# PresiUniv at FinCausal 2025 Shared Task: Applying Fine-tuned Language Models to Explain Financial Cause and Effect with Zero-shot Learning

**Medha Jeenoor, Madiha Aziz, Saipriya Dipika Vaidyanathan, Avijit Samantraya, Sandeep Mathias**
Department of Computer Science and Engineering
Presidency University, Bangalore
**Correspondence:** sandeepalbert@presidencyuniversity.in

## Abstract

Transformer-based multilingual question-answering models are used to detect causality in financial text data. This study employs BERT (Devlin et al., 2019) for English text and XLM-RoBERTa (Conneau et al., 2020) for Spanish data, which were fine-tuned on the SQuAD datasets (Rajpurkar et al., 2016) (Rajpurkar et al., 2018). These pre-trained models are used to extract answers to the targeted questions. We design a system using these pre-trained models to answer questions, based on the given context. The results validate the effectiveness of the systems in understanding nuanced financial language and offers a tool for multi-lingual text analysis. Our system is able to achieve SAS scores of 0.75 in Spanish and 0.82 in English.

## 1 Introduction

As the growing connectivity of global markets and the rising use of multiple languages in communication continue, the need for a model that can interpret text data has become increasingly important. Question Answering (QA) is a key component in extracting or identifying relevant data across domains. Traditionally, QA models have been trained separately for individual languages, resulting in fragmented systems that are costly to maintain and difficult to scale. Although some multilingual models such as Typologically Diverse Question Answering (TyDiQA) (Clark et al., 2020) and Multilingual Knowledge Questions and Answers (MKQA) (Longpre et al., 2021) have been introduced in recent years, they often struggle with maintaining accuracy in non-English languages or processing large datasets efficiently. These limitations underscore the gap between current technologies and the demands of modern multilingual applications (Lioutas et al., 2020).

In light of this, we decided to participate in the 2025 FinCausal Shared Task. The goal of the shared task is to create a strong and effective multilingual system for English and Spanish that can cater to international markets.

## 2 Problem Statement

The FinCausal 2025 Shared Task[1] focuses on the extraction of causal relationships from financial reports (Moreno-Sandoval et al., 2025). The task involves processing financial reports to identify explicit and implicit causal relationships between financial events, entities, or market factors. Participants are required to develop models that can accurately detect these causal links, taking into account the complex, and often ambiguous nature of the financial language.

This task expands on earlier work in extracting causal relationships, which has been studied in areas like event extraction (Angeli et al., 2010) and causal inference in news data. Unlike prior editions of the shared task, this edition challenges participants to handle diverse financial contexts with increased accuracy and scalability from financial reports (Moreno-Sandoval et al., 2025).

The aim of the Shared Task is to advance the field of financial event analysis by providing robust, scalable methods for causal extraction in real-world financial data (Moreno-Sandoval et al., 2025).

## 3 Related Work

Research on multilingual Question-Answering has advanced significantly, frequently as a result of shared assignments that address many aspects of the QA pipeline. The Question Answering for Machine Reading Evaluation (QA4MRE) (Peñas et al., 2013) tasks which were organized at CLEF from 2011 to 2013, focused on machine reading comprehension across languages, was one of the first multilingual QA challenges. The best-performing

[1] https://www.lllf.uam.es/wordpress/fincausal-25/

*Proceedings of the Joint Workshop of the 9th FinNLP, the 6th FNP, and the 1st LLMFinLegal*, pages 259–264
January 19–20, 2025. ©2025 Association for Computational Linguistics

systems used hybrid strategies to enhance their reasoning abilities across multilingual texts by fusing rule-based techniques with machine learning models.

The advent of datasets like TyDi QA (Clark et al., 2020) marked a turning point for multilingual QA by emphasizing typological diversity. This dataset aimed to provide a benchmark for systems handling typologically distinct languages, such as Swahili and Finnish. Participants in shared tasks built on TyDi QA used techniques ranging from fine-tuned transformer-based models, such as Multilingual BERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020) to multi-task learning for better performance on low-resource languages.

