Sarang at FinCausal 2025: Contextual QA for Financial Causality Detection Combining Extractive and Generative Models

Avinash Trivedi[1], Gauri Toshniwal[1], Sivanesan Sangeetha[1], S.R. Balasundaram[1]

[1] NIT Trichy, India

Correspondence: avinashtrivedi.2008@gmail.com

Abstract

This paper describes our approach for the FinCausal 2025 English Shared Task, aimed at detecting and extracting causal relationships from the financial text. The task involved answering context-driven questions to identify causes or effects within specified text segments. Our method utilized a consciousAI RoBERTa-base encoder model, fine-tuned on the SQuADx dataset. We further fine-tuned it using the FinCausal 2025 development set. To enhance the quality and contextual relevance of the answers, we passed outputs from the extractive model through Gemma2-9B, a generative large language model, for answer refinement. This hybrid approach effectively addressed the task's requirements, showcasing the strength of combining extractive and generative models. We (Team name: *Sarang*) achieved outstanding results, securing 3rd rank with a Semantic Answer Similarity (SAS) score of 96.74% and an Exact Match (EM) score of 70.14%.

1 Introduction

Causality within financial documents is necessary for understanding financial markets and making informed decisions. Manually extracting causal relationships from financial data is both tedious and time-consuming. Automating this process enhances efficiency and enables the analysis of large volumes of data that would be impractical to handle manually. The FinCausal 2025 shared task (Moreno-Sandoval et al., 2025), part of the Financial Narrative Processing Workshop, focuses on advancing methods for detecting causal relationships in financial texts. The task involves identifying and extracting causes and effects within given segments from financial annual reports, with datasets provided in both English and Spanish. This year's edition introduces a shift from traditional extractive methods to a generative AI framework. Participants must answer abstractive questions about causes or effects, with evaluations based on exact matching and semantic similarity metrics. We started with prompt engineering with Zero-shot and Few-shot Prompting to efficiently explore various LLMs, namely llama3.2-1b-instruct, Llama-3.2-3B-Q8, Llama-3.1-8B-Instruct-Q8_0,mistral-ins-7b-q4, gemma-2-2b-it, gemma-2-9b-it, gemma-2-27b, etc. Our best-performing system is Fine-tuning + Refinement using Gemma2-9B.

The rest of the paper is as follows. Section 2 contains related work, section 3 describes the dataset, section 4 describes our methodology, section 5 contains experimental results, section 6 describes strengths and weaknesses, section 7 provides feedback on the dataset, and section 8 includes conclusions and future work.

2 Related Work

The necessity for precise identification of cause-effect links in domain-specific situations has made the extraction of causal relationships in financial documents a critical task in natural language processing. The FinCausal shared tasks, conducted between 2020 and 2022, have significantly contributed to the advancement of research in financial text analysis by establishing benchmarks for detecting and extracting causal relationships within financial texts. With each successive edition of the event, It introduced more complex datasets and refined evaluation metrics, driving progress and innovation in this domain. The 2020 shared task (Mariko et al., 2020) laid the groundwork by offering a foundational dataset and benchmarks for causal extraction. Subsequent editions in 2021 (Mariko et al., 2021) and 2022 (Mariko et al., 2022) introduced increasingly intricate causal chains, highlighting the limitations of purely extractive approaches and promoting the adoption of hybrid architectures for enhanced performance.

In recent years, hybrid methods that integrate extractive and generative models have demonstrated potential in overcoming these challenges. Authors

Proceedings of the Joint Workshop of the 9th FinNLP, the 6th FNP, and the 1st LLMFinLegal, pages 242–247
January 19–20, 2025.

