Exploring the Effectiveness of Multilingual and Generative Large Language Models for Question Answering in Financial Texts

Ali Al-Laith
Copenhagen University, Denmark
alal@di.ku.dk

Abstract

This paper investigates the use of large language models (LLMs) for financial causality detection in the FinCausal 2025 shared task, focusing on generative and multilingual question answering (QA) tasks. Our study employed both generative and discriminative approaches, utilizing GPT-4o for generative QA and BERT-base-multilingual-cased, XLM-RoBerta-large, and XLM-RoBerta-base for multilingual QA across English and Spanish datasets. The datasets consist of financial disclosures where questions reflect causal relationships, paired with extractive answers derived directly from the text. Evaluation was conducted using Semantic Answer Similarity (SAS) and Exact Match (EM) metrics. While the discriminative XLM-RoBerta-large model achieved the best overall performance, ranking 5th in English (SAS: 0.9598, EM: 0.7615) and 4th in Spanish (SAS: 0.9756, EM: 0.8084) among 11 team submissions, our results also highlight the effectiveness of the generative GPT-4o approach. Notably, GPT-4o achieved promising results in few-shot settings, with SAS scores approaching those of fine-tuned discriminative models, demonstrating that the generative approach can provide competitive performance despite lacking task-specific fine-tuning. This comparison underscores the potential of generative LLMs as robust, versatile alternatives for complex QA tasks like financial causality detection.

1 Introduction

The increasing complexity of financial documents necessitates advanced methodologies to extract and analyze causality within such texts. The FinCausal 2025 shared task introduced a hybrid question-answering (QA) framework for detecting causal relationships in financial disclosures across English and Spanish languages. The task required participants to address a combination of extractive and generative QA challenges. Questions were formulated abstractly, focusing on either the cause or the effect of a relationship, while answers were required to be extracted directly from the provided financial texts.

Evaluation of the task was based on two metrics: Exact Match, which measures the strict correctness of answers, and Semantic Answer Similarity (SAS), which evaluates the semantic alignment between predicted answers and ground truths. The multilingual nature of the task, combined with the hybrid QA format, offered a unique opportunity to test the performance of state-of-the-art models in addressing causality detection across different linguistic contexts.

This paper outlines our approach to the task, which involved experimenting with multiple pre-trained large language models (LLMs), including GPT-4o, XLM-Roberta (base and large), and BERT-base-multilingual-cased. The results demonstrate the effectiveness of XLM-Roberta-large, which achieved the best performance among the tested models, securing a 5th-place rank in English and 4th-place rank in Spanish. These findings highlight the importance of leveraging multilingual large language models for nuanced tasks like financial causality detection. The code and dataset are available in GitHub: https://github.com/yemen2016/FinCausal-2025

2 Related Work

The task of causal relationship detection in financial texts has garnered significant attention in recent years, particularly with the rise of advanced Natural Language Processing (NLP) models (Ghosh and Naskar, 2022). Early approaches in this domain often relied on rule-based systems and traditional machine learning methods, such as Support Vector Machines (SVMs) and decision trees (Verma et al., 2021), to detect causal patterns in financial reports

Proceedings of the Joint Workshop of the 9th FinNLP, the 6th FNP, and the 1st LLMFinLegal, pages 230–235

and news articles. These models, however, required extensive feature engineering and often struggled to capture the complex nuances of causal relationships in the language of finance. In recent years, the advent of deep learning and transformer-based models, such as BERT and its multilingual variants, has revolutionized this field by providing models capable of understanding and extracting contextual information with little to no manual feature extraction (Yang et al., 2019).

A significant body of work in causal relationship extraction from financial text has focused on the use of pre-trained large language models like BERT and its multilingual variants (Wan and Li, 2022). Researchers have fine-tuned these models on domain-specific datasets, achieving state-of-the-art results in both causal relationship extraction and other financial text analysis tasks, such as sentiment analysis and event extraction (Mariko et al., 2020). For instance, studies have shown that XLM-R and XLM-Roberta models, which are trained on a diverse set of multilingual corpora, can generalize well to a variety of languages, including English and Spanish, making them ideal for multilingual financial text analysis tasks (Akermi et al., 2020). Fine-tuned PLMs have been demonstrated to achieve competitive performance, outperforming traditional machine learning approaches, particularly when working with large and complex datasets like financial reports (Jin et al., 2023).

