Extracting Financial Causality through QA: Insights from FinCausal 2025 Spanish Subtask

Marcelo J. Moreno Aviles and **Alejandro Vaca Serrano**
LenguajeNatural.AI
Madrid, Spain
{marcelo.moreno, alejandro.vaca}@lenguajenatural.ai

Abstract

This paper addresses causality detection in financial documents for the Spanish subtask of the FinCausal 2025 challenge. The task involved identifying cause-effect relationships using an extractive question-answering framework. We compared span extraction and generative approaches, with the latter demonstrating superior performance. Our best model, SuperLeNIA, achieved a Semantic Answer Similarity (SAS) score of 0.979 and an Exact Match score of 0.816 on the blind test set.

1 Introduction

Understanding causality in financial documents is crucial for informed decision-making, as it involves identifying true cause-and-effect relationships beyond surface-level correlations. By detecting these, organizations can uncover risks, enhance audit compliance, and gain insights into market trends for more effective strategies. In previous editions of FinCausal (Moreno-Sandoval et al., 2023; Mariko et al., 2022, 2021, 2020), participants identified cause-and-effect spans within causal sentences, typically using pre-trained BERT transformers in a BIO token classification setup. For example, the top-ranked team in the 2023 FinCausal Spanish Subtask, BBVA AI (Algarra and Muelas, 2023), adapted BIO tagging to label each span as C (cause), E (effect), or N (none) and used RoBERTa Base BNE transformer (Gutiérrez-Fandiño et al., 2022).

This edition of FinCausal (Moreno-Sandoval et al., 2025) framed the task as an extractive question-answering problem, where a question based on the cause or effect had to be answered by extracting the relevant part of the relationship. This change allowed the task to be approached either as an extractive question-answering task using span extraction (Keskar et al., 2019), or as a generative task by fine-tuning large language models (LLMs). The challenge included both Spanish and English subtasks, with this paper focusing on the Spanish subtask. We initially tested both approaches using baseline models: the pre-trained RoBERTa Base BNE (Gutiérrez-Fandiño et al., 2022) and our custom LeNIA model (Serrano, 2024b) based on Qwen2 (Yang et al., 2024). Our tests showed that the generative approach performed better, and after further experiments with various LLMs, a private model achieved a Semantic Answer Similarity (SAS) (Risch et al., 2021) of 0.979 and an Exact Match score of 0.816 in the blind test. This paper outlines the complete process from start to finish.

2 Methodology

2.1 Dataset

The training dataset consisted of 2000 data points extracted from a corpus of Spanish financial annual reports from 2014 to 2018. It contained four columns: ID, Text, Question, and Answer. The dataset was divided into two subsets: train and test, containing 1600, and 400 data points, respectively.

ID: 3873
Text: *Durante el verano, tanto los índices en Europa como en Estados Unidos se vieron severamente castigados a raíz de las dudas sobre el crecimiento económico global.*
Question: *¿Cuál es la razón de que los índices en Europa y Estados Unidos se vieran severamente castigados durante el verano?*
Answer: *las dudas sobre el crecimiento económico global*

Figure 1: Example Data Point in the Spanish Subtask

The example data point shown in Figure 1 demonstrates how questions and answers are formulated: in this case, the question is focused on the effect, and the answer extracts the cause. In other

Proceedings of the Joint Workshop of the 9th FinNLP, the 6th FNP, and the 1st LLMFinLegal, pages 265–270
January 19–20, 2025.

instances, the roles are reversed, with the question focused on the cause and the answer providing the corresponding effect. The question is always paraphrased from the context, while the answer is directly extracted from it.

The lower quartile for the word count in the answer was 12, while the upper quartile was 27, indicating answers are relatively short. The max words in an answer was 105, meaning that for most models a *max new tokens* of 256 would be enough during inference.

2.2 Text pre-processing

To prepare the dataset for training, both span extraction and text generation require distinct formats for fine-tuning.

We adapted the SQuAD (Rajpurkar et al., 2016) format for span extraction, keeping the original columns with slight modifications: **id**, **context**, **question**, and **answers**. The answers field contains the **answer_start** (the start position of the answer in the context) and the corresponding **text**. A minor issue was found in 36 data points, where answers didn't exactly match the context due to discrepancies like extra words, grammatical variations, or whitespace differences. To address this, we used Algorithm 1 to extract the closest matching answer, as described in the model inference section.

As for the generative task, we adapted the dataset to fit a conversational format designed for large language models. The conversational format included a brief **system message** explaining the task, which sets the assistant's behavior.

2.3 Baseline Models

The baseline models for this study were selected based on their proven effectiveness in Spanish language tasks. RoBERTa Base BNE (Gutiérrez-Fandiño et al., 2022) has demonstrated strong performance across various Spanish tasks and performed well in the previous FinCausal edition. LeNIA, a generative model, is relatively small for its type, yet it has consistently outperformed other models of similar or greater size across several Spanish language tasks.

RoBERTa Base BNE

The RoBERTa Base BNE (Gutiérrez-Fandiño et al., 2022) model is based upon the original RoBERTa base model (Liu et al., 2019) and has been pre-trained on the largest available Spanish corpus. The version used for our baseline was the RoBERTa Base BNE fine-tuned on the SQAC dataset (Gutiérrez-Fandiño et al., 2022) which is a dataset for Spanish Question Answering based on the SQuAD format (Rajpurkar et al., 2016). The model size is **125 M** parameters.

LeNIA

The LeNIA Model (Serrano, 2024b) is our public model built on the Qwen2 architecture (Yang et al., 2024). It was pre-trained using a corpus of supervised Spanish tasks formatted as FLAN-style instructions (Wei et al., 2022). Subsequent fine-tuning was performed on a variety of Spanish instruction-following datasets and enhanced with a mix of public and proprietary data from Lenguaje-Natural.AI. The model size is **1.5 B** parameters.

Fine-Tuning Models

Hyperparameter	Roberta BNE	LeNIA
Learning Rate	3e-05	5e-05
Epochs	1	1
Batch Size	16	8

Table 1: Hyperparameters for fine-tuning baseline models.

For Roberta Base BNE, standard fine-tuning techniques with Transformers library (Wolf et al., 2020) were employed for a span extraction task, with the dataset formatted according to the SQuAD format as described in Section 2.2. It was trained in ~2 minutes on a Colab instance with an NVIDIA T4 GPU (16 GB VRAM).

For LeNIA, given its larger size, QLoRA (Dettmers et al., 2023) was employed to efficiently fine-tune the model for the generative task, utilizing a chat-based dataset format as described in Section 2.2. The fine-tuning process was conducted using the AutoTransformers library (Serrano, 2024a), which integrates functionalities from both Transformers (Wolf et al., 2020) and PEFT (Mangrulkar et al., 2022) libraries, enabling seamless implementation of QLoRA. The following parameters were used in the QLoRA configuration for targeting all linear modules: rank $r = 128$, $\alpha = 32$, LoRA dropout of 0.1, and 4-bit quantization. It was trained in ~15 minutes on a Colab instance with an NVIDIA L4 GPU (24 GB VRAM).

Some of the relevant hyperparameters for fine-tuning each model are summarized in Table 1. These choices were based on prior experience with similar tasks and model architectures. For instance, a single epoch was selected for both models, as question answering models typically exhibit signs of overfitting after just one epoch of training.

Baseline Results

Model	SAS	Exact Match
Roberta Base BNE	0.820	0.256
LeNIA	0.917	0.553

Table 2: Baseline models results on the test set.

The results presented in Table 2 clearly show that the generative approach to question answering for financial causality significantly outperformed the span extraction approach. Specifically, the generative model achieved an improvement of 0.097 in Semantic Answer Similarity (SAS) and a substantial increase of 0.297 in Exact Match on our test set.

Initial Inference Experiments

Despite achieving a high score, we noticed that our LeNIA model did not always extract the answer directly from the text. Specifically, 72 out of 400 predicted answers were not found in the context, usually due to minor changes in words. To improve results before experimenting with new models, we implemented two strategies. First, we adjusted the *temperature* at inference to 0.1 to reduce randomness in the predictions.

Algorithm 1 Find Closest Answer in Context

Input: Context ctx, Predicted answer ans
if ans is in ctx **then**
  return ans
else
  Define n as word count in ans
  Generate n-grams from ctx
  Use RapidFuzz to match ans with ctx n-grams
  return best match based on similarity score
end if

Secondly, we used Algorithm 1, which uses PolyFuzz library (Grootendorst, 2020), to find the closest match for the predicted answer when it is not directly present in the context.

Strategy	SAS	Exact Match
Temp (0.1)	0.964	0.753
Temp (0.1) + Alg. 1	0.964	0.775

Table 3: LeNIA results on test set with inference strategies.

Implementing these two strategies, the results for LeNIA improved, as shown in Table 3. With these two simple adjustments, the SAS improved by 0.047, while the Exact Match increased by a significant 0.223. These results highlight two key insights: first, that selecting the right inference parameters can have a substantial impact, and second, that ensuring the answer is directly extracted from the context is crucial for achieving a high Exact Match.

2.4 Intermediate Models

Building on these insights, a range of model architectures with varying sizes was explored. For illustrative purposes, only three distinct models, including the best one, each with distinct architectures, sizes, along with their performance will be discussed: LeNIA (2.3), Llama 3.2 Instruct and SuperLeNIA (a private model).

Llama 3.2-3B Instruct

The Llama 3.2-3B Instruct model is built on the Llama 3 architecture (Dubey et al., 2024) and fine-tuned for multilingual dialogue tasks. Pretrained on a mix of publicly available data, it supports multiple languages, including Spanish. The model size is **3.21 B** parameters. This section omits the 8B parameter Llama 3.2 version as it did not achieve the best performance compared to models of similar size.

SuperLeNIA

The SuperLeNIA model is based on a combination of publicly available multilingual models ranging from **7B** to **8B** parameters. Just like the public LeNIA, it was pre-trained using a corpus of supervised Spanish tasks formatted as FLAN-style instructions (Wei et al., 2022) and fine-tuning was performed on a variety of Spanish instruction-following datasets and enhanced with a mix of public and proprietary data from LenguajeNatural.AI. According to internal evaluations, SuperLeNIA outperforms GPT-4o and GPT-4 Turbo in

various Spanish tasks, thus, making it a suitable choice.

Fine-tuning

For the fine-tuning process, the same methodology was applied to both Llama 3.2 and SuperLeNIA, utilizing a generative task framework with QLoRA. The configurations used for fine-tuning these models were consistent with those detailed for LeNIA (2.3).

Llama 3.2 was trained in ~20 minutes on a Colab instance with an NVIDIA L4 GPU (24 GB VRAM). The SuperLeNIA model was trained in ~10 minutes on a cloud instance with an NVIDIA H100 GPU (80 GB VRAM).

Inference Hyperparameter Tuning

As noted in section 2.3, the inference parameters proved to be important. To improve performance, each model was hyper-tuned during inference, using the Optuna (Akiba et al., 2019) framework, with the following hyper-parameters:

- *Temperature*: Controls the model's output randomness, with higher temperatures yielding more diverse responses and lower temperatures making it more deterministic.
- *Top p*: Refers to nucleus sampling (Holtzman et al., 2020), where the model selects from the smallest set of top probabilities whose cumulative sum is greater than or equal to 'p'.
- *Min p*: Sets a minimum threshold for the probability of the next token.

In each iteration, Optuna employs Bayesian optimization, specifically the tree-structured Parzen estimator (TPE) (Bergstra et al., 2011), to select a new set of hyperparameters. For efficient inference, vLLM (Kwon et al., 2023) was employed, enabling the models to generate 10 predictions per sample in each iteration. The prediction with the highest cumulative log probability was then selected and processed using Algorithm 1.

Model	Temperature	Top P	Min P	SAS	Exact Match
LeNIA	0.35	0.92	0.10	0.964	0.775
Llama 3.2	0.06	0.87	0.19	0.968	0.800
SuperLeNIA	0.56	0.74	0.01	0.978	0.835

Table 4: Intermediate model results with hyperparameter tuning.

