LeedsMEng26: Qwen + Gemini for FinCausal 2026 Causality Detection in Financial Narrative Texts

Ayomide Ivienagbor*, **Idrees Asad***, **Rijul Shrestha***
Yasemin Bal*, **Zahaab Nadeem***, **Zaid Shahrouri***
*University of Leeds, Leeds, United Kingdom
sc22ai@leeds.ac.uk, sc22i2a@leeds.ac.uk, sc22r2s@leeds.ac.uk,
sc22y2b@leeds.ac.uk, sc22zn@leeds.ac.uk, sc21z2s@leeds.ac.uk

Abstract

This paper presents the LeedsMEng26 system for the FinCausal 2026 shared task Moreno-Sandoval et al. (2026a) on financial causality detection in narrative texts. The task is formulated as extractive question answering over English and Spanish financial reports, where systems must return a verbatim span from the context that answers an abstractive question about a cause or an effect. We propose a two-stage pipeline consisting of candidate span generation followed by span verification and boundary refinement under a strict extractiveness constraint. We evaluate both an extractive RoBERTa-based baseline and instruction-tuned large language models. Results show that Qwen-2.5-1.5B-Instruct is a stronger candidate generator than the RoBERTa baseline, and that a second-stage verifier further improves answer boundary accuracy and overall adequacy. Our best configuration, Qwen-2.5-1.5B-Instruct with Gemini-2.5-flash refinement, achieved an adequacy score of 4.7000 for English and 4.6143 for Spanish. These findings suggest that a modular generation-and-verification pipeline is effective for extractive financial causality detection.

Keywords: Financial Narrative Processing, Causality Detection, Natural Language Processing, Question Answering, Multilingual NLP, Financial Text Analytics

1. Introduction

Financial narratives such as annual reports, earnings announcements, management commentaries, and regulatory filings play a central role in how firms communicate with investors, regulators, and the public. These documents often describe not only what has happened but also why it occurred. They do this through explicit or implicit causal statements, such as linking changes in performance to macroeconomic events or strategic decisions. Understanding these causal links is essential for interpreting corporate performance, assessing risk, and forming expectations about future developments. Manually analysing narratives at scale is costly and time-consuming, particularly given the volume and growing complexity of financial disclosures. This has motivated increasing interest in applying natural language processing (NLP) to financial text.

Within this emerging area, financial causality detection focuses on identifying cause–effect relations in financial documents. Automatically extracting such relations can support tasks such as risk analysis, fraud detection, forecasting, and explainable decision support. It also enhances transparency by making implicit reasoning patterns in corporate communication more explicit and machine-interpretable. Despite these benefits, financial text presents several challenges: it is often technical, domain-specific, and highly contextual, and causal language may be expressed in subtle, indirect or multi-sentence forms. Moreover, financial narratives frequently combine quantitative data with qualitative explanations, requiring models to integrate numerical reasoning with linguistic understanding. Recent shared tasks, including the FinCausal track at the Financial Narrative Processing (FNP) workshop, provide benchmark datasets and evaluation protocols for studying these problems in a controlled setting, thereby enabling systematic comparison of modelling approaches.

This paper describes the LeedsMEng26 system for the FinCausal 2026 shared task (Moreno-Sandoval et al., 2026b), which formulates financial causality detection as question answering over English and Spanish financial texts. Given an abstractive question and a short context paragraph, the system must return an extractive span from the context that answers either the cause or the effect. We study both extractive and instruction-tuned generative approaches under a strict extractiveness constraint, and propose a two-stage pipeline: (i) candidate span generation and (ii) span verification and boundary refinement using a verifier LLM constrained to copy a contiguous substring.

2. Related Work

The FinCausal shared tasks have progressively advanced research on causality detection in financial narratives, moving from span extraction in earlier editions toward question-answering and more generative evaluation settings in recent years. The current edition, FinCausal 2026, continues the multilingual English–Spanish setting and retains the use of annotated contexts paired with abstractive questions and extractive answers. However, it introduces two notable changes: a new random parti-

Proceedings of the 7th Financial Narrative Processing Workshop @ LREC 2026, pages 160–168

tioning of the 2026 dataset and an LLM-as-a-judge evaluation metric, which scores responses on a 1–5 adequacy scale and replaces the previous Semantic Answer Similarity (SAS) and Exact Match (EM) metrics used in 2025 (Moreno-Sandoval et al., 2025).

