Financial Causal QA via Instruction and Prompt Tuning of Gemma3-12B

Avinash Trivedi, Chindukuri Mallikarjuna
SRM University-AP,
Amaravati, Andhra Pradesh 522240, India
avinashtrivedi.2008@gmail.com

Abstract

In this paper we present a novel methodology that harnesses the power of prompt tuning applied directly to Gemma3-12B, a state-of-the-art generative large language model to enhance performance on complex natural language processing challenges. Instead of relying solely on extensive retraining, our approach leverages carefully crafted input prompts to steer the pre-trained Gemma-12B towards generating outputs with superior contextual accuracy and interpretability. Our experimental evaluation employed a composite LLM Score metric that quantifies both semantic coherence and relevance; under this framework, our system (Team Name: *Sarang*) achieved a score of 4.54, ranking 9th in the shared task. Furthermore, in the competitive task evaluation, our method demonstrated the potential of prompt tuning as a viable alternative to traditional fine-tuning approaches. This study not only demonstrates the practical benefits of integrating prompt engineering with large language models but also opens avenues for future research aimed at further optimizing model performance in domain-specific applications.

Keywords: FinCausal, Prompt Tuning, LLM Score

1. Introduction

Natural question answering systems which are expected to facilitate decision-making and market research should have a sophisticated sense of cause-and-effect structure embedded in financial narrative. Annual reports, earnings statements, etc., often describe events, antecedents and the consequences of those events such that it requires a keen sense of cause and effect. These linkages are computationally intensive and impractical to extract manually when faced with the large volume of modern financial text corpora. Therefore, automated causal question-answering reduces such limitations, simplifying the processing of financial information and making the produced financial intelligence more readable and understandable. It is on this basis that the creation of systems that can respond to causal questions derived through financial narratives has become one of the critical research areas.

The FinCausal 2026 Shared Task (Moreno-Sandoval et al., 2026a), which is structured into the Financial Narrative Processing Workshop, is focused on the improvement of the methodologies of causal question-answering in financial texts. The dataset has been carefully designed in terms of triadic settings of context, inquiry, and response whereby the participants are required to make an inference of the missing causal determinant on a financial passage and the query related to that passage. These interrogatives are highly abstracted and hence compel the examinees to identify antecedent or consequent items whereas the intended responses are specific extractive spans obtained out of the contextual contents. Therefore, the construct can be described as a graceful combination of classical span extraction and question answering with reasoning, and, as a consequence, such an elevated standard of subtle understanding of financial stories. Although modern large language models have made significant progress, it is still an extremely challenging task to identify causal relationships in the financial discourse: causes and effects are incorporated in a systematic way in such texts, specialised terminological registers are used, and long-term causal relationships are cultivated. The 2026 update makes the complexity intrinsic, with increased granularity of causal processes, and re-organization of ways of inquiry, to require a high level of inferential skill. This means that researchers have to come up with mechanisms that can be used to traverse complex causal networks, beyond surface pattern recognition. What adds to these inherent difficulties is the addition of a new judgment measure based on the use of Large Language Models as adjudicators (LLM-as-a-judge). This change of paradigm shifts the focus of emphasis on specific span recall to a general evaluation of the semantic fidelity and reasoning profundity. Correspondingly, analysis will be based on the ability of a system to encode cause-and-effect relationships with probable coherence, and not merely match annotated spans.

2. Literature study

The methodic identification and parsing out of causal relationships in the financial discourse has

Proceedings of the 7th Financial Narrative Processing Workshop @ LREC 2026, pages 169–174
May 16, 2026. ©2026 European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

