**Proceedings Paper:**
Alqarni, A., Stevenson, M. and Laksito, A. (2026) Sheffield NLP at FinCausal 2026: A comparative study of RAG approaches and fine-tuning for causal Q&A in financial texts. In: Sandoval, A.M. and Martinez, P., (eds.) Proceedings of The 7th Financial Narrative Processing Workshop (FNP 2026). The 7th Financial Narrative Processing Workshop (FNP 2026) @ LREC 2026, 16 May 2026, Palma de Mallorca, Spain. , pp. 125-131. ISBN: 9781952148255.

# Sheffield NLP at FinCausal 2026: A Comparative Study of RAG Approaches and Fine-Tuning for Causal Q&A in Financial Texts

**Aali Alqarni, Mark Stevenson and Arif Laksito**
School of Computer Science, University of Sheffield
Sheffield, United Kingdom
{aaalqarni1, mark.stevenson, alaksito1}@sheffield.ac.uk

**Abstract**

This paper describes our approach to the FinCausal 2026 shared task, which addresses causal question answering from financial documents in English and Spanish. We investigated the effectiveness of fine-tuned generative models combined with Retrieval-Augmented Generation (RAG). Our approach compares five retrieval strategies across base and fine-tuned GPT models (GPT-4.1-mini). RAG-based few-shot selection performed better than random sampling, particularly for the base model. In the FinCausal 2026 official run, this approach was ranked first in both the English and Spanish sub-tasks, obtaining LLM scores of 4.8140 and 4.8131 out of 5, respectively.

**Keywords:** Question and Answering (Q&A), Causality, Large Language Model (LLM), Generative Pre-trained Transformer (GPT), Retrieval-Augmented Generation (RAG)

## 1. Introduction

Financial documents, including earnings reports and financial analysis articles, contain rich causal relationships that drive market movements and trends. Understanding causality in these documents is fundamental for risk assessment, investment decision making, and market analysis (Cormack et al., 2009). Manual extraction of these relations is time-consuming and labour intensive. Automated methods offer a solution but face challenges due to the complexity of financial language and the diversity of causal expressions (Li et al., 2024; Cao et al., 2022).

To advance research in this area, the FinCausal 2026 shared task (Moreno-Sandoval et al., 2026) builds on the 2025 edition, which introduced a question and answering (Q&A) framework for extracting causal relationships from financial texts (Moreno Sandoval et al., 2025). The core objective of the 2026 edition remains the identification of events that explain financial outcomes such as revenue changes, market movements, or corporate decisions, while also introducing more rigorous benchmarks. The dataset has been expanded with over 500 complex causal cases, including multi-element causal chains, and questions have been rephrased to reduce reliance on sentence-level similarity. The task has also shifted to LLM-as-a-judge evaluation rather than the Exact Match (EM) and Semantic Answer Similarity (SAS) measures used previously (Risch et al., 2021).

### 1.1. Task Description

**Objective:** Given a natural language question, $Q$, and corresponding financial text context, $C$, systems must extract a verbatim span, $A$, from $C$ that captures the underlying causal relationship.

**Example:**

$Q$: "What explains their strong performance in health and safety?"

$C$: "As a result of **these very high standards and relentless focus**, we have a strong performance in health and safety" ($A$ indicated by **bold font**).

## 2. Related Work

The FinCausal shared tasks have been organised as a benchmark for evaluation methods that detect causal relationships in financial texts. Early editions (2020-22) formulated causality detection as either causal sentence classification or cause–effect span identification, with most systems relying on extractive models such as BERT and BiLSTM-CRFs (Mariko et al., 2020, 2022). The 2023 edition extended the task to a multilingual setting by requiring systems to identify causal relationships in financial texts written in English and Spanish. This edition marked a notable transition to the use of generative large language models (LLMs) such as GPT (Moreno-Sandoval et al., 2023). The 2025 edition reframed the task as a Q&A problem, requiring systems to extract answer spans from financial contexts given a natural language question.