Earlier editions focused more on span- and token-level causality extraction. FinCausal 2023 expanded the task across English and Spanish, using span-level Exact Match (EM) and token-level weighted F1 for evaluation, while encouraging multilingual and prompt-based approaches (Moreno-Sandoval et al., 2023). FinCausal 2022 focused on causality in quantified financial facts, with the winning SPOCK system using an ensemble of RoBERTa-Large sequence-tagging models with the BIO scheme (Mariko et al., 2022). Overall, the FinCausal series shows a progression from structured cause–effect extraction to multilingual reasoning, QA-based formulations, and more flexible generative evaluation.

(Moreno-Sandoval et al., 2025) introduces FinCausal 2025 competition entries and it was used as a guide to see where the most recent advancements for the task were. Participants adopted a range of approaches, spanning discriminative extractive QA models, generative LLMs with prompt engineering (simple, CoT, few-shot), and varying use of fine-tuning and quantization. Notably, strong scores were achieved even without fine-tuning in some cases, highlighting that fine-tuning was not the only route to competitive performance.

The (Trivedi et al., 2025) system paper was useful for our work because it provided a strong, task-aligned example of a hybrid extractive to generative refinement pipeline (RoBERTa-based span prediction followed by Gemma2-9B refinement) that directly targets common QA boundary and coherence errors while remaining competitive on the official SAS/EM metrics.

3. Dataset and Task

The FinCausal 2026 task is formulated as a generative question answering problem over financial annual reports, where systems must identify either the cause or the effect corresponding to an abstractive question. We decided to participate in both sub-tasks, training models on both the English and Spanish texts.

The training data are drawn from the FinCausal 2026 dataset Moreno-Sandoval et al. (2026b), which is provided in CSV format, with each file containing 2000 rows. Each instance was in the following form: (i) **ID** - its unique identifier; (ii) a **context**, corresponding to a paragraph extracted from a financial report; (ii) an **abstractive question**, asking for either the cause or the effect of a described event; and (iii) an **answer**, which is an extractive span taken precisely from the context. Although the question is abstractive, the expected output is a text span grounded strictly in the provided paragraph.

The 2026 edition introduces a revised and expanded dataset, including more complex causal structures and rephrased questions designed to require deeper reasoning. The training and test sets are randomly partitioned from this updated dataset. The task focuses on explanatory cause–effect relations within the text, especially where specific events result in measurable financial outcomes.

For our submission, we approached the problem from both an extractive and a generative perspective. We first experimented with span-based extraction models to exploit the extractive nature of the gold answers, and subsequently explored fine-tuning and combining large language models to assess whether generative approaches could better capture complex causal structures.

4. Methodology

4.1. Task Formulation

Financial causality extraction is approached as an extractive question answering (QA) task. Each instance consists of a financial context c and a question q specifying a causal relation. The system produces an answer span a that must appear verbatim within c. We apply this extractive constraint strictly to all system variants. This ensures that predictions are always drawn from the provided context.

We report span-level Exact Match (EM) and Semantic Answer Similarity (SAS) metrics. EM evaluates strict boundary correctness. SAS accounts for minor boundary deviations. The official leaderboard score is also tracked where applicable.

4.2. Pipeline Overview

The proposed approach utilises a two-stage framework to enhance boundary accuracy and preserve the extractive constraint:

1. **Candidate generation:** an extractive QA model produces a candidate span $\hat{a}$ from the context.

2. **Span verification and refinement:** an instruction-following LLM receives $(c, q, \hat{a})$ and either confirms $\hat{a}$ or corrects span boundaries, while being constrained to return a verbatim substring of c.

This design distinctly separates relevant evidence retrieval in stage 1 from boundary correction

and consistency verification in stage 2. Preliminary error analysis revealed that boundary truncations, such as missing causal qualifiers, and boundary overruns, such as the inclusion of adjacent clauses, are the most frequent failure modes in pure extractive models. The refinement stage specifically addresses these errors while maintaining strict adherence to the input context.

Figure 1 illustrates the two-stage candidate generation and refinement pipeline used in our system.

All systems employ a unified preprocessing & post-processing pipeline implemented using the Hugging Face `transformers` library. Inputs are constructed by concatenating the question and context with the model-specific tokeniser. For extractive QA models, start and end position labels are derived from the gold answer span.