Alongside fine-tuned models, there has been growing interest in leveraging generative models, such as GPT-3 and GPT-4, for causal relationship detection (Kim et al., 2023). Unlike extractive models, which pull information directly from the text, generative models produce new text based on the input provided, offering more flexibility in handling abstract and complex questions. While GPT-3 and GPT-4 have been primarily used in conversational AI, recent studies have explored their potential in tasks like question answering (QA) (Rodrigues et al., 2024; Zhang et al., 2023; Kalpakchi and Boye, 2023), euphemism detection (Firsich and Rios, 2024; Keh, 2022). Research has shown that generative models can be particularly useful in scenarios where few-shot learning is beneficial, as they can adapt to new tasks with minimal training data. However, while generative models show promise, they often require careful prompt engineering to achieve optimal results, as their performance can vary depending on the context and number of examples provided (Xiao et al., 2022; Pan et al., 2024).

3 Methodology

3.1 Dataset

Financial Causality Detection (FINCausal 2025) shared task is the dataset used in this experiment which comprises financial disclosures in English and Spanish and is structured for a hybrid question-answering task(Moreno-Sandoval et al., 2025). Each example includes four components: an identifier (ID), a context (Text), a question, and an answer. The context is a paragraph extracted from financial annual reports. Questions are designed abstractly, focusing on either the cause or effect within the text. For instance, questions might ask, "Why did X (effect) happen?" or "What is the consequence (effect) of X (cause)?" The answers are extracted verbatim from the context, adhering to an extractive approach. In cases involving complex causal relationships, such as chains or non-linear connections, up to two questions are included for clarity. This dual-language dataset challenges models to combine abstractive question generation with precise extractive answering, making it a robust resource for evaluating financial causality detection systems. We merged the training and development datasets for both English and Spanish, resulting in a combined training set of 3,999 samples. The testing set, comprises 999 samples, were kept separate. This facilitates independent performance evaluation in both English and Spanish languages during the testing phase.

3.2 Experimental Setup

The evaluation metrics in the shared task is Semantic Answer Similarity (SAS) and Exact Match, with SAS serving as the primary ranking metric. We utilized the following models in our experiments:

- Generative QA: GPT-4o
- Multilingual QA: XLM-Roberta (base and large), and BERT-base-multilingual-cased

Generative QA: GPT-4o For the Generative QA setup, GPT-4o (**model: gpt-4o-2024-08-06**) was utilized with a series of prompting techniques to evaluate its effectiveness in detecting financial causal relationships. The experiments included both zero-shot and few-shot prompting approaches. In the zero-shot setup, the model was queried without any prior examples, while the zero-shot with context experiment added relevant contextual information from the financial text. Few-shot prompting

involved providing the model with 2, 4, or 8 randomly selected examples to guide its responses. These examples served as templates, enabling the model to better understand the expected format and structure of the answers. Each configuration was evaluated in both English and Spanish to ensure the approach's robustness across languages. This experimental design aimed to examine how incremental exposure to examples impacted the model's performance, particularly in terms of its semantic answer similarity (SAS) scores.

Multilingual QA For multilingual QA, we fine-tuned XLM-Roberta (base and large) and BERT-base-multilingual-cased on both English and Spanish datasets. These models were trained to identify cause-effect relationships by aligning questions with answer spans in the text. Tokenization was performed using model-specific tokenizers to ensure compatibility, and the training objectives were adjusted to optimize for extractive answers. The multilingual QA models were trained with the following hyperparameters: Learning Rate: 2×10^{-5}, Batch Size: 16 per device, Epochs: 10, and Weight Decay: 0.01. The training process was conducted on a single GPU, and the datasets for both languages were used in all phases (training and development).

3.3 Pre-trained Language Models

In this research, we use the following four models:

1. **GPT-4o**: GPT-4o is a generative large language model designed to excel in conversational and question-answering tasks[1]. It is based on a transformer architecture with billions of parameters, fine-tuned for contextual understanding and generative capabilities. The model supports various prompting techniques, including zero-shot, few-shot, and context-aware prompting, allowing flexible adaptation to specific QA scenarios. Its capacity to process natural language queries and generate extractive answers aligns it with complex tasks such as financial question answering.

2. **XLM-Roberta-Base**: XLM-Roberta-Base, part of the XLM-R family, is a robust multilingual transformer model pre-trained on CommonCrawl data in 100 languages (Conneau et al., 2019). Unlike its predecessor XLM, XLM-R is optimized for performance by removing tasks like translation language modeling during pre-training. It employs a masked language model (MLM) objective and features 12 layers with 270 million parameters, enabling it to handle diverse linguistic structures effectively. Its balanced performance across multiple languages makes it suitable for cross-lingual and multilingual applications.

3. **XLM-Roberta-Large**: XLM-Roberta-Large is an advanced version of XLM-Roberta-Base, featuring 24 transformer layers and 550 million parameters (Conneau et al., 2019). This model achieves superior multilingual understanding by leveraging the same CommonCrawl corpus but with significantly larger capacity and depth. Its pre-training strategy, focused exclusively on the MLM objective, enhances its ability to capture complex linguistic patterns and long-range dependencies across languages. The large-scale architecture makes it particularly effective for high-resource and multilingual settings, albeit at a higher computational cost.

4. **BERT-Base-Multilingual-Cased**: BERT-Base-Multilingual-Cased is a transformer-based model pre-trained on a multilingual corpus of 104 languages, including English and Spanish (Devlin et al., 2018). The model uses a cased vocabulary, preserving capitalization, which is crucial for languages where case impacts meaning. It is trained using masked language modeling (MLM) and next-sentence prediction tasks, enabling it to understand contextual relationships in multilingual text. Its architecture consists of 12 transformer layers with 110 million parameters, making it computationally efficient for multilingual tasks.

3.4 Experimental Results

The evaluation results, as shown in Table 1, provide insights into the performance of both fine-tuned pre-trained language models (PLMs) and generative models for causal relationship detection in financial disclosures. Two key metrics were used: Semantic Answer Similarity (SAS) and Exact Match (EM). SAS measures the cosine similarity between the embeddings of predictions and references, while EM assesses the proportion of predictions that perfectly match the ground truth.

[1] https://openai.com/

	English		Spanish	
GPT-4o Prompting Technique	**SAS.**	**EM.**	**SAS.**	**EM.**
Zero-Shot	0.77	0.002	0.82	0.002
Zero-Shot w context	0.77	0.002	0.82	0.002
Few Shot - Random Examples (2)	0.92	0.387	0.92	0.341
Few Shot - Random Examples (4)	0.93	0.505	0.94	0.425
Few Shot - Random Examples (8)	0.94	0.515	0.94	0.487
Fine-tuned PLM Models	**SAS.**	**EM.**	**SAS.**	**EM.**
BERT-Base-Multilingual-Cased	0.93	0.517	0.87	0.629
XLM-Roberta-Base	0.94	0.725	0.97	0.739
XLM-Roberta-Large	**0.96**	**0.762**	**0.98**	**0.808**

Table 1: Semantic Answer Similarity (SAS) and Exact Match (EM) Results on English and Spanish Testing Sets.

GPT-4o Prompting Technique: The generative GPT-4o model demonstrated substantial variability depending on the prompting technique used. In zero-shot settings, GPT-4o performed poorly, with SAS scores of 0.77 for English and 0.82 for Spanish and minimal EM scores of 0.002 in both languages. However, the model showed considerable improvement when provided with few-shot examples. For instance, using eight examples, GPT-4o achieved SAS scores of 0.94 for both languages and EM scores of 0.515 for English and 0.487 for Spanish. This demonstrates the importance of providing targeted examples to enhance GPT-4o's performance.