**LLM-based causal extraction.** Recent work in cause–effect span detection tasks has explored GPT-based approaches for causal extraction from financial texts. Shukla et al. (2023) combined Retrieval-Augmented Generation (RAG) with few-shot prompting using GPT-4, while the LTRC IIITH team explored chain-of-thought (CoT) prompting to enhance reasoning performance (Moreno-Sandoval et al., 2023). These approaches suggested that in-context learning can effectively iden-



*Proceedings of the 7th Financial Narrative Processing Workshop @ LREC 2026*, pages 125–131

| ID | Context | Question |
|---|---|---|
| **English** | | |
| 29 | **At the start of the year, we also reduced the size of our transport fleet by 10% in light of network changes. At this point, we decided against further reducing the size of the fleet due to the future demand from new contracts**. As a result, the business carried significant excess costs of under-utilised trucks during the current financial year. | What accounts for the business carrying significant excess costs derived from under-utilized trucks during the current financial year? |
| **Spanish** | | |
| 76 | Por el contrario, **la Comunidad de Estados Independientes registró un descenso de la producción del 28%. Esta reducción fue menos acusada que la de la demanda interna** debido a las exportaciones. | ¿Qué efecto tuvieron las exportaciones? |

Table 1: Example entries from the English and Spanish datasets. Answers are highlighted in **bold** within the context.

tify cause–effect relationships without fine-tuning.

**LLM-based Q&A.** Niess et al. (2025) compared the performance of extractive models, such as BERT, with generative LLMs, including Llama, in zero-shot and few-shot settings for causal Q&A. Their findings indicated that instruction-tuned LLMs without task-specific fine-tuning were competitive. However, a fine-tuned Llama model trained on a multilingual dataset significantly outperformed both approaches.

Beyond stand-alone model comparisons, recent work has explored the integration of LLMs within retrieval-augmented Q&A frameworks to enhance factual grounding and reduce hallucinations. For instance, Yang et al. (2026) proposed Structured-Semantic RAG (SSRAG), a hybrid architecture that incorporates LLM-based query augmentation, prompt-guided source routing, and unified graph–vector retrieval to improve answer faithfulness across open-domain Q&A benchmarks. In the medical domain, Aljohani and Alsanoosy (2026) proposed a modular hybrid RAG framework that integrates sparse retrieval (BM25) with dense biomedical retrieval (MedCPT; Jin et al. 2023) to enhance medical Q&A. Their approach demonstrated substantial improvements in retrieval recall, precision, and generation faithfulness across PubMedQA, MedMCQA, and MedQA-US benchmarks (Pal et al., 2022; Jin et al., 2019, 2021). Despite these advances, limited work has systematically examined hybrid retrieval approaches for domain-specific causal Q&A in financial texts, especially in multilingual settings. This gap highlights the need for retrieval strategies that are specifically designed for causal financial Q&A tasks.

## 3. FinCausal 2026 Dataset

The FinCausal 2026 dataset (Moreno-Sandoval et al., 2026) contains English text from UK financial reports dated 2017 and Spanish text from Spanish financial reports dated 2014 to 2018. Entries consist of an abstractive question asking about a cause or effect, a context passage, and a verbatim answer extracted from the context.

The dataset includes different types of causality. Some have explicit causal markers such as 'due to' and 'because', while others require reasoning from the context to identify implicit causal connections. It includes complex cause-and-effect relationships, such as causal chains of three or more elements where multiple events are linked sequentially (e.g., $A$ causes $B$, which leads to $C$). Questions can be classified into cause-seeking (e.g., 'What caused the...?'), effect-seeking (e.g., 'What was the impact of...?'), and others that do not follow a specific causal pattern (e.g., 'What did these considerations entail?').

**Dataset Description.** Context passages vary in length from a single sentence to multiple sentences, and answer spans range from a short phrase to multiple clauses. Table 1 shows representative examples from the English and Spanish datasets.

**Dataset Statistics.** The dataset comprises 2,000 labelled training instances per language, with 500 blind test instances for English and 503 for Spanish. We applied an 80:20 split to the training data. This resulted in 1,600 instances per language for fine-tuning and retrieval indexing, and 400 per language for development (including ablation studies and prompt engineering).

## 4. Methodology

### 4.1. Model Selection

A generative approach was adopted following previous work which demonstrated that fine-tuned generative models notably outperform traditional extractive Q&A systems (Moreno Sandoval et al., 2025). Figure 1 shows the overall system architecture.