become a critical project in the context of financial natural language processing. The FinCausal shared tasks have provided a significant impact on this pattern, providing strict guidelines and increasing complex evaluation models. In 2020, the first FinCausal attempt was launched, introducing the first annotated corpus on financial causality detection, which defines two fundamental subtasks: sentence-level classification and definition of cause-effect spans(Mariko et al., 2020) In 2021 and 2022, further cycles of improvement were made, along with annotation guidelines, increasing the data coverage, and advancing more complex causal patterns, including quantified facts and transformation-based relations (Mariko et al., 2021, 2022). These initial implementations demonstrated the effectiveness of transformer-based encoders in combination with highly structured span-extraction systems, and at the same time, the difficulty in representing implicit causality and cross-sentence reasoning. Based on this background, the 2023 version of FinCausal expanded its criterion to include multilingual tracks and redefined the role of the span-oriented extraction as the part of question answering or sequence generating (Moreno-Sandoval et al., 2023). The 2025 version also better specified its aims by incorporating a multilingual causal question-answer model that may be assessed by Exact Match (EM) and Semantic Answer Similarity (SAS) scores (Moreno-Sandoval et al., 2025). This redefinition marked the end of pure extraction and an incorporative reasoning and generation of answers. FinCausal is closely related to the development of financial question answering research, which places greater emphasis on financial reasoning and multi-step inference, at the cost of more traditional financial reporting (Chen et al., 2021). Taken together, these changes reflect a larger change in causal recognition, that is, surface-level identification to a more abstract financial reasonability that incorporates text with quantitative information. The effectiveness of hybrid architectures which combine extractive precision and generative reasoning is further supported by the recent literature. (Pilault et al., 2020) proposed a model in which an extractive item picks the relevant evidence to modulate a transformer-based generative model and thus showing a better contextual consistency. According to (Luo et al., 2022), extractive and generative QA models have been systematically compared, with the first type proving much more successful in generalisation in limited settings, while the latter type possesses the benefits of abstraction. The NeurIPS EfficientQA competition (Min et al., 2021) unveiled the trade-off between the computation efficiency and results, where well optimised lightweight extractive models can be competitive with state-of-the-art results. Basic transformer models like RoBERTa (Liu et al., 2019) have also enhanced the abilities of domain adaptation.

3. Dataset for FinCausal2026

This paper uses English version of FinCausal 2026 Question Answering dataset (Moreno-Sandoval et al., 2026b) published as part of FinCausal shared task, which is dedicated to detecting causal links in financial narratives. The task is expected to test systems, which are able to read financial texts and identify the cause or effect of financial events. The dataset is assembled out of the financial reports and corporate disclosures in which causal relations are often found in the descriptions of the company performance, market trends and economic situation. The dataset consists of 2000 training instances and 500 instances for testing. In the training dataset each instance includes *ID*, *Context*, *Question* and *Answer*. Whereas test data contains same attributes except *Answer*.

Considering a financial context and a causal question, the goal is to derive the right cause or effect.

4. Methodology

4.1. Few-shot prompting and Finetuning of Language Model

As a baseline, We tried few-shot prompting of various LLMs. Later started the finetuning of *consciousAI/question-answering-roberta-base-s*, then changed the checkpoint to *deepset/deberta-v3-large-squad2* followed by prompt based enhancement steps as in Fig 1, inspired from (Trivedi et al., 2025) to improve the response received from finetuned model. This technique was giving LLM score of 4.424. We also tried prompt tuning and from there we found our best performing system discussed in section 4.2.

4.2. System Submission

The current section outlines the proposed structure of the FinCausal Question Answering (QA) task. Fig 2 representing the architecture of the model submission. The current research involves a methodology that integrates the few-shot learning, automated prompt optimization and a large language model to identify causal relationships in finance in a systematic manner. The pipeline includes four main elements, dataset utilisation, few-shot prompt construction, prompt tuning, and model inference, which together allow performing sound causal reasoning on financial texts.

Prompt

```
{"role": "system",
"content": "You are a helpful assistant
that provides accurate and improved an-
swers."},
{"role": "user",
"content": """You are given a Context, a
Question, and an Answer.
1. If the Answer is 100% correct and is ex-
tracted verbatim from the Context, return
the exact same Answer.
2. If the Answer is incorrect or not fully
extracted from the Context, return an im-
proved version of the Answer that is ex-
tracted verbatim from the Context.
Context: {context}
Question: {question}
Answer: {answer} """}
```

Figure 1: Prompt for enhancement step

4.2.1. Few-Shot Prompt Construction

Few-shot learning has proven to be effective in instructing large language models to perform specialised reasoning (Brown et al., 2020; Wei et al., 2022). In the current methodology, few instances of the dataset are integrated into prompts as demonstrations. Each instance in the dataset is characterized by a financial context, a causal question, and the answer that clearly outlines the cause effect relationship. Such demonstrations offer implicit information to the language model, and it is able to identify how causal relationships are formulated in financial texts and how the appropriate responses can be produced.