Predicted spans undergo lightweight normalisation as follows: (i) whitespace trimming and deduplication, and (ii) minor punctuation trimming at span boundaries where it does not alter content. The extractive constraint is enforced through substring verification, requiring the produced answer to appear as a contiguous substring in the given context. If a refinement model outputs text that violates this constraint, the system reverts to the pre-refinement candidate.

4.3. Baseline 1: Extractive QA with RoBERTa

The first baseline employs `question-answering-roberta-base-s-v2`, a RoBERTa-base encoder fine-tuned for extractive QA. Given (c, q), the model produces probability distributions over context tokens for the start and end indices of the answer span. The highest-scoring (i, j) pair is selected, and the corresponding substring is returned verbatim.

We fine-tuned the model on the official FinCausal English training data. We optimised based on the extractive QA objective, minimising cross-entropy loss over the gold start and end token positions. This baseline serves as a strictly extractive reference point with no generative components.

4.4. Baseline 2: RoBERTa + GPT Refinement

As a result, it was concluded that RoBERTa was sensitive to boundary errors that affected the consistency of meanings. Early truncation and overrun errors were the main issues; consequently, a second-stage refinement step is introduced using GPT-3.5 as a verifier.

The refiner is provided with the context, question, and candidate span predicted by RoBERTa. It is instructed to:

- Output *only* the final answer span (no explanation),
- Copy the span verbatim from the context (no paraphrasing),
- Keep the candidate unchanged if correct, and
- Correct boundary errors by expanding or contracting to the shortest correct substring when multiple valid spans exist.

The method ensures compliance with the context and achieves improved span boundary precision. If the refined span is not a substring of the context, the original candidate is used.

4.5. Baseline 3: Qwen-2.5-1.5B-Instruct as Candidate Generator

The third baseline replaces the RoBERTa extractor with Qwen-2.5-1.5B-Instruct, a decoder-only instruction-tuned LLM with multilingual capability. Unlike encoder-style QA models that predict token positions, Qwen generates text directly; therefore, constrained prompting is used to enforce extractive behaviour.

The candidate-generation prompt explicitly requires the model to output a verbatim substring from the context and nothing else. Despite these directives, we still observed boundary drift and occasional paraphrasing. These were similar to those seen with RoBERTa when the model attempted to be 'helpful.'

Low-Rank Adaptation (LoRA) is employed, updating only low-rank matrices inserted into selected attention and feed-forward layers while keeping the base weights frozen, enabling Qwen to adapt to the task with limited computational resources and resulting in reduced training memory and cost compared to full fine-tuning.

Reinforcement learning (RL) post-training is explored using a composite reward. The reward encourages extractive correctness and semantic fidelity. It combines: (i) Exact Match (EM), (ii) a semantic similarity component between the prediction and the gold span, and (iii) an auxiliary judge score reflecting span correctness and boundary precision. At inference time, we use conservative decoding (low temperature) to reduce variability and discourage paraphrasing. Qwen-2.5-1.5B-Instruct was also fine-tuned using Low-Rank Adaptation (LoRA) for parameter-efficient fine-tuning. (Schulman and Lab, 2025)