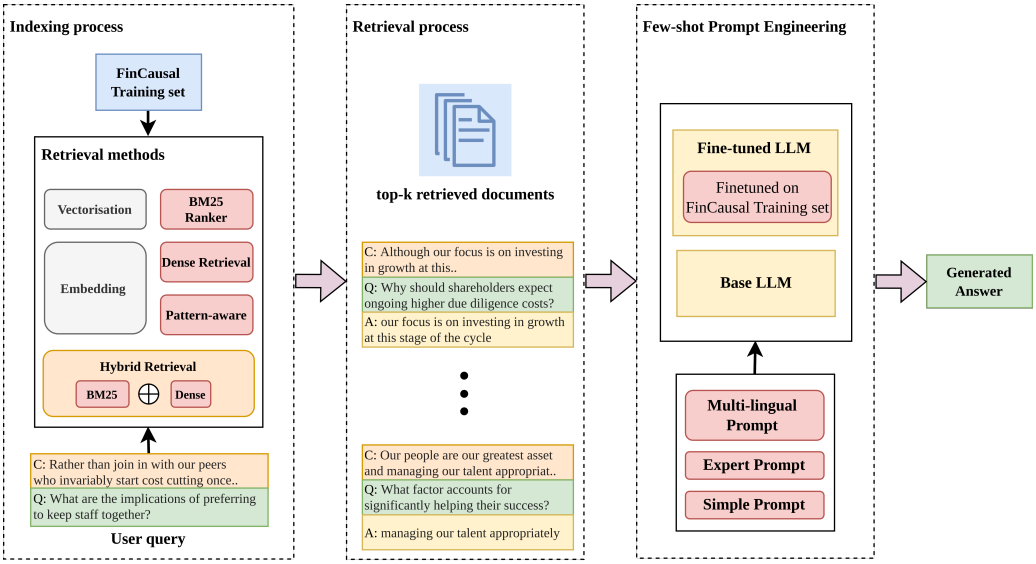


Figure 1: Overview of system architecture. The pipeline consists of three stages: (1) **Indexing**, where the training set is indexed using multiple retrieval methods (Random, RAG-BM25, RAG-Dense, RAG-Pattern, and RAG-Hybrid); (2) **Retrieval**, where the top-$k$ most relevant training examples are retrieved for each test query; and (3) **Few-shot Prompt Engineering**, where the retrieved examples are combined with the test query and passed to either the base or fine-tuned LLM with different prompt strategies to generate the answer.

## 4.2. Prompt Engineering

Prompt engineering was employed to guide the GPT model to answer causal questions from the provided passages. The model was configured with a temperature of 0.1 and top-p of 1 to ensure deterministic outputs across multiple runs, and a maximum token limit of 512. We began with zero-shot prompting and progressively refined our approach using few-shot examples from the training set. To meet the task's extractive requirements, we implemented strict instructional constraints: the model was explicitly directed to answer the given question by extracting the answer verbatim from the context $C$ and it was strictly prohibited from paraphrasing.

Three prompting strategies were explored, as illustrated in Figure 1: (1) a *simple prompt*, which provides minimal instructions to extract the answer from the context; (2) an *expert prompt*, which assigns the model a domain-expert role (e.g., "You are a financial causal analysis expert") and includes detailed extraction rules; and (3) a *multilingual prompt*, which extends the expert prompt with language-aware instructions to handle both English and Spanish inputs (e.g., "The passage and question may be in Spanish.").[1]

Two strategies were explored for selecting few-shot examples:

**1) Random sampling.** We sampled $k$ examples randomly from the training set. The model was provided with $k$ question-context-answer triplets as examples before being asked to answer the target question. We experimented with $k \in \{5, 10\}$.

**2) RAG-based approaches.** The most relevant training examples for each test instance were retrieved using the RAG approach (Lewis et al., 2020). Multiple retrieval strategies were explored: (1) BM25 (Robertson and Zaragoza, 2009), a sparse lexical retrieval method widely used for relevance ranking in information retrieval (RAG-BM25); (2) a dense semantic retrieval approach that encodes texts into dense vector representations and retrieves the most semantically similar training examples based on vector similarity (RAG-Dense) (Karpukhin et al., 2020); (3) a pattern-aware variant (RAG-Pattern), in which each question is first classified into a causal template (CAUSE, EFFECT, or OTHER), and retrieval is restricted to examples of the same type; and (4) a hybrid approach that combines RAG-BM25 and RAG-Dense using Reciprocal Rank Fusion (RRF) (Cormack et al., 2009) to merge the ranked lists into a single unified ranking (RAG-Hybrid).