4.2.2. Prompt Optimization using MIPROv2

To further enhance timely efficacy, the system will use MIPROv2 prompt optimization through the DSPy teleprompter framework (Khattab et al., 2022, 2024). DSPy provides a programmatic interface that is structured in such a way that it can be optimized and interactions with language models co-ordinated. The MIPROv2 teleprompter automatically optimizes prompts by sequentially sampling a space of instruction and few-shot examples. Through iterative evaluation, the system will pick highly effective prompts, thus improving the ability of the model to identify causal relationship and provide accurate answers for FinCausal QA task.

4.2.3. Model Integration and Inference

Gemma3:12B large language model(Gemma and DeepMind, 2024) is the refined version that performs the prompts and provides the strong language understanding and reasoning skills. The implementation of the model makes use of DSPy, which is used together with Ollama to enable efficient local execution and enable a controlled interaction with the model. In the process of inference, the system obtains a context and a causal query. The optimized prompt along with the selected few-shot examples are sent to the language model using the DSPy framework. The model then performs contextual reasoning on the financial text and then gives out the final answer thus discovering the relevant causal relationship.

Our final model was build on prompt tuning of Gemma3-12B using MIPROv2 of DSPy (Khattab et al., 2022, 2024). The tuned system prompt is in Fig 3 and best parameters are listed in Table 1

Hyperparameter	Value
auto	medium
max_bootstrapped_demos	10
max_labeled_demos	10
model	gemma3:12b

Table 1: MIPROv2 Hyperparameters

5. Experimental Results

The results of our experiments few-shot prompting, prompt tuning and including finetuning of *consciousAI/question-answering-roberta-base-s* and *deepset/deberta-v3-large-squad2* are listed in Table 2.

Technique	Model	LLM Score
Finetuning	roberta based	4.136
Finetuning + Enhancement	roberta based	4.288
Finetuning	deberta based	4.292
Finetuning + Enhancement	deberta based	4.302
Few-shot	gemma2:latest	4.424
Few-shot	qwen3:latest	4.418
Few-shot + light optimization	gemma3:12b	4.448
Few-shot + medium optimization	**gemma3:12b**	**4.54**

Table 2: Performance comparison on test set

The experimental findings prove the existence of a performance improvement between the conventional transformer-based fine-tuning strategies and LLM based few-shot strategies augmented with prompt tuning. First, both the RoBERTa-based and DeBERTa-based models perform competitively among the fine-tuning methods. The initial fine-tuned RoBERTa model has a score of 4.136 that is enhanced to 4.288 using prompt based response enhancement step mentioned in Fig 1. In the same manner, the fine-tuned DeBERTa model achieves 4.292, which is slightly higher than RoBERTa and it reaches 4.302 after applying enhancement step. These outcomes demonstrate the idea that the architectural variation (DeBERTa versus RoBERTa) comes with minor benefits, whereas the enhancement strategies are both able to provide incremental improvements to both models.

Conversely, the few-shot LLM-based models have a visible lead over all fine-tuned encoder mod-

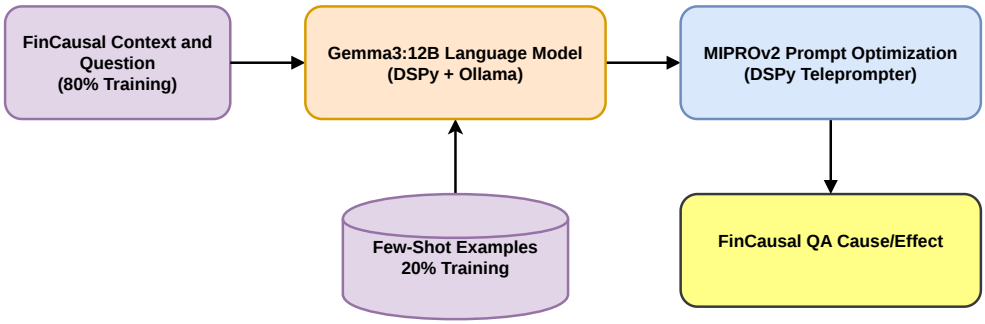

Figure 2: Model architecture

Prompt

```
Your input fields are:
1. context (str): Financial report excerpt
containing the causal relation
2. question (str): Question asking for the
cause or effect

Your output fields are:
1. answer (str): Exact causal span copied
from the context

All interactions will be structured in the
following way, with the appropriate values
filled in.

[[ ## context ## ]]
{context}

[[ ## question ## ]]
{question}

[[ ## answer ## ]]
{answer}

[[ ## completed ## ]]

In adhering to this structure, your objec-
tive is:

You are a financial causal reasoning expert.

The answer to the question (cause or effect)
is ALWAYS explicitly stated in the provided
context.

Extract the exact text span from the context
that answers the question.

Rules:
- The answer MUST be copied verbatim from the
context.
- Do NOT paraphrase.
- Do NOT add any extra words.
- Return only the precise causal phrase.
```