In 2020, the MKQA shared task (Longpre et al., 2021) emphasized the evaluation of systems on a translated version of the Natural Questions dataset. The challenge revealed that translation-based evaluation often introduces biases, as noted by the top participants. These teams leveraged cosine similarity measures and context-sensitive embeddings from pre-trained models to tackle semantic drift during translation (Longpre et al., 2021).

The SemEval 2022 Multilingual News Article Similarity shared task required systems to handle domain-specific and multilingual inquiries. In order to enhance performance in a variety of settings, winning entries combined cross-lingual retrieval models with Retrieval-Augmented Generation (RAG) frameworks (Lewis et al., 2021). An effective technique for solving contextual ambiguity and improving substitute generation in multilingual contexts is prompt engineering on large-language models (Guo et al., 2023). These shared tasks and their evolving methodologies have significantly shaped the development of efficient QA systems, demonstrating the interaction between dataset design, evaluation strategies, and model capabilities in advancing multilingual NLP.

## 4 Dataset

The dataset used in the shared task has 2 tracks for 2 different languages - English and Spanish - consisting of data from financial annual reports in those languages. Further details of the dataset can be found in Moreno-Sandoval et al. (2025).[2]

The Shared Tasks organisers provided three sets of data for both languages. The reference and training datasets have 4 columns namely “ID”, “Text”, “Question” and “Answer”. The “ID” column is an identifier for each instance of the data. The “Text” column contained the context which has both, the cause and the effect. The “Question” column was the question that was asked, and the “Answer” column is the expected answer. The testing dataset had the first 3 columns as the training dataset, and the shared task was to predict the answer. Questions in the dataset required participants to use the given text data to either identify the cause(s) given the effect(s) or vice versa for the financial data. All the columns are delimited by a semicolon (;).

| Dataset Type | English | Spanish |
|---|---|---|
| Reference | 101 | 101 |
| Training | 2000 | 2001 |
| Testing | 499 | 501 |

Table 1: Details of the Dataset in both languages.

Table 1 summarizes the number of data points for each language (where each data point consists of the “ID”, “Text”, “Question”, etc. fields).

## 5 System

In this section, we describe our system.

### 5.1 Resources Used

The resources which we used for the question-answering tasks in our project involve:

- Transformer-based pre-trained models (Eg. BERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020)) for generating the answers for the provided context-question pairs.
- Python libraries for input and output data processing in CSV format.
- Transformers library from Hugging Face (Wolf et al., 2020) for accessing and executing the QA pipelines.

For each of the languages, we used different pre-trained language models. For English, we used the BERT large model fine-tuned on the SQuAD (Rajpurkar et al., 2016) dataset[3]. For Spanish, we used a variant of XLM-RoBERTa (Conneau et al., 2020) which was pre-trained on the SQuAD 2.0 (Rajpurkar et al., 2018) dataset[4].

[2] Further details of the competition are found here.

[3] English model name: google-bert/bert-large-uncased-whole-word-masking-finetuned-squad