As presented in Table 4, LeNIA demonstrated no improvement with Hyperparameter Tuning as compared to results on Table 3, while Llama 3.2 achieved a slightly better performance than LeNIA across both metrics. SuperLeNIA outperformed both models, exceeding their scores by a margin of at least 0.01 across both metrics. Thus, SuperLeNIA was chosen as the final model. As observed, the *temperature* for our best-performing model was not particularly low. This could indicate that a more deterministic inference approach may have occasionally restricted the generation of alternative sequences that aligned more closely with the correct answer or that the hyperparameter search wasn't exhaustive enough. Future work should consider conducting a more comprehensive parameter search.

2.5 Error Analysis

All models exhibited similar types of errors at inference. They frequently produced overly long responses (75% or more of the context), indicating difficulty in discerning the most relevant information. Minor phrasing differences like adding unnecessary introductory words (e.g., starting with "a la" instead of just "la") occurred often, impacting exact matches despite their small differences. Additionally, overly short responses, though less common, occasionally missed essential context. These issues significantly affected exact match scores but had a less pronounced impact on SAS.

3 Results

The blind test set, used for submitting predictions for evaluation in the FinCausal 2025 Competition, consisted of 500 data points. The SuperLeNIA model achieved a SAS score of **0.979**, attaining **3rd** place among participating teams. Additionally, it attained an Exact Match score of **0.816**, ranking **4th** in this metric.

4 Conclusion

This paper presented a comprehensive approach to addressing the FinCausal 2025 Spanish subtask, which required extracting causality relationships in financial texts using a question-answering framework. By focusing on financial causality, this work highlights how LLMs can potentially play a role in understanding cause-effect relationships within financial contexts, enabling more accurate analysis and decision-making.

We explored multiple model architectures, fine-tuning methodologies, and inference optimization strategies. Our experiments demonstrated the effectiveness of generative models over span extraction models, with the SuperLeNIA model achieving the highest performance among the models evaluated. The results emphasize the importance of model selection, inference hyperparameter tuning, and text-processing techniques in QA tasks.

Future works could explore the integration of the model into a retrieval-augmented generation (RAG) system. Making it useful for uncovering root causes of risks, improving audit compliance, and providing deeper insights into market trends through its ability to extract causality.

5 Limitations

This study has several limitations that require attention. First, the methods developed are primarily tailored to Spanish financial documents, which may limit their effectiveness in other languages with different syntactic structures or more complex morphology.

Additionally, the approach may not generalize well to all types of financial documents or causality relationships. Financial documents can vary widely depending on the industry, region, or specific financial context, and the model may need further fine-tuning or domain adaptation to handle the nuances of different financial contexts. The temporal limitation is also a factor, as financial trends, regulations, and language usage may have evolved after 2018, potentially affecting the model's applicability to more recent documents.

Moreover, the context and answers were relatively short, but as document length increases, capturing and extracting causal relationships over extended contexts may become challenging. This issue may require additional pre-processing and testing the models capabilities of processing longer texts.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631.

Alberto Algarra and David Muelas. 2023. BBVA AI Factory at FinCausal 2023: a RoBERTa Fine-tuned Model for Causal Detection . In *2023 IEEE International Conference on Big Data (BigData)*, pages 2798–2801, Los Alamitos, CA, USA. IEEE Computer Society.

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. 2011. Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Maarten Grootendorst. 2020. Polyfuzz: Fuzzy string matching, grouping, and evaluation.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquin Silveira-Ocampo, Casimiro Pio Carrino, Carme Armentano-Oller, et al. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, page 39–60.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. *Preprint*, arXiv:1904.09751.

Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Unifying question answering, text classification, and regression via span extraction. *Preprint*, arXiv:1904.09286.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, et al. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, et al. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

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.

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*.

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 *2023 IEEE International Conference on Big Data (BigData)*, pages 2855–2860.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *Preprint*, arXiv:1606.05250.

Julian Risch, Timo Möller, Julian Gutsch, and Malte Pietsch. 2021. Semantic answer similarity for evaluating question answering models. *Preprint*, arXiv:2108.06130.

Alejandro Vaca Serrano. 2024a. Autotransformers: A library for automatic training and benchmarking of transformer models. `https://github.com/lenguajenatural-ai/autotransformers`.

Alejandro Vaca Serrano. 2024b. Lenia chat.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, et al. 2022. Finetuned language models are zero-shot learners. *Preprint*, arXiv:2109.01652.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, et al. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.