Figure 3: Tuned system prompt

els. The few-shot set up using Gemma2 also gives 4.424, and Qwen3 gives 4.418, indicating the high level of generalisation of large instruction-tuned models without task-specific fine-tuning. This indicates that prompt-based learning on modern LLMs performs better at this task environment than in traditional fine-tuning. Lastly, better performance is further improved by optimisation of bigger LLMs. The few-shot plus light optimisation experiment with Gemma3 (12B) has a result of 4.448 and medium optimisation has the most optimal result of 4.54, which is the best overall result. The development of this process underlines the fact that prompt tuning methods significantly enhance the efficiency of a model. Overall, the outcomes show that there is a definite trend: bigger LLMs with organised prompt tuning outperform classic fine-tuned transformer baselines, achieving the best results in an experimental environment.

6. Conclusions and Future Work

Within this investigation we have delineated several experimental paradigms, including few-shot learning, fine-tuning procedures, and prompt tuning. Our most effective submission i.e. Gemma3-12B prompt tuning achieved an LLM Score of 4.54.

Looking ahead, future work will concentrate on overcoming computational resource constraints, thereby permitting an in-depth exploration of prompt tuning strategies for larger LLMs. Furthermore, investigating data augmentation techniques to fine-tune the deberta based checkpoint represents another promising research direction. Finally, the potential integration of LLM agents remains a viable avenue for subsequent experimental inquiries.

7. Bibliographical References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. Finqa: A dataset of numerical reasoning over

financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3697–3711. Association for Computational Linguistics.

Team Gemma and Google DeepMind. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive NLP. *arXiv preprint arXiv:2212.14024*.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. Dspy: Compiling declarative language model calls into self-improving pipelines.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Man Luo, Kazuma Hashimoto, Semih Yavuz, Zhiwei Liu, Chitta Baral, and Yingbo Zhou. 2022. Choose your qa model wisely: A systematic study of generative and extractive readers for question answering. In *Proceedings of the 1st Workshop on Semiparametric Methods in NLP: Decoupling Logic from Knowledge*, pages 7–22, Dublin, Ireland and Online. Association for Computational Linguistics.

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, Marseille, France. European Language Resources Association.

Dominique Mariko, Hanna Abi Akl, Estelle Labidurie, Stephane Durfort, Hugues de Mazancourt, and Mahmoud El-Haj. 2021. The financial document causality detection shared task (FinCausal 2021). In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 58–60, Lancaster, United Kingdom. Association for Computational Linguistics.

Sewon Min et al. 2021. Neurips 2020 efficientqa competition: Systems, analyses and lessons learned. In *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *PMLR*, pages 86–111.

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, Abu Dhabi, UAE. Association for Computational Linguistics.

Antonio Moreno-Sandoval, Jordi Porta, Yanco Torterolo, Alexia Stanescu, Melina Chatzi, and Sofía Roseti. 2026a. 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.

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

Antonio Moreno-Sandoval, Jordi Porta Zamorano, Blanca Carbajo-Coronado, Doaa Samy, Dominique Mariko, and Mahmoud El-Haj. 2023. The financial document causality detection shared task (fincausal 2023). In *Proceedings of the 2023 IEEE International Conference on Big Data (BigData 2023)*, pages 2855–2860, Sorrento, Italy. IEEE.

Jonathan Pilault, Raymond Li, Sandeep Subramanian, and Chris Pal. 2020. On extractive and abstractive neural document summarization with transformer language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pages 9308–9319. Association for Computational Linguistics.

Avinash Trivedi, Gauri Toshniwal, Sivanesan Sangeetha, and SR Balasundaram. 2025.

Sarang at fincausal 2025: Contextual qa for financial causality detection combining extractive and generative models. 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 242–247.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.