For RAG-BM25, every question–context pair in the training dataset was indexed using a BM25 retriever with PyStemmer-based language-specific stemming for both English and Spanish.[2] For RAG-Dense, embeddings were generated for the same training examples using OpenAI's `text-`

[1] Prompt templates: https://github.com/aaalqarni/fincausal-2026/blob/main/prompts/README.md

[2] https://pypi.org/project/PyStemmer/

`embedding-3-large`[3] model and indexed using LlamaIndex's in-memory vector store (Liu, 2022). For RAG-Pattern, each question was first classified into the causal class using regular expression patterns, and retrieval was restricted to training examples of the same template class. At inference time, each test question–context pair was embedded using the same encoder, and the top-$k$ most similar training examples were retrieved and provided as few-shot examples to the base and the fine-tuned GPT-4.1-mini models.

### 4.3. Fine-tuning

Following the approach of the top-performing teams in FinCausal 2025 (Moreno Sandoval et al., 2025), GPT-4.1-mini was fine-tuned on the provided training data using OpenAI's fine-tuning API. The training set was formatted as question-context-answer in JSONL format, where each entry contained the question, context, and expected extractive answer. The training process ran for 3 epochs with a learning rate multiplier of 2 and a batch size of 3. This configuration was specifically chosen to accelerate convergence and to avoid overfitting, given the limited size of the dataset.[4]

## 5. Experimental Setup

**Experimental Stages.** Experiments were conducted in three different stages. First, various prompt templates and few-shot configurations ($k \in \{0, 5, 10\}$) were evaluated on the development set using the base GPT-4.1-mini model. Second, GPT-4.1-mini was fine-tuned on the 3,200 training instances and its performance was benchmarked against the base version across all retrieval strategies. Third, ablation studies were conducted on the five retrieval methods described in Section 4, varying the number of few-shot examples. In the hybrid configuration, RRF was used with equal weighting between the lexical (RAG-BM25) and semantic (RAG-Dense) retrieval examples.

**Final Submission.** For the final submission, the retrieval indexes were rebuilt using all 2,000 training instances per language, and predictions were generated on the blind test sets using the best-performing configuration from the development experiments.

**Evaluation Metrics.** EM and SAS (Risch et al., 2021) are reported for the development set. EM measures whether the predicted answer exactly matches the gold answer string. SAS measures the semantic similarity between the predicted and gold answers using a cross-encoder model. For SAS, `all-MiniLM-L6-v2` is used for English and `paraphrase-multilingual-MiniLM-L12-v2` for Spanish. The official competition ranking uses LLM-as-a-judge scoring on the blind test set, which rates system outputs on a 1–5 adequacy scale.

[3] https://developers.openai.com/api/docs/models/text-embedding-3-large

[4] https://developers.openai.com/api/docs/guides/model-optimization

## 6. Results and Discussion

For English, the simple prompt performed best, while the multilingual prompt produced the best results for Spanish. Table 2 presents our results across all configurations for both English and Spanish sub-tasks. Results showed a large performance gap between the base and fine-tuned models. Fine-tuning GPT-4.1-mini on bilingual (EN+ES) data increased English EM from .3533 to .8875 and Spanish EM from .0250 to .8625. This improvement was consistent across all retrieval configurations, suggesting that fine-tuning helps the model follow the strict verbatim extraction requirement rather than relying only on in-context examples.

For the fine-tuned model, all retrieval strategies produced comparable results. English EM ranged from .8650 to .8875, and Spanish EM from .8425 to .8625. The difference between the best configuration RAG-Dense ($k$=5) and the worst was below 2 percentage points in both languages. In contrast, the base model benefited more from retrieval. For example, RAG-Hybrid ($k$=10) achieved .5950 EM in English compared to .3533 in the zero-shot setting. This suggests that retrieval approaches play a larger role when the model is not fine-tuned, whereas fine-tuning reduces the additional gains from complex retrieval strategies. RAG-Pattern showed higher EM for the base model in English (RAG-Pattern $k$=10: .7550 vs RAG-Dense $k$=10: .5875), but underperformed on SAS and showed no consistent gains for the fine-tuned model, suggesting that fine-tuning implicitly captures causal directionality.