Interestingly, the results indicate that GPT-4o's few-shot approach with eight examples nearly matches the SAS performance of fine-tuned models, though it still falls short in EM. This adaptability positions GPT-4o as a competitive alternative in scenarios where fine-tuning is not feasible, albeit with slightly lower precision in exact matching.

Fine-Tuned Pre-trained Language Models (PLMs): Among the fine-tuned PLMs, XLM-Roberta-Large consistently outperformed other models in both English and Spanish, achieving the highest SAS scores of 0.96 and 0.98, respectively. This model also achieved the best EM results, with 0.762 for English and 0.808 for Spanish. These results highlight the model's robustness and ability to extract accurate and nuanced causal relationships from financial texts.

The smaller XLM-Roberta-Base model also performed strongly, particularly in Spanish, with an SAS of 0.97 and an EM of 0.739. Although slightly behind its larger counterpart, this model demonstrated its efficiency for multilingual tasks. The BERT-Base-Multilingual-Cased model, while still effective, had lower performance, with SAS scores of 0.93 and 0.87 for English and Spanish, respectively, and EM scores of 0.517 and 0.629. This suggests that model size and pre-training strategies significantly influence performance in these tasks.

Comparative Insights: Fine-tuned models consistently outperformed GPT-4o in zero-shot configurations, highlighting the superiority of task-specific training for extractive question answering. However, in few-shot settings, GPT-4o demonstrated competitive performance, particularly with eight examples, narrowing the gap with fine-tuned models. This underscores GPT-4o's adaptability and effectiveness in scenarios where fine-tuning large PLMs is computationally expensive, resource-intensive, or impractical.

Language-Specific Observations Across all models, Spanish texts exhibited higher SAS and EM scores compared to English, with XLM-Roberta-Large achieving particularly strong results. These findings suggest that Spanish financial texts may possess structural or lexical characteristics that are more conducive to causal relationship detection or that the training data provided better representation for Spanish. This disparity underscores the importance of tailoring model development and evaluation to specific languages.

4 Discussion of Results

The experimental results highlight the comparative performance of fine-tuned pre-trained language models (PLMs) and GPT-4o prompting techniques for detecting causal relationships in financial texts

across English and Spanish. For both Semantic Answer Similarity (SAS) and Exact Match (EM), fine-tuned models demonstrated superior performance, with XLM-Roberta-Large emerging as the best-performing model. It achieved the highest SAS scores (0.96 for English and 0.98 for Spanish) and EM scores (0.762 for English and 0.808 for Spanish), showcasing its capability to handle complex extractive question-answering tasks. These results underscore the strength of leveraging large-scale multilingual PLMs for tasks requiring precision and contextual understanding.

Among the fine-tuned models, XLM-Roberta-Base also performed strongly, particularly in Spanish, where it achieved a high SAS of 0.97 and an EM of 0.739. BERT-Base-Multilingual-Cased, while slightly behind, still delivered competitive results, particularly in English, with an EM of 0.517. This demonstrates that even smaller, less computationally intensive models can perform effectively, particularly when fine-tuned on specific tasks.

In contrast, GPT-4o, while initially less effective in zero-shot configurations (SAS: 0.77 and EM: 0.002 for both English and Spanish), showed significant improvement under few-shot settings. By incorporating up to eight random examples during prompting, GPT-4o achieved SAS scores of 0.94 for both languages, with corresponding EM scores of 0.515 for English and 0.487 for Spanish. These results illustrate GPT-4o's adaptability and potential in resource-constrained environments where extensive fine-tuning of large PLMs is not feasible. However, the relatively lower EM scores in comparison to fine-tuned PLMs suggest that GPT-4o, while versatile, may not yet match the precision offered by task-specific models in exact-match scenarios.

The disparity in performance between English and Spanish, particularly for fine-tuned models, further underscores the influence of language-specific characteristics on model effectiveness. Spanish financial texts consistently yielded higher SAS and EM scores, suggesting better alignment between the models and linguistic nuances of Spanish financial disclosures. This finding highlights the need for tailored approaches and datasets to ensure optimal performance in multilingual environments.

In summary, the results demonstrate the complementary strengths of fine-tuned PLMs and generative models. Fine-tuned models excel in accuracy and task-specificity, while GPT-4o offers a flexible alternative, particularly when fine-tuning is infeasible. Future research could explore hybrid methodologies that combine the robustness of fine-tuned models with the adaptability of generative techniques, potentially enhancing performance across diverse tasks and languages.

5 Conclusion

This study investigated the effectiveness of fine-tuned pre-trained language models (PLMs) and generative prompting techniques for causal relationship detection in financial disclosures in English and Spanish. The results underscore the complementary strengths of both approaches in addressing this challenging task.

The GPT-4o generative model showcased impressive adaptability, particularly in few-shot configurations, where its SAS scores approached those of fine-tuned PLMs. Despite lower EM scores, GPT-4o's ability to perform competitively without extensive fine-tuning makes it a valuable alternative in scenarios with limited resources or time constraints. These results reinforce the versatility of generative language models, particularly when used with targeted prompting techniques.

On the other hand, fine-tuned PLMs, particularly XLM-Roberta-Large, demonstrated superior performance, achieving the highest scores in both Semantic Answer Similarity (SAS) and Exact Match (EM) metrics. These results highlight the advantages of leveraging large-scale multilingual PLMs for tasks requiring high precision and contextual understanding. The performance of smaller models, such as XLM-Roberta-Base and BERT-Base-Multilingual-Cased, also underscores the potential of fine-tuned PLMs to deliver strong results even with reduced computational demands.

Notably, the consistently higher performance on Spanish financial texts highlights the impact of language-specific nuances in financial disclosures and emphasizes the importance of tailored datasets and approaches in multilingual contexts.

Overall, this work demonstrates the value of using fine-tuned PLMs and generative approaches for extractive question answering tasks. Future research could focus on hybrid methodologies, integrating the precision of fine-tuned models with the adaptability of generative models, to further enhance causal relationship detection in financial texts across diverse languages and domains.

References

Imen Akermi, Johannes Heinecke, and Frédéric Herledan. 2020. Transformer based natural language generation for question-answering. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 349–359.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Todd Firsich and Anthony Rios. 2024. Can gpt4 detect euphemisms across multiple languages? In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 65–72.

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.

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.

Sedrick Scott Keh. 2022. Exploring euphemism detection in few-shot and zero-shot settings. *arXiv preprint arXiv:2210.12926*.

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.

Dominique Mariko, Hanna Abi Akl, Estelle Labidurie, Stephane Durfort, Hugues De Mazancourt, and Mahmoud El-Haj. 2020. Financial document causality detection shared task (fincausal 2020). *arXiv preprint arXiv:2012.02505*.

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*.

Siduo Pan, Ziqi Zhang, Kun Wei, Xu Yang, and Cheng Deng. 2024. Few-shot generative model adaptation via style-guided prompt. *IEEE Transactions on Multimedia*.

Luiz Rodrigues, Filipe Dwan Pereira, Luciano Cabral, Dragan Gašević, Geber Ramalho, and Rafael Ferreira Mello. 2024. Assessing the quality of automatic-generated short answers using gpt-4. *Computers and Education: Artificial Intelligence*, 7:100248.

Devika Verma, Ramprasad Joshi, Shubhamkar Joshi, and Onkar Susladkar. 2021. Study of similarity measures as features in classification for answer sentence selection task in hindi question answering: Language-specific v/s other measures. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 747–756.

Chang-Xuan Wan and Bo Li. 2022. Financial causal sentence recognition based on bert-cnn text classification. *The Journal of Supercomputing*, pages 1–25.

Jiayu Xiao, Liang Li, Chaofei Wang, Zheng-Jun Zha, and Qingming Huang. 2022. Few shot generative model adaption via relaxed spatial structural alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11204–11213.

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. *arXiv preprint arXiv:1902.01718*.

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. *arXiv preprint arXiv:2310.13265*.