[4] Spanish model name: deepset/xlm-roberta-large-squad2

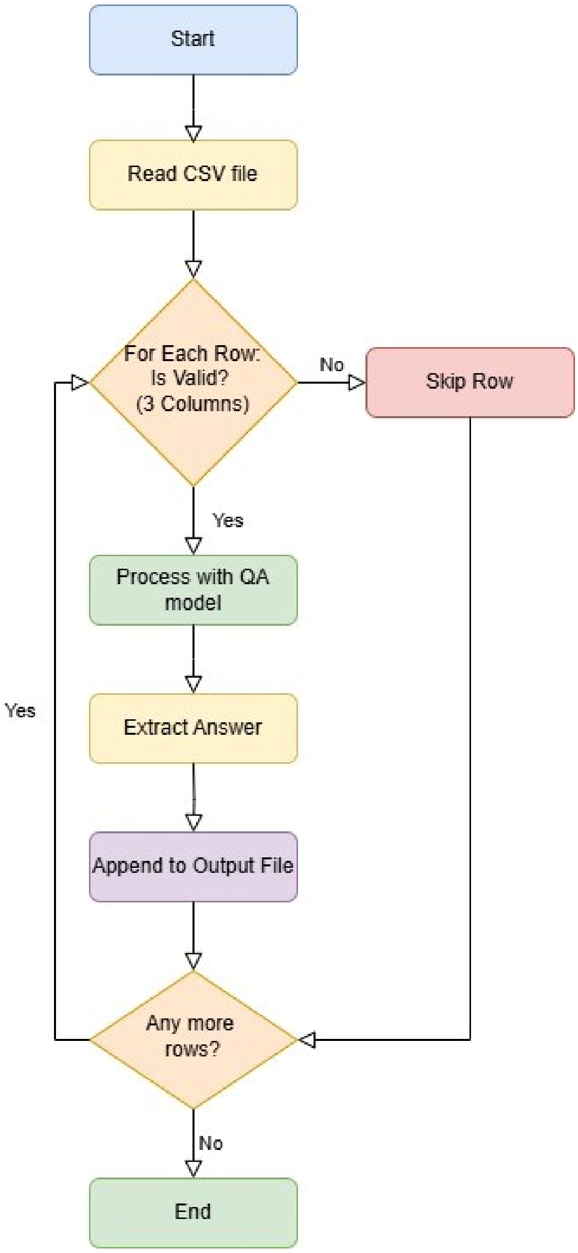

Figure 1: Workflow of our system.

### 5.2 Workflow

Figure 1 describes our workflow. In our task, we perform zero-shot learning by using the pre-trained language models which have been finetuned on the SQuAD datasets.

For each row, we first check if the row is valid (i.e. it has 3 columns, corresponding to the "ID", "Text", and "Question"). We then extract the context and question from the row, and generate a response from the pre-trained language model (either XLM-RoBERTa or BERT). After that, we add the relevant row to our output file. Prior to submission, we add the header and submit the file for evaluation on CodaLab.

For example, consider that we have the following row from the English dataset: "1882.b;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.;What has motivated the increase in UK overheads by £1.0 million or 5.6%?".

Our system will generate the line: "1882.b;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.;What has motivated the increase in UK overheads by £1.0 million or 5.6%?;investment in support of our strategic initiatives."

### 5.3 Evaluation Metrics

The shared task systems were evaluated on 2 evaluation metrics - Semantic Answer Similarity (SAS) (Risch et al., 2021) and Exact Match (EM) (Baker, 1978). SAS evaluates the semantic similarity between the predicted and reference answers, while EM reflects the verbatim match accuracy.

## 6 Results and Analysis

In this section, we report and analyze our results.

### 6.1 Comparison with Different Pre-trained Language Models

Table 2 shows the comparison of different systems which we explored for selecting our model. We achieved SAS: 0.8241 and EM: 0.2244 for English, and SAS: 0.7520 and EM: 0.0140 for Spanish. Based on the results, we selected the BERT Large model which was fine-tuned on the SQuAD dataset for English and the XLM-RoBERTa model fine-tuned on SQuAD 2.0 for Spanish.

Some of the other systems that we tried - RoBERTa (for English) (Liu et al., 2019), Helsinki-NLP MarianMT ((Tiedemann et al., 2023), (Tiedemann and Thottingal, 2020)) and GPT 4o-mini[5] (for Spanish) - did not perform as well.

### 6.2 Error Analysis

Our model, BERT, pre-trained on SQuAD dataset, excelled in handling straightforward question-answer pairs. However, the Exact Match (EM)

[5] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

| Language | Large / Pre-Trained Language Model | SAS Score |
|---|---|---|
| English | bert-large-uncased-whole-word-masking-finetuned-squad | **0.824** |
| English | deepset/roberta-base-squad2 | 0.818 |
| Spanish | Helsinki-NLP MarianMT translation models (translating to English) | 0.713 |
| Spanish | deepset/xlm-roberta-squad2 | **0.752** |
| Spanish | OpenAI GPT 4o-mini (temperature=0.3) | 0.735 |

Table 2: Comparison of different systems that we tried. The best performing systems are in **boldface**.

| Team | SAS |
|---|---|
| TU Graz Data Team | 0.9841 |
| Team nirvanatear | 0.9801 |
| LenguajeNatural.AI | 0.9787 |
| LaithTeam | 0.9756 |
| CUFE | 0.9755 |
| Aukbc | 0.9607 |
| Semantists | 0.9555 |
| OraGenAIOrganisation | 0.9219 |
| RGIPT – India | 0.8987 |
| **PresiUniv** | **0.7520** |
| Yanco | 0.7244 |