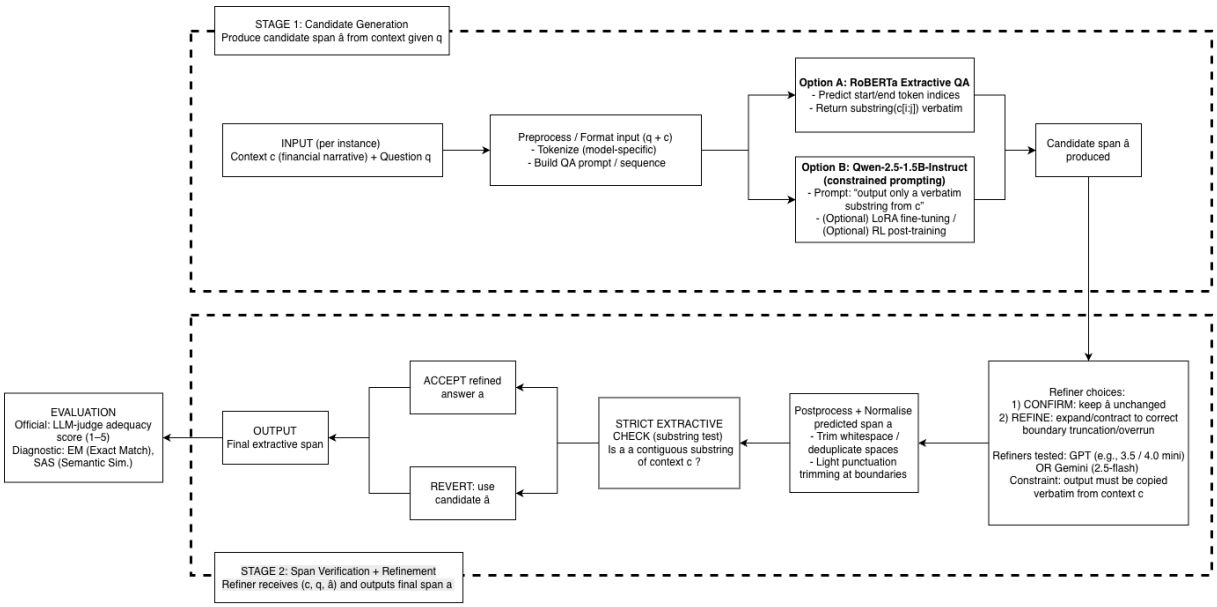

Figure 1: Two-stage pipeline for financial causality extraction.

4.6. Hybrid: Qwen + GPT Refinement

Qwen generates an initial candidate span under strict prompting, and the verifier model performs second-stage verification and boundary correction, replicating the two-stage process used in the RoBERTa + verifier configuration.

We perform substring verification after refinement as before. If GPT-based refinement (GPT-3.5 and GPT-4.0 mini) violates extractiveness, the system reverts to the original Qwen candidate. This hybrid approach uses Qwen's strong instruction-following together with GPT-3.5 and GPT-4.0 mini for verification, reducing hallucination and improving boundary precision. The following configuration further explores this two-stage strategy by integrating a different refinement model, as outlined next.

4.7. Final System: Qwen + Gemini Refinement

Ultimately, we used Gemini-2.5-Flash as the refinement model. Qwen generates a candidate span using strict extractive prompting, then provides the prediction to Gemini and is instructed to verify correctness and adjust the boundaries if needed.

Similar to previous hybrids, we apply minimal normalisation and substring verification, reverting to the candidate span if refinement breaks extractiveness. This configuration provided the most consistent boundary corrections and the strongest overall performance across English and Spanish settings. The next sections outline the experimental setup supporting these results.

4.8. Refinement Prompt

The refinement stage relies on a constrained prompt designed to correct boundary errors while preserving the extractive constraint.

The verifier model receives the context, question, and candidate span produced by Qwen and is instructed to return only the final corrected answer.

The prompt enforces several rules: (i) the output must be a verbatim span from the context, (ii) no explanations or additional text may be produced, (iii) the candidate span should be returned unchanged if it is already correct, and (iv) if multiple spans are possible, the shortest correct substring should be selected.

A shortened version of the prompt is shown below.

> You are correcting a short answer. Output only the final answer. Return a verbatim span from the context. Do not add explanations. If the answer is already correct, return it unchanged. If multiple spans are possible, choose the shortest correct span.

The full prompt, including the in-context examples used during inference, is provided in Appendix A. A Spanish version of the prompt with identical constraints was used for Spanish inputs.

4.9. Experimental Setup

- **Data:** Experiments used the FinCausal 2025 dataset. Training was performed using only the English data. Spanish examples were evaluated without additional training.

- **Data Split:** The English dataset consisted of 2000 instances, which were divided into 75% training and 25% validation data.
- **RoBERTa Training:** The parameters used were learning rate of 3×10^{-5}, weight decay 0.01, and a linear scheduler with warmup ratio 0.1. The batch size was 4 with gradient accumulation of 2 (effective batch size 8). The best model was selected based on validation loss.
- **Qwen Fine-tuning:** Qwen-2.5-1.5B-Instruct was fine-tuned using LoRA with rank $r = 16$, $\alpha = 32$, and dropout 0.0. Bias parameters were not adapted, and gradient checkpointing was enabled using Unsloth.
- **Decoding:** Low-temperature decoding was used during inference. Outputs were constrained to be substrings of the input context to ensure extractive answers.
- **Reproducibility:** Random seeds and preprocessing were kept consistent across all experiments.

