Enhancing Causal Relationship Detection Using Prompt Engineering and Large Language Models

Pulkit Chatwal[1] **Amit Agarwal**[2] **Ankush Mittal**[3]
[1]Rajiv Gandhi Institute of Petroleum Technology, Jais, India
[2]AICoE, Wells Fargo International Solutions Private Limited, Bangalore, India
[3]COER University, Roorkee, India
pulkitchatwal@gmail.com, aagarwal3@iitr.ac.in, dr.ankush.mittal@gmail.com

Abstract

Causal relationships are essential for understanding financial systems, offering insights into market dynamics, regulatory impacts, and organizational decisions. Traditional approaches to detecting causality in financial texts often struggle with the nuanced and domain-specific language of such disclosures. The FinCausal 2025 shared task provided a benchmark for evaluating advanced methods in multilingual financial causality detection.

In this work, we employ prompt engineering with large language models (LLMs) to identify causal relationships in financial disclosures across languages. Our system achieved Semantic Answer Similarity (SAS) scores of 0.9086 in English and 0.8987 in Spanish, with Exact Match (EM) scores of 0.5110 and 0.0619, respectively. These results demonstrate the potential of LLMs for tackling the challenges of causality detection in multilingual and domain-specific contexts, while also identifying areas for future refinement.

1 Introduction

Causal relationships are central to understanding complex systems, particularly in domains like finance, healthcare, and policy. In the financial sector, these relationships help interpret market dynamics, regulatory changes, and organizational decisions. Financial disclosures often reveal such causal links, but detecting them in unstructured text is challenging. Traditional NLP methods, relying on linguistic patterns or machine learning, struggle with the nuanced language of financial texts. The rise of multilingual NLP underscores the need for models that handle diverse languages with limited annotated resources. Recently, large language models (LLMs) and prompt engineering have revolutionized NLP, enabling domain-specific tasks like causality detection in under-resourced languages. Unlike extractive methods, prompt engineering allows LLMs to reason about complex cause-and-effect relationships. In finance, robust causality detection is crucial for explaining market events and organizational outcomes in increasingly globalized and complex narratives. This paper explores prompt engineering with LLMs to detect causal relationships in financial disclosures, contributing to this critical field.

2 Related Work

Detecting causal relationships in text has been a longstanding challenge in natural language processing (NLP). Several studies have made significant strides in this domain, exploring various methods, languages, and applications. (Blanco et al., 2008) developed a supervised method for extracting explicit causal relations using syntactic patterns and machine learning. Their approach, while foundational, focused on a limited set of patterns such as verb phrases and causal relators like because and since, making it less adaptable to complex real-world contexts. (Yang et al., 2022) expanded this by surveying causality extraction methods, including knowledge-based and deep learning techniques, emphasizing the potential of deep learning to handle implicit and inter-sentential causality, despite challenges like data scarcity and computational demands. Transformer-based models have proven highly effective in diverse NLP tasks, such as age detection across social media platforms (Sankar et al., 2024) and automating farmer query resolution with AgriLLM (Didwania et al., 2024). (Reimann, 2021) extended causality detection into multilingual settings, demonstrating how zero-shot and few-shot transfer learning using transformer-based models could address data limitations for languages like German and Swedish. However, performance still depended heavily on the availability and quality of training data. (Feder et al., 2022)explored the integration of causal inference in NLP,

Proceedings of the Joint Workshop of the 9th FinNLP, the 6th FNP, and the 1st LLMFinLegal, pages 248–252
January 19–20, 2025. ©2025 Association for Computational Linguistics

addressing spurious correlations and proposing causal debiasing techniques to improve model robustness and interpretability. Leadership traits during natural hazards have been studied using social media data (Agarwal and Toshniwal, 2020), and SMS-based FAQ systems have tackled noisy text challenges (Agarwal et al., 2015), showcasing NLP's versatility across domains. Prompt engineering, as highlighted by (Marvin et al., 2023), has become a transformative method for adapting large language models to domain-specific tasks. They emphasized its versatility in complex NLP tasks, providing a strong foundation for exploring causal detection using LLMs. (Xiao et al., 2024) reviewed the application of NLP in financial analytics, noting its effectiveness in automating tasks such as financial report analysis and risk assessment, underscoring the relevance of NLP in finance. In this work, we build on these advancements by introducing a hybrid QA approach for detecting causal effects in financial disclosures. Leveraging multilingual datasets and prompt engineering techniques such as zero-shot, few-shot, and chain-of-thought prompting, we adapt the Llama 3.2 model to generate both extractive and generative responses, advancing causal detection in financial NLP beyond existing methods.

3 Dataset

The Financial Causality Detection (FinCausal 2025) task was introduced to advance research in identifying causal relationships within financial narratives. The task requires models to determine the cause or effect from financial reports in English and Spanish. Each data point consists of a *context* (a paragraph from financial reports), a *question* (targeting the cause), and an *answer* (verbatim text extracted from the context). The evaluation is generative, using exact matching and similarity metrics.

3.1 Dataset Overview

The English dataset is sourced from 2017 UK financial annual reports (*UCREL corpus*), while the Spanish dataset comprises financial reports from 2014 to 2018. Both datasets are structured to ensure comparability for testing multilingual models. A summary of the dataset splits is provided in Table 1.

Language	Training	Reference	Input
English	1999	100	498
Spanish	2000	100	500

Table 1: Summary statistics of the datasets.

4 Methodology

4.1 LLaMA 3.2 Model

LLaMA 3.2 was chosen for its advanced multilingual processing capabilities and exceptional performance in both extractive and generative QA tasks. Its ability to handle nuanced financial disclosures in English and Spanish made it well-suited for causality detection. The model's architecture also supports effective integration of advanced prompt engineering techniques, enabling precise causal inference (Dubey et al., 2024).

4.2 Prompt Engineering Techniques

Prompt engineering played a pivotal role in optimizing the model's performance. This study employed different strategies tailored to the complexities of English and Spanish datasets. The prompt engineering methods used were informed by best practices outlined in (Sahoo et al., 2024).

4.2.1 English: Four Techniques

Zero-Shot Prompting: The model was provided with the text and question directly, with no prior examples. This method served as a baseline, handling straightforward causalities well, but struggling with more complex or multi-step relationships, as it lacked guidance for reasoning through intricate scenarios.

Prompt: *"Answer each question precisely, using only the information provided in the text."*

Few-Shot Prompting: We provided the model with 10 carefully selected examples of causal relationships to help it understand and generalize causal patterns. By showing a few examples, the model improved its accuracy, particularly for extractive question-answering tasks that align with the examples.

Prompt: *"Answer each question precisely, using only the information provided in the text. Below are 10 examples demonstrating this process."*

Chain of Thought (CoT): Intermediate reasoning steps were introduced to help the model understand and process multi-step causalities. This approach improved the model's ability to break down complex cause-effect relationships into manageable components, particularly for indirect or multiple causal factors.

Prompt: *Prompt: "You are a highly skilled assistant with expertise in financial and corporate contexts. Your role is to identify the specific answer to a question by carefully analyzing the given text. Follow these steps to ensure accuracy and alignment:*

- Identify relevant keywords in the text that answer the question directly.
- Extract only the necessary information, focusing on precision and alignment with the question.
- Rephrase the answer if needed to ensure conciseness and relevance without losing meaning.

Few-Shot + Chain of Thought: This approach used the same prompt as the CoT technique, but incorporated 10 examples from the Few-Shot technique. These examples guided the model in understanding causal patterns and improved performance when reasoning through complex multi-step causalities. This combined method proved to be the most effective for both direct and complex causal relationships.