Table 3: Results on the Spanish Dataset, ranked by SAS. Our system's best performance is in **boldface**.

| Team | SAS |
|---|---|
| Team nirvanatear | 0.9779 |
| TU Graz Data Team | 0.9732 |
| Sarang | 0.9674 |
| Aukbc | 0.9604 |
| Semantists | 0.9598 |
| LaithTeam | 0.9598 |
| CUFE | 0.9595 |
| OraGenAIOrganisation | 0.9244 |
| RGIPT – India | 0.9086 |
| **PresiUniv** | **0.8241** |
| Yanco | 0.7373 |

Table 4: Results on the English Dataset, ranked by SAS. Our system's best performance is in **boldface**.

score was impacted by the extractive nature of the task. Our answers directly extracted the relevant phrase rather than forming complete sentences tailored to the question.

Consider the following example from the dataset: "I joined Columbus because I believed in the underlying assets and I recognized quickly that I would be able to build a strong, capable team around me." For the question "What led him to join Columbus?", the answer generated by our model was "I believed in the underlying assets", as opposed to a more contextualized sentence like "He believed in the underlying assets and felt that he could strongly contribute." While this approach impacted the EM score, the SAS score remained high as the extracted answer phrases were semantically aligned with the ground truth, even if not the same.

### 6.3 Comparison with Other Teams

Tables 3 and 4 show the comparison of our system with various other submitted systems. In both languages, we achieved a peak performance SAS (Risch et al., 2021) score in excess of 0.75. This was achieved without using any training data, and only the pre-trained language models which were fine-tuned on the SQuAD (Rajpurkar et al., 2016) and SQuAD 2.0 (Rajpurkar et al., 2018) datasets.

## 7 Conclusion and Future Work

Our transformer models demonstrated the capability to extract and predict cause-effect relationships from financial data. This system not only enhances the analytical process of complex multilingual financial documents, but also fosters data-driven decision-making to promote economic stability. While the model did not achieve the best overall performance, it exhibited a strong semantic understanding of the data. However, further refinements and fine-tuning would help us achieve better verbatim matching and a better understanding with domain-specific nuances in diverse datasets.

In the future, we plan to enhance our model by incorporating an explainability module to provide human-readable explanations for causal predictions, thereby improving user trust and interpretability. We also plan to explore the model's multilingual capabilities by including additional languages and implementing cross-lingual transfer learning to address linguistic nuances more effectively.

## References

Gabor Angeli, Percy Liang, and Dan Klein. 2010. A simple domain-independent probabilistic approach to generation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 502–512, Cambridge, MA. Association for Computational Linguistics.

Theodore P Baker. 1978. A technique for extending rapid exact-match string matching to arrays of more than one dimension. *SIAM Journal on Computing*, 7(4):533–541.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *Preprint*, arXiv:2301.07597.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Vasileios Lioutas, Ahmad Rashid, Krtin Kumar, Md. Akmal Haidar, and Mehdi Rezagholizadeh. 2020. Improving Word Embedding Factorization for Compression Using Distilled Nonlinear Neural Decomposition. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2774–2784, Online. 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. *Preprint*, arXiv:1907.11692.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

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.*

Anselmo Peñas, Eduard Hovy, Pamela Forner, Alvaro Rodrigo, Richard Sutcliffe, and Roser Morante. 2013. Qa4mre 2011-2013: Overview of question answering for machine reading evaluation. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization: 4th International Conference of the CLEF Initiative, CLEF 2013, Valencia, Spain, September 23-26, 2013. Proceedings 4*, pages 303–320. Springer.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. 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.

Julian Risch, Timo Möller, Julian Gutsch, and Malte Pietsch. 2021. Semantic answer similarity for evaluating question answering models. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 149–157, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato Yves Scherrer, Raul Vazquez, and Sami Virpioja. 2023. Democratizing neural machine translation with OPUS-MT. *Language Resources and Evaluation*, (58):713–755.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 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.