5. Experiments and Results

5.1. Evaluation Protocol

The FinCausal 2026 shared task uses an LLM-as-a-judge evaluation protocol. Each system prediction is scored on a 1–5 adequacy scale. The judge's score rewards answers that fully address the question using evidence from the context. It penalises truncations, overruns, and non-extractive outputs.

In addition to the official adequacy score, we report **Exact Match (EM)** and **Semantic Answer Similarity (SAS)** as *diagnostic* metrics to analyse span boundary quality and semantic correspondence during development. Unless stated otherwise, EM/SAS are computed by comparing predictions against available labelled data and are used for internal evaluation rather than leaderboard ranking.

5.2. English Results

System Configuration	Score
RoBERTa + GPT-3.5	3.8200
Qwen-2.5-1.5B-Instruct + GPT-3.5	4.5080
Qwen-2.5-1.5B-Instruct	4.6060
Qwen-2.5-1.5B-Instruct + GPT-4.0 mini	4.6240
Qwen-2.5-1.5B-Instruct + Gemini-2.5-flash	4.7000

Table 1: Best performing English submission per system configuration (LLM-judge adequacy score).

Table 1 summarises the best-performing English submission for each system configuration. Overall, performance improves monotonically as the candidate generator becomes more instruction-aligned and as verification-based refinement is introduced.

The RoBERTa-based hybrid baseline (**RoBERTa + GPT-3.5**) achieves an adequacy score of **3.8200**. While the extractive QA model reliably returns substrings from the context, qualitative inspection indicates that it repeatedly faces minor span boundary errors (e.g., missing causal qualifiers or including adjacent clauses), which reduces adequacy under judge-based scoring.

Replacing the encoder-based extractor with an instruction-tuned decoder model yielded substantial improvement. **Qwen-2.5-1.5B-Instruct** achieves **4.6060**, indicating that constrained, instruction-following generation is better aligned with the causality-oriented QA prompts and the judge's emphasis on completeness and adequacy.

Adding a second-stage verifier provides further gains. **Qwen + GPT-4.0 mini** reaches **4.6240**, indicating that refinement helps correct residual boundary variances and improves answer adequacy.

The best-performing English configuration is **Qwen + Gemini-2.5-flash**, achieving **4.7000**. This result supports the effectiveness of a generation verification pipeline in which a strong refiner corrects subtle span boundary errors while preserving the extractive constraint.

For diagnostic analysis, Table 3 reports EM and SAS on labelled English data. Both metrics increase consistently across configurations (EM from **0.3500** to **0.7415**; SAS from **0.8850** to **0.9460**), indicating that adequacy gains are accompanied by improvements in boundary accuracy and semantic correspondence, rather than reflecting superficial formatting differences.

5.3. Spanish Results

System Configuration	Score
RoBERTa + GPT-3.5	3.9264
Qwen-2.5-1.5B-Instruct + GPT-4.0	4.4692
Qwen-2.5-1.5B-Instruct	4.5030
Qwen-2.5-1.5B-Instruct + Gemini-2.5-flash	4.6143

Table 2: Best performing Spanish submission per system configuration (LLM-judge adequacy score).

Spanish results (Table 2) follow similar trends, with the strongest performance again obtained by verification-based refinement using Gemini-2.5-Flash.

The RoBERTa + GPT baseline achieves **3.9264**. The standalone **Qwen-2.5-1.5B-Instruct** model im-

proves adequacy to **4.5030**, demonstrating strong cross-lingual generalisation in Spanish under extractive prompting.

Unlike in English, adding GPT-based refinement slightly reduced performance (**4.4692** vs **4.5030**), suggesting that the refiner may occasionally overcorrect boundaries or produce outputs that are less well aligned with the judge's adequacy preferences for Spanish instances. In contrast, **Qwen + Gemini-2.5-flash** yields the best Spanish score of **4.6143**, indicating that Gemini provides more reliable verification and boundary correction in the bilingual setting.

5.4. Overall Analysis

Across both languages, two observations are deduced. First, instruction-tuned candidate generation (Qwen) better aligns with the task format and produces more adequate spans with limited prompting. Secondly, a second-stage verifier further improves robustness using correcting boundary truncations and overruns, with **Gemini-2.5-flash** providing the most reliable refinement among tested models. RoBERTa achieved lower scores on both the English and Spanish tasks, indicating lower effectiveness than the alternative approaches evaluated.