4.2.2 Spanish: Few-Shot

The Few-Shot technique was used to provide the model with 10 carefully selected examples of causal relationships. These examples helped the model understand how causal connections are expressed in Spanish financial documents, allowing it to generalize and answer new questions based on the same pattern.

Prompt: *"You are a Spanish financial analyst assistant experienced in analyzing corporate, legal, and financial documents. Your task is to answer questions in Spanish concisely and factually, based only on the information provided in the text. Do not interpret, infer, or add information not explicitly stated. Provide direct answers without extra details. Below are 10 examples demonstrating this process."*

4.3 Example Outputs

4.3.1 English: Example Outputs

Example Context: "Underlying Group EBITDA declined by 10.1% to £10.0m (2016: £11.2m). This decline has been driven by an increase in UK overheads of £1.0m (5.6%), due to investment in support of our strategic initiatives and well-publicised cost headwinds."

Question: "What has motivated the increase in UK overheads by £1.0 million, or 5.6%?"

Generated Answers:

- Zero-Shot: "investment in support of our strategic initiatives and well-publicised cost headwinds"
- Chain of Thought: "investment in support of our strategic initiatives and well-publicised cost headwinds"
- Few-Shot: "investment in support of our strategic initiatives and well-publicised cost headwinds"
- Few-Shot + Chain of Thought: "investment in support of our strategic initiatives and well-publicised cost headwinds"

4.3.2 Spanish: Example Outputs

Example Context: "Por este motivo, diferentes áreas han participado en un programa formativo diseñado para mejorar la gestión de las reclamaciones y para familiarizarse con la nueva herramienta que apoya la gestión de las mismas."

Question: "¿Qué ha ocurrido por este motivo?"

Generated Answer: "diferentes áreas han participado en un programa formativo diseñado para mejorar la gestión de las reclamaciones y para familiarizarse con la nueva"

4.4 Comparative Performance of Prompt Engineering Techniques

Table 2: Comparative Performance of Prompt Engineering Techniques

Language	Method	SAS	EM
English	Zero-Shot	0.758	0.292
English	Few-Shot	0.877	0.470
English	CoT	0.844	0.545
English	Few-Shot + CoT	**0.908**	**0.511**
Spanish	Few-Shot	**0.898**	0.0619

5 Results

5.1 Evaluation Metrics

To evaluate model performance, the shared task organizers used two primary metrics: Exact Match (EM) and Semantic Alignment Score (SAS).

Language	Metric	Score
English	SAS	0.9086
English	EM	0.5110
Spanish	SAS	0.8987
Spanish	EM	0.0619

Table 3: Test scores of our systems, provided by the organizers.

Exact Match (EM): This metric measures the percentage of cases where the model's predicted answer matches the ground truth exactly. **Formula:**

$$EM = \frac{\text{Number of exact matches}}{\text{Total number of examples}} \times 100$$

Semantic Alignment Score (SAS): SAS assesses the semantic similarity between the predicted and ground truth answers. This metric uses cosine similarity between the embeddings of the two answers, allowing for partial credit when the generated answer is semantically correct but not an exact match. **Formula:**

$$SAS = \text{Cosine Sim}(Emb_{\text{predicted}}, Emb_{\text{ground truth}})$$

These metrics ensure a comprehensive evaluation by accounting for both exact correctness and semantic closeness.

5.2 Results and Analysis

The best scores achieved by our approach for the English and Spanish datasets, using the Few-Shot + Chain of Thought and Few-Shot methods respectively, are summarized in Table 3.

English Dataset: For the English dataset, the highest scores achieved are an Exact Match (EM) of 0.5110 and a Semantic Alignment Score (SAS) of 0.9086. These results demonstrate the effectiveness of the Few-Shot + Chain of Thought approach, which successfully captures the semantic meaning of causal relationships while providing reasonable precision in exact matches. The high SAS score indicates the model's strong ability to semantically align with the context of causal effects, making this method particularly well-suited for understanding complex causal relationships in English.

Spanish Dataset: For the Spanish dataset, the best scores achieved are an EM of 0.0619 and a SAS of 0.8987. Although the EM score is lower, the SAS score reflects the model's capacity to semantically align with the causal information. The Few-Shot method proved to be effective in this context, leveraging example-based learning to identify causal relationships in Spanish. While exact matches were harder to achieve, the method excelled in capturing the overall semantic alignment, which is crucial for multilingual tasks.

These results highlight the strengths of the Few-Shot + Chain of Thought approach for English, where complex causalities benefit from reasoning steps, and the Few-Shot approach for Spanish, which excels in example-driven learning to align with semantic information.

6 Conclusion

In this study, we explored the effectiveness of Few-Shot and Few-Shot + Chain of Thought prompting techniques for identifying causal relationships in financial and corporate texts. The results highlight that Few-Shot + Chain of Thought achieved superior performance for English, excelling in capturing complex causal relationships through structured reasoning steps. For Spanish, the Few-Shot approach demonstrated strong semantic alignment, effectively leveraging example-based learning to adapt to linguistic nuances. These findings underline the importance of tailoring prompt engineering techniques to the specific characteristics of each language.

To further enhance the performance of causal relationship detection, we plan to fine-tune LLaMA 3.2 and evaluate additional state-of-the-art large language models (LLMs), such as Phi3 (Abdin et al., 2024) Mistral (Jiang et al., 2023)and Qwen(Bai et al., 2023). These models will be tested for their capabilities in handling multilingual causal inference tasks, comparing their strengths in understanding domain-specific language and nuanced causal reasoning. Future efforts will also focus on expanding the dataset to include more languages and diverse financial scenarios, enabling broader applicability and improved generalization of the models.

By combining fine-tuning techniques with the exploration of diverse LLM architectures, we aim to advance semantic understanding and causal reasoning in multilingual financial text analysis, paving the way for more robust applications in financial decision-making and narrative generation.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Amit Agarwal, Bhumika Gupta, Gaurav Bhatt, and Ankush Mittal. 2015. Construction of a semi-automated model for faq retrieval via short message service. In *Proceedings of the 7th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 35–38.

Amit Agarwal and Durga Toshniwal. 2020. Identifying leadership characteristics from social media data during natural hazards using personality traits. *Scientific reports*, 10(1):2624.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Eduardo Blanco, Nuria Castell, and Dan I Moldovan. 2008. Causal relation extraction. In *Lrec*, volume 66, page 74.

Krish Didwania, Pratinav Seth, Aditya Kasliwal, and Amit Agarwal. 2024. Agrillm: Harnessing transformers for farmer queries. *arXiv preprint arXiv:2407.04721*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Ggaliwango Marvin, Nakayiza Hellen, Daudi Jjingo, and Joyce Nakatumba-Nabende. 2023. Prompt engineering in large language models. In *International conference on data intelligence and cognitive informatics*, pages 387–402. Springer.

Sebastian Michael Reimann. 2021. Multilingual zero-shot and few-shot causality detection.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

Thadavarthi Sankar, Dudekula Suraj, Mallamgari Reddy, Durga Toshniwal, and Amit Agarwal. 2024. Iitroorkee@ smm4h 2024 cross-platform age detection in twitter and reddit using transformer-based model. In *Proceedings of The 9th Social Media Mining for Health Research and Applications (SMM4H 2024) Workshop and Shared Tasks*, pages 101–105.

Jue Xiao, Jiangshan Wang, Wenqing Bao, Tingting Deng, Shuochen Bi, et al. 2024. Application progress of natural language processing technology in financial research. *Financial Engineering and Risk Management*, 7(3):155–161.

Jie Yang, Soyeon Caren Han, and Josiah Poon. 2022. A survey on extraction of causal relations from natural language text. *Knowledge and Information Systems*, 64(5):1161–1186.