. 2024a. Ensemble watermarks for large language models. *Preprint*, arXiv:2411.19563.

Georg Niess and Roman Kern. 2024b. Stylometric watermarks for large language models. *Preprint*, arXiv:2405.08400.

Archiki Prasad, Trung Bui, Seunghyun Yoon, Hanieh Deilamsalehy, Franck Dernoncourt, and Mohit Bansal. 2023. MeetingQA: Extractive question-answering on meeting transcripts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15000–15025, Toronto, Canada. Association for Computational Linguistics.

Houssam Razouk, Leonie Benischke, Daniel Garber, and Roman Kern. 2024a. Increasing the accessibility of causal domain knowledge via causal information extraction methods: A case study in the semiconductor manufacturing industry. *arXiv preprint arXiv:2411.10172*.

Houssam Razouk, Leonie Benischke, Georg Niess, and Roman Kern. 2024b. Evaluating large language models for causal modeling. *Preprint*, arXiv:2411.15888.

Anik Saha, Jian Ni, Oktie Hassanzadeh, Alex Gittens, Kavitha Srinivas, and Bulent Yener. 2022. Spock at fincausal 2022: Causal information extraction using span-based and sequence tagging models. In *Proceedings of the 4th Financial Narrative Processing Workshop@ LREC2022*, pages 108–111.

Saptarshi Sengupta, Wenpeng Yin, Preslav Nakov, Shreya Ghosh, and Suhang Wang. 2024. Exploring language model generalization in low-resource extractive qa. *arXiv preprint arXiv:2409.18446*.

Wonjin Yoon, Richard Jackson, Aron Lagerberg, and Jaewoo Kang. 2022. Sequence tagging for biomedical extractive question answering. *Bioinformatics*, 38(15):3794–3801.

A Appendix

```
**LLM Prompt for the Financial Document Causality Detection Task**

---

**Task Description:**

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

---

**Instructions:**

1. **Read the Context Carefully:**
   - Understand the events and relationships described in the context.

2. **Understand the Question:**
   - Determine whether the question is asking for a cause or an effect.
   - Identify the specific event or statement the question refers to.

3. **Extract the Answer Verbatim:**
   - Locate the exact sentence or phrase in the context that answers the question.
   - **The answer must be copied word-for-word from the context.**
   - Do not paraphrase, summarize, or add any external information.

4. **Provide Only the Answer:**
   - **Do not include any introductions, explanations, or formatting.**
   - **Output only the extracted answer, and nothing else.**

---

**Examples:**

{formatted_examples}

---

**Your Task:**

*Context:*

{text}

*Question:*

{question}

*Answer:*

[Provide only the exact answer extracted from the context.]

---

**Remember:**

- **Output only the answer. Do not include any additional text. Do not include *
    Answer:* in your answer.**
- **The answer must exactly match a portion of the context.**
- **Do not add introductions, explanations, or any extra information.**
- **any extra symbols like " or .**
- **Do not copy • from the context to the answer.**
"""
```

Figure 3: Final LLM prompt created iteratively.