Overall, the hybrid **Qwen + Gemini-2.5-Flash** system achieved the best results in both English and Spanish, indicating that generation verification pipelines are effective for extractive financial causality tasks evaluated using an adequacy-oriented LLM judge.

Model	EM Accuracy	SAS
RoBERTa + GPT-3.5	0.3500	0.8850
Qwen-2.5-1.5B-Instruct + GPT-3.5	0.6295	0.9155
Qwen-2.5-1.5B-Instruct	0.6855	0.9366
Qwen-2.5-1.5B-Instruct + Gemini-2.5-flash	0.7415	0.9460

Table 3: Diagnostic EM and SAS results for English (computed on labelled data for development).

6. Strengths and Limitations

A key strength of our work is that, despite limited time and resources, we achieved strong competitive performance, placing 7th in the English category with only marginal differences from teams ranked above us. This result demonstrates that our approach was effective and robust even under practical constraints.

Our pipeline also benefits from a clear and structured design that combines the reliability of extractive question answering with the contextual reasoning benefits of instruction-tuned Large Language Models and verification/refinement steps applied after the initial prediction. By progressing from a strong extractive baseline (RoBERTa) to an instruction-following model (Qwen) and then adding refinement stages (GPT-3.5, GPT-4.0 mini, and Gemini-2.5-Flash), we were able to isolate which components contributed most to performance gains and reduce common QA errors such as span boundary drift. In particular, our two-stage candidate and verifier setup made the system more reliable. The verifier checks the initial span and fixes common issues such as answers being too short (truncated) or too long, while also ensuring the final output stays strictly extractive. This iterative setup also made our experiments more interpretable, since each stage had a clear and measurable role in improving Exact Match and semantic alignment.

Despite the effectiveness of our pipeline, several constraints limited the scope of our experiments and likely capped performance:

- **Time constraints due to schedule clashes.** The shared task timeline overlapped with our university exam period, causing us to begin experimentation later than planned and reducing the time available for broader hyperparameter searches and ablation studies.
- **Limited compute access.** We did not have access to the university's powerful GPUs, which restricted our ability to train larger models, run longer fine-tuning schedules, or explore more compute-intensive approaches. A single NVIDIA RTX 3050 was used for the reinforcement learning finetuning of Qwen-2.5-1.5B.
- **Restricted access to the newest proprietary LLM APIs.** We were unable to use the latest frontier LLM APIs. Access to stronger models for refinement and verification could plausibly have improved boundary correction and reduced rare failure cases.
- **Training focused only on English.** Our main training and optimisation effort was concentrated on the English dataset rather than fully training separate systems for both English and Spanish. This likely reduced Spanish performance relative to what could be achieved with language-specific fine-tuning and validation.

7. Dataset Feedback

One limitation of the dataset is its relatively small size, which limited the amount of supervision avail-

able for adapting the models to the task. While extractive models remained reasonably stable, larger generative models were more sensitive to the limited supervision and showed less consistent performance. This limitation influenced our decision to adopt a multi-stage approach with a separate verification step, rather than relying solely on direct fine-tuning. A larger training set would likely allow more effective fine-tuning of generative models and lead to more stable and reliable results overall.

8. Conclusion and Future Work

Across the project, we demonstrate that a modular extractive QA pipeline is an effective approach for financial causality extraction, achieving a 7th-place ranking in the English track with only marginal differences from higher-ranked teams. Starting from an extractive baseline (RoBERTa) and progressively incorporating instruction-following modelling (Qwen) and refinement stages (GPT-3.5, GPT-4.0 mini, and Gemini-2.5-Flash), we achieved consistent improvements while keeping outputs strictly extractive. In particular, the candidate-verifier design helped reduce span boundary drift by correcting answers that were too short (truncated) or too long, and the staged structure made it clear which components were responsible for the improvements in Exact Match and semantic alignment.

For future work, this system will serve as a foundation for a Masters-level group project and will be expanded in both scope and capability. We aim to evaluate whether the approach generalises beyond financial reports to other domains such as medical or construction text, and to train and fine-tune stronger models using improved computational resources to further boost performance.

References

Dominique Mariko, Kim Trottier, and Mahmoud El-Haj. 2022. The financial causality extraction shared task (fincausal 2022). In *Proceedings of the 4th Financial Narrative Processing Workshop @ LREC 2022*, pages 105–107, Marseille, France. European Language Resources Association.