**RAG benefits (for Spanish).** The base model showed lower zero-shot performance on Spanish (EM = .0250) compared to English (EM = .3533), likely due to a higher prevalence of English financial corpora in the pre-training data. However, RAG strategies narrowed this gap. Specifically, RAG-Dense ($k$=10) increased the Spanish EM score to .5075, a substantial relative improvement over the zero-shot baseline. This suggests that RAG few-shot examples are particularly beneficial in lower-resource scenarios.

**Official Blind Test Results.** On the official blind test set, evaluated using LLM-as-a-judge, the fine-tuned model with RAG-Dense ($k$=5) achieved the highest English score (4.8140), while RAG-Dense ($k$=10) obtained the best Spanish score (4.8131).

| | | | English | | | Spanish | | |
|---|---|---|---|---|---|---|---|---|
| | | | Validation | | Blind | Validation | | Blind |
| **Model** | **Mode** | $k$ | **EM** | **SAS** | **LLM Score** | **EM** | **SAS** | **LLM Score** |
| GPT-4.1-mini | Zero-shot | 0 | .3533 | .8605 | 4.4600 | .0250 | .8876 | 4.3002 |
| | Random | 5 | .5450 | .9083 | 4.4620 | .3450 | .9282 | 4.3857 |
| | | 10 | .5850 | **.9354** | 4.4760 | .3925 | .9383 | 4.4076 |
| | RAG-BM25 | 5 | .5675 | .9113 | 4.4920 | .4400 | .9393 | 4.4513 |
| | | 10 | .5800 | .9135 | 4.5280 | .4850 | .9395 | 4.5189 |
| | RAG-Dense | 5 | .5550 | .9136 | 4.5780 | .4525 | .9347 | 4.4712 |
| | | 10 | .5875 | .9151 | 4.5840 | **.5075** | .9408 | **4.5030** |
| | RAG-Pattern | 5 | .7300 | .8593 | 4.6220 | .4075 | .9406 | 4.4135 |
| | | 10 | **.7550** | .8622 | **4.6500** | .4175 | **.9432** | 4.4334 |
| | RAG-Hybrid | 10 | .5950 | .9152 | 4.5940 | .4575 | .9372 | 4.4712 |
| Finetuned GPT-4.1-mini (EN+ES) | Zero-shot | 0 | .8800 | .9755 | 4.7440 | .8550 | .9734 | 4.7853 |
| | Random | 5 | .8800 | .9763 | 4.7940 | .8600 | .9735 | 4.7952 |
| | | 10 | .8675 | .9776 | 4.7860 | .8575 | .9732 | 4.8012 |
| | RAG-BM25 | 5 | .8775 | .9741 | 4.7980 | .8550 | .9740 | 4.7932 |
| | | 10 | .8825 | .9770 | 4.7920 | .8475 | .9734 | 4.8012 |
| | RAG-Dense | 5 | **.8875** | **.9790** | **4.8140** | **.8625** | .9720 | 4.8032 |
| | | 10 | .8825 | .9759 | 4.7940 | .8425 | .9698 | **4.8131** |
| | RAG-Pattern | 5 | .8650 | .8796 | 4.7380 | .8500 | **.9780** | 4.7714 |
| | | 10 | .8700 | .8801 | 4.7080 | .8550 | .9780 | 4.7773 |
| | RAG-Hybrid | 10 | .8850 | .9773 | 4.7980 | .8600 | .9724 | 4.7932 |

Table 2: Results across English and Spanish sub-tasks. EM (Exact Match) and SAS (Semantic Answer Similarity) are computed on the 400-instance validation set per language. LLM denotes the official blind test score evaluated by an LLM-as-judge on a 1–5 scale. Bold values indicate the best result per metric within each model group.

| **Model** | **Mode** | **k** | **English** | **Spanish** |
|---|---|---|---|---|
| Finetuned GPT-4.1-mini | RAG-Dense | (5, 10) | **4.8140** | **4.8131** |
| *Our Ranking* | | | *1st (tie)* | *1st* |

Table 3: Official rankings based on LLM scores (out of 5) on the blind test set for the FinCausal 2026 shared task. The system used 5-shot prompting for English and 10-shot prompting for Spanish.