in (Pilault et al., 2020) proposed a method where an extractive step selects relevant information, which is then summarized and used to condition a transformer language model for text generation. Further, a systematic comparison of generative and extractive readers by (Luo et al., 2022) highlighted that extractive readers often outperform generative ones in short-context QA tasks and exhibit better out-of-domain generalization. NeurIPS 2020 EfficientQA competition (Min et al., 2021) highlights the balance between efficiency and accuracy in QA systems. The competition demonstrated that well-tuned lightweight extractive models can deliver performance close to state-of-the-art performance while avoiding the high computational costs of larger generative models. These findings are especially valuable for scaling hybrid architectures in practical financial applications. Expanding on previous research, our method utilizes RoBERTa (Liu et al., 2019), fine-tuned on the SQuADx and FinCausal 2025 datasets, to accurately identify causal links. To enhance contextual relevance and semantic coherence, we integrate a generative refinement step powered by Gemma2-9B. This hybrid approach effectively combines extractive and generative strategies, achieving high scores in both semantic similarity and exact match evaluations.

3 Dataset for FinCausal2025

In the provided development set, we could load 1996 rows, excluding a few bad entries. We prepared two variants of this data, one as it is, i.e. 1,996 samples and another cleaned version of 1,985 samples, which contains only those entries where the answer is a sub-string of context.

The Development sets are provided in a CSV file format with the following headers: ID, Text, Question and Answer, separated by semicolons (;). Table 1 contains sample data, below is a description of each field:

- **ID**: Example identifier.
- **Context**: The original paragraph extracted from the annual reports.
- **Question**: Designed to identify the other part of the causal relationship, whether cause or effect. The question is always abstractive.
- **Answer**: The answer will be the cause or effect previously questioned, extracted verbatim from the text, making it extractive.

The evaluation dataset includes only the ID, Context and Question fields; we are supposed to extract an Answer.

4 Methodology

4.1 Zero-shot and few-shot prompting on LLMs

Initially, we started with various manually crafted prompts on LLMs; later, we refined those prompts and applied Zero-shot and Few-shot prompting. We have observed significant performance boost after prompt refining but with limitations in achieving optimal performance beyond a certain SAS and exact match score. After that, any further change in prompts led to performance reductions. All these attempts used the entire development set to decide a better choice of prompt and LLM. The experiments have been performed on llama.cpp[1] server using model-specific GPT-Generated Unified Format (GGUF) files[2]. We observed the best configuration is gemma-2-9b-it + better prompt + post-processing. Its corresponding prompt is available in the Appendix B.

4.2 Best performing system

The architecture of our best-performing system is illustrated in Fig 1. It consists of two stages, Extractive QA and Answer enhancement. The first step involves preprocessing the raw input data using text normalization, which lowercases and eliminates excess white spaces. Next, removing samples where the answer is not in the context using answer verification. The processed data is then converted into SQuAD format to fine-tune a QA model (consciousAI/question-answering-roberta-base-s)[3], enabling it to extract precise answers. The fine-tuning configurations are detailed in Table 3. The second stage focuses on enhancing the extracted answers. Post-processing removes unwanted characters (e.g., full stops and commas) for cleaner outputs. The processed answers are then refined using a large language model (gemma2-9b-it (Gemma Team, 2024)), ensuring improved quality and alignment with the context. Another round of post-processing removes extraneous prefixes (e.g., "Answer:") to produce polished final outputs.

This two-stage system ensures high-quality answers by combining robust preprocessing, fine-tuned extraction, and enhancement by utilizing an instruction-following prompt in Fig 3 of Appendix A with the gemma2-9b-it model.

ID	Text	Question	Answer
3337	Overall, Group trading continues to be subdued in large part due to legacy issues	What is the main reason why the Group trading continues to be subdued?	legacy issues
3375	Developments in the year: Change of tax laws or practices as a result of base erosion and profit shifting initiatives ("BEPS").	What caused a change of tax laws or practices?	base erosion and profit shifting initiatives ("BEPS")

Table 1: Sample development data

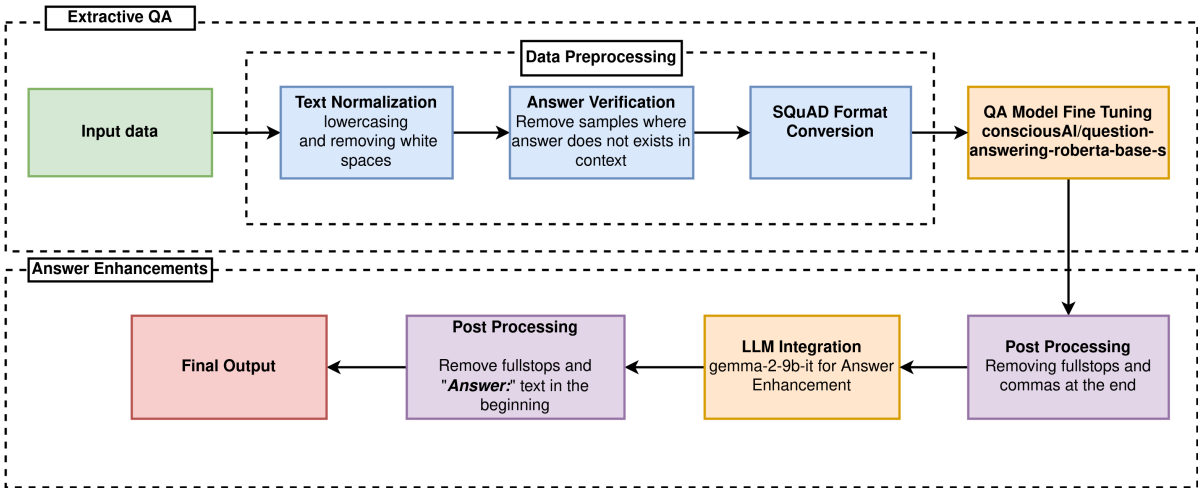

Figure 1: Fine-tuning and Answer enhancement based system architecture

5 Experimental Results

Table 2 contains experimental results of Zero-shot and Few-shot prompting on LLMs. The best-performing model was gemma2-9b-it, but its performance was capped at SAS of 0.9117 and an Exact match of 0.5711 on Evaluation set (**baseline-1** as in Table 4). So we tried Dynamic few-shot prompting, where few-shot examples are considered in the evaluation prompts based on semantically similar Context and Question in the development set. We retrieved this similar examples by calculating the cosine similarity between the concatenated form of sentence embeddings[4] of the Evaluation Context and Question, and the Development Context and Questions (**baseline-2** as in Table 4). Later, we tried two LLMs, both gemma2-9b-it, with different prompts, one acting as Child and another as Parent LLM, the response of Child LLM is appended to the prompt of Parent LLM to correct the child's reply if necessary. It did not perform well, resulting in a reduction in both SAS and Exact Match scores, Since it was only a one-time correction by the Parent.

Model	SAS	Exact Match
llama-3.1-8B-Q8_0	0.8853	0.1608
llama-3.2-3B-Q8	0.6623	0.0220
llama3.2-1b-instruct	0.6623	0.0220
mistral-ins-7b-q4	0.8701	0.2344
mistralLite.Q6_K	0.4229	0.0976
gemma-2-2b-it	0.8660	0.1903
gemma-2-9b-it	0.8974	0.2870
gemma-2-9b-it + post processing	0.9067	0.4549
gemma-2-9b-it + better prompt + post processing	**0.9340**	**0.6052**

Table 2: Performance comparison on development set

Inspired from (Lester et al., 2021), we tried prompt tuning and fine-tuning of gemma-2-2b-it, but in both cases, i.e. prompt tuning and fine-tuning using Quantized Low-Rank Adaptation (QLoRA) (Hu et al., 2021), gemma-2-2b-it was not behaving as expected, so we dropped the idea of prompt tun-

[1]https://github.com/ggerganov/llama.cpp/tree/master
[2]https://huggingface.co/models?library=gguf
[3]https://huggingface.co/consciousAI/question-answering-roberta-base-s
[4]https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

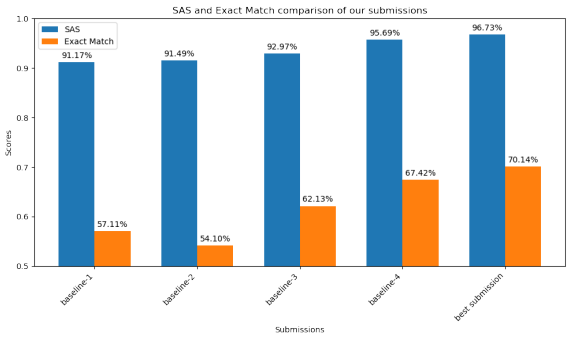

Figure 2: Comparison of SAS and Exact Match among our submissions

ing, which we strongly believe could have given much better result provided we prompt tune either 9B or 27B variant of Gemma2.

Next, we fine-tuned RoBERTa-base on the Development set, using a 90:10 train/validation split. This achieved a SAS score of 0.465 and an exact Match score of 0.071 on validation set. Later we changed the model checkpoint to consciousAI/question-answering-roberta-base-s[5] which is encoder-only (roberta-base) (Liu et al., 2019) with QuestionAnswering LM Head, fine-tuned on SQUADx (Rajpurkar et al., 2016). We tried fine-tuning the consciousAI checkpoint as well and observed a further performance boost (**baseline-3** as in Table 4). The same fine-tuning was tried on the cleaned version of the development set and observed further improvement in SAS and exact match (**baseline-4** as in Table 4). Our final approach involved passing the baseline-4 answers through the enhancement step, utilizing an instruction-following prompt in Fig 3 with the gemma-2-9b-it model to achieve improved results. This configuration achieved the best performance, yielding a SAS score of 0.9674 and an Exact Match score of 0.7014 (**Best system submission** as in Table 4). Fig 2 depicts the comparison of our submissions.

Hyperparameter	Value
learning_rate	2e-5
per_device_train_batch_size	8
per_device_eval_batch_size	8
num_train_epochs	3
weight_decay	0.01
logging_steps	10

Table 3: Hyperparameters values.

[5]https://huggingface.co/consciousAI/question-answering-roberta-base-s

Model	SAS	Exact Match
baseline-1	0.9117	0.5711
baseline-2	0.9149	0.5410
baseline-3	0.9297	0.6213
baseline-4	0.9569	0.6742
Best performing submission	**0.9673**	**0.7014**

Table 4: Major system submissions

6 Strength and Weaknesses

Table 5 contains the unique strengths and limitations of the three approaches. LLM-based models effectively leverage prompt engineering, achieving over 91% SAS, but struggle with issues like text overflow. Meanwhile, PLM-based models excel at identifying the precise start and end of answers, making them suitable for tasks requiring accurate localization, although they occasionally fail to detect an answer entirely. PLM+LLM models combine the advantages of both, addressing many individual weaknesses. However, they still face difficulties in pinpointing the exact start and end of answers, leading to lower exact Match scores.

To overcome these challenges, improved preprocessing techniques to handle text overflow and targeted fine-tuning for boundary detection could further refine the performance of models.

Model	Strength	Weakness
LLM Based	Showing the power of Prompt engineering with > 91% SAS	Text overflow
PLM Based	Able to locate start and end of Answer better than LLM	Sometimes unable to find answer
PLM+LLM	Utilize the best of both worlds to overcome each other's weaknesses	Unable to locate exact start and end of Answer, leads to less Exact Match

Table 5: Strength and Weaknesses of attempted approaches

7 Feedback on the Dataset

Table 6 contains observed issues in the dataset. To address these issues, dataset can be refined by standardizing formatting inconsistencies, such as fixing spacing and hyphenation (e.g., "re-financing" to "refinancing"), removing unnecessary quotation marks and phrases like "Remuneration Policy" or

"Life on land" from answers, and ensuring the context is relevant and aligns with the answers. Additionally, errors or irrelevant responses can be identified and corrected, non-text characters like \xa0 eliminated, and instances where the context mirrors the question can be restructured for clarity. These refinements can be implemented through a combination of automated processes and manual review to improve the dataset for future editions of FinCausal.

ID	Appeared in Context	Appeared in Answer
5221	"natural"	natural
5364.3	one-off	one off
4047	currency- denominated	currency-denominated
3965	""	Extra "Remuneration Policy" in the beginning
5269.3.b	re- financing	re-financing
2564	Context itself is question	
3373	""	Extra "Life on land" in beginning of answer
4093.a	""	not in context + wrong answer
6014.b	reserve	re serve
2587	"natural"	natural
3681.a	\xa0	

Table 6: Issues with development dataset

8 Conclusions and Future Work

We have described all our experimented approaches: Zero-shot, Few-shot, Dynamic Few-shot prompting on various LLMs, Parent-Child LLM, Fine-tuning and a combination of fine-tuning + answer enhancement using LLM. Our best submission achieved an SAS of 0.9674 and an Exact Match score of 0.7014, outperforming initial baselines. In addition, we performed a comparative analysis of the gap in the Exact match.

Future work will focus on resource constraints to fully explore the prompt tuning of larger models. Also, It will be interesting to explore data augmentation to fine-tune the consciousAI checkpoint. In addition, trying LLM agents can not be ruled out.

References

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size.

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.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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. *CoRR*, abs/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, Barcelona, Spain (Online). COLING.

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, Jordan Boyd-Graber, Chris Alberti, Danqi Chen, Eunsol Choi, Michael Collins, Kelvin Guu, Hannaneh Hajishirzi, Kenton Lee, Jennimaria Palomaki, et al. 2021. Neurips 2020 efficientqa competition: Systems, analyses and lessons learned. In *NeurIPS 2020 Competition and Demonstration Track*, pages 86–111. PMLR.

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*.

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)*, pages 9308–9319, Online. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

A Prompt for Enhancement Step

Prompt

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

Figure 3: Prompt for enhancement step

B Better Prompt

Prompt

{ "role": "user",
"content": """ ### Instruction
You will be given a financial text in ### Context.

Definitions
- Cause: The reason why an event occurs.
- Effect: The event that happens as a result of the cause.

Context: {context}
Question: {question}
Answer: """ }

Figure 4: User prompt

Prompt

{ "role": "system",
"content": """ You are an AI assistant specialized in Finance Causal extraction. Your task is to identify and return either the cause or effect as requested, verbatim, from the provided financial text.

Guidelines:
- Focus on extractive responses only, do not add or modify text outside the given context.No added words or rephrasing.
- Ensure responses follow the cause-and-effect relationship: a cause precedes an effect, and an effect follows a cause.

Examples:

Example 1:
CONTEXT: Nationwide is in robust financial health, having achieved profits of over £1 billion for the third consecutive year. Profits allow us to maintain our financial strength, invest with confidence, and return value to members through pricing and service.
QUESTION: What is the effect of achieving profits of over £1 billion for the third consecutive year?
ANSWER: Nationwide is in robust financial health

Example 2:
CONTEXT: All the Directors are resident in the UK, bringing a wide range of skills to the Board. Given the Company's small size and that the Board is comprised of only five Directors, all are members of the Audit Committee and the Nomination and Remuneration Committee.
QUESTION: What is the impact of the Company's small size and having a Board comprised of only five independent Directors?
ANSWER: the Board considers it sensible for all the Directors to be members of the Audit Committee and of the Nomination and Remuneration Committee

Example 3:
CONTEXT: Following a thorough and comprehensive review, we believe that our Remuneration Policy continues to be appropriate, and are therefore proposing the Policy remains broadly unchanged. In recognition of emerging best practice, we have updated our Policy to reduce the pension contribution for new Executive Director appointments to 15QUESTION: What impact had the thorough and comprehensive review?
ANSWER: we believe that our Remuneration Policy continues to be appropriate

Example 4:
CONTEXT: As the Board consists entirely of non-executive directors it is considered appropriate that matters relating to remuneration are considered by the Board as a whole, rather than a separate remuneration committee. All directors are considered independent with the exception of Oliver Bedford who is an employee of Hargreave Hale Limited and is not therefore independent.
QUESTION: What is the reason Oliver Bedford is the only director not deemed as independent?
ANSWER: All directors are considered independent with the exception of Oliver Bedford who is an employee of Hargreave Hale Limited
""" }

Figure 5: System prompt