Antonio Moreno-Sandoval, Blanca Carbajo Coronado, Jordi Porta Zamorano, Yanco Amor Torterolo Orta, and Doaa Samy. 2025. The financial document causality detection shared task (FinCausal 2025). In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 214–221, Abu Dhabi, UAE. Association for Computational Linguistics.

Antonio Moreno-Sandoval, Jordi Porta, Yanco Torterolo, Alexia Stanescu, Melina Chatzi, and Sofía Roseti. 2026a. The Financial Document Causality Detection Shared Task (FinCausal 2026). In *Proceedings of the 7th Financial Narrative Processing Workshop (FNP 2026) at LREC 2026*, Palma de Mallorca, Spain. ELRA.

Antonio Moreno-Sandoval, 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)*, Sorrento, Italy. IEEE.

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

John Schulman and Thinking Machines Lab. 2025. Lora without regret. *Thinking Machines Lab: Connectionism*.

Avinash Trivedi, Gauri Toshniwal, Sivanesan Sangeetha, and S. R. Balasundaram. 2025. Sarang at FinCausal 2025: Contextual QA for financial causality detection combining extractive and generative models. In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 242–247, Abu Dhabi, UAE. Association for Computational Linguistics.

A. Appendix

A.1. English Refinement Prompt

The following prompt was used in the second-stage refinement step of the Qwen + Gemini pipeline.

You are correcting a short answer. **Rules:**

- Output **only** the final answer.
- Do **not** add explanations.
- Return a **verbatim span** from the context.
- If the answer is already correct, return it unchanged.
- Prefer the full sentence containing the answer if needed.
- If multiple spans are possible, choose the shortest correct one.
- Goal: maximise Exact Match.

Examples:

Context: Additionally, during the year the Group commenced work, under its EPCC contract with CGE, on four biogas-based power generation plants. As previously stated, due to financial constraints, progress was slower than initially expected and work has been temporarily suspended, awaiting the Company finalising an arrangement with CGE.

Question: What did financial constraints bring about?

Answer: progress was slower than initially expected and work has been temporarily suspended, awaiting the Company finalising an arrangement with CGE

Context: The Directors believe this growth is driven by consumer preferences moving away from chain and branded pubs and towards pubs with an individual identity and an ambience which reflects the local market.

Question: What explains the growth, according to the Directors?

Answer: consumer preferences moving away from chain and branded pubs and towards pubs with an individual identity and an ambience which reflects the local market

Context: The nature of the Group's operations creates an ongoing demand for fuel and therefore the Group is exposed to movements in market fuel prices. The Group enters into commodity derivative instruments to hedge such exposure where it makes commercial and economic sense to do so.

Question: What explains why the Group is exposed to movements in market fuel prices?

Answer: The nature of the Group's operations creates an ongoing demand for fuel

A.2. Spanish Refinement Prompt

For Spanish inputs, a language-adapted version of the refinement prompt was used to ensure consistent instruction following. The structure and constraints remained identical to the English prompt, but the instructions and examples were provided in Spanish.

Estás corrigiendo una respuesta corta.

Reglas:

- Devuelve **solo** la respuesta final.
- No añadas explicaciones.
- La respuesta debe ser un fragmento **verbatim** del contexto.
- Si ya es correcta, devuélvela igual.
- Si es necesario, prefiere la oración completa que contiene la respuesta.
- Si hay múltiples opciones, elige la más corta correcta.
- Objetivo: maximizar Exact Match.

Ejemplos:

Contexto: Debido a restricciones financieras, el progreso fue más lento de lo esperado y el trabajo ha sido suspendido temporalmente hasta que la empresa finalice un acuerdo.

Pregunta: ¿Qué provocaron las restricciones financieras?

Respuesta: el progreso fue más lento de lo esperado y el trabajo ha sido suspendido temporalmente hasta que la empresa finalice un acuerdo

A.3. Model Versions

Table 4: Model versions used in the experiments.

Model	Version / Checkpoint	Date Used
Qwen	Qwen2.5-1.5B-Instruct	March 2026
RoBERTa QA	question-answering-roberta-base-s-v2	March 2026
GPT-3.5	GPT-3.5	March 2026
GPT-4	GPT-4.0	March 2026
Gemini	Gemini 2.5 Flash	March 2026