These configurations were ranked first among all participating systems for both sub-tasks, as illustrated in Table 3. A small variance was observed across fine-tuned settings (4.7440–4.8140 for English; 4.7853–4.8131 for Spanish), indicating stable performance across different RAG approaches. These results demonstrate that fine-tuned GPT-4.1-mini combined with RAG-Dense is the most effective approach for causal Q&A in financial texts.

**Error Analysis.** Error analysis was conducted on the development set, as gold answers for the official test set were not released to participants. We analysed the mismatched predictions from the best-performing fine-tuned configuration RAG-Dense ($k$=5) and identified two main types of errors: (1) *span boundary errors*, where the model extracted a slightly longer or shorter span than the gold answer, typically by including or omitting a leading clause; and (2) *causal direction confusion*, where the model confused the direction of causality when multiple cause–effect relations were present in the context. For instance:

$Q$: "Why did Legendary increase its stake in Virtual Stock from 6.8% to 7.2%?"

$C$: "Subsequent to this investment and also in July 2017, Legendary increased its stake in Virtual Stock from 6.8% (subsequent to **the dilution due to Notion's investment**) to 7.2% increasing the carrying value of its investment in Virtual Stock to £4.3m." ($A$ indicated by **bold font**)

$A$ (predicted): *increasing the carrying value of its investment in Virtual Stock to £4.3m*.

These errors highlight the difficulty of distinguishing between causal explanations and subsequent financial outcomes in complex financial narratives.

## 7. Conclusion and Future Work

This paper presented a systematic comparison of RAG approaches combined with fine-tuning for causal Q&A in financial texts in English and Spanish. Fine-tuning GPT-4.1-mini on bilingual data improved performance across both languages. RAG-Dense further improved performance, achieving the best results in both sub-tasks. RAG-Pattern improved base model performance but showed no consistent gains after fine-tuning, suggesting that fine-tuning reduced reliance on causal pattern matching. Our system ranked first in both sub-tasks. Future work will explore more advanced RAG approaches and extend the approach to other domains and languages.

## Code Availability

Code to reproduce experimental results reported in this paper is publicly available[5].

## Bibliographical References


Bushra Aljohani and Tawfeeq Alsanoosy. 2026. Enhancing medical question answering with llms via a hybrid retrieval-augmented generation framework. *Information*, 17(2):133.

Lang Cao, Shihua Zhang, and Juxing Chen. 2022. Cbcp: A method of causality extraction from unstructured financial text. In *Proceedings of the 2021 5th International Conference on Natural Language Processing and Information Retrieval*, NLPIR '21, page 135–140, New York, NY, USA. Association for Computing Machinery.

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 758–759, New York, NY, USA. Association for Computing Machinery.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

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. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Ying Li, Xiaosha Xue, Zhipeng Liu, Peibo Duan, and Bin Zhang. 2024. Implicit-causality-exploration-enabled graph neural network for stock prediction. *Information*, 15:743.

Jerry Liu. 2022. LlamaIndex.

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, Stéphane Durfort, Hugues de Mazancourt, and Mahmoud El-Haj. 2020. Financial document causality detection shared task (fincausal 2020).

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. 2026. 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. To appear.

Antonio Moreno-Sandoval, Jordi Porta-Zamorano, Blanca Carbajo-Coronado, Doaa Samy, Dominique Mariko, and Mahmoud El-Haj. 2023. The


[5] https://github.com/aaalqarni/fincausal-2026

financial document causality detection shared task (fincausal 2023). In *2023 IEEE International Conference on Big Data (BigData)*, pages 2855–2860.

Georg Niess, Houssam Razouk, Stasa Mandic, and Roman Kern. 2025. Addressing hallucination in causal Q&A: The efficacy of fine-tuning over prompting in LLMs. 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 253–258, Abu Dhabi, UAE. 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.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Neelesh K Shukla, Raghu Katikeri, Msp Raja, Gowtham Sivam, Shlok Yadav, Amit Vaid, and Shreenivas Prabhakararao. 2023. Investigating large language models for financial causality detection in multilingual setup. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2866–2871. IEEE.

Tianyi Yang, Nashrah Haque, Vaishnave Jonnalagadda, Yuya Jeremy Ong, Zhehui Chen, Yanzhao Wu, Lei Yu, Divyesh Jadav, and Wenqi Wei. 2026. Augmenting question answering with a hybrid rag approach.

## Language Resource References

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering.

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

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering.