The Financial Document Causality Detection Shared Task (FinCausal 2026)

Antonio Moreno-Sandoval, Jordi Porta, Yanco Torterolo, Alexia Stanescu, Melina Chatzi, Sofía Roseti
Universidad Autónoma de Madrid
Laboratorio de Lingüística Informática
{antonio.msandoval, jordi.porta}@uam.es
{yanco.torterolo, maria.stanescu, melina.chatzi, sofia.roseti}@estudiante.uam.es

Abstract

The Financial Document Causality Detection shared task (FinCausal) is a competition organized within the Financial Narrative Processing (FNP) workshop series. It aims to identify the causal relationship between a question and its answer in a given financial context. The dataset is built from real annual reports drafted by Spanish IBEX 35 companies and several UK companies. The task includes two subtasks, one in English and one in Spanish. It is formulated as an Extractive Question-Answering (EQA) task in which, given a context (C) and a question (Q), participants must extract the verbatim answer span (A). The 2026 edition introduces several changes to increase task difficulty, including the reformulation of 10% of the questions to require deeper reasoning and a stronger emphasis on multi-step causal chains with three or more elements, achieved by removing overly simple cases and adding 500 new complex fragments per language. Another innovation is the adoption of an LLM-as-a-judge metric on a 1–5 scale, based on a rubric designed to align better with human preferences than Semantic Answer Similarity (SAS) and Exact Match (EM). This edition was hosted as part of the LREC conference in Palma de Mallorca, Spain.

Keywords: causal detection, EQA task, financial documents, LLM-as-a-judge

1. Introduction

The Financial Document Causality Detection Shared Task (FinCausal) is a long-running competition organized within the Financial Narrative Processing workshop series. The task focuses on text-internal causality in financial documents. Our objective is not to verify the factual truth of financial statements, but to evaluate how systems identify causes and effects as they are expressed in text. In its first editions, the shared task included only an English subtask (Mariko et al., 2021, 2022).

In the 2023 edition, causality detection was formulated as a span-extraction task (Moreno-Sandoval et al., 2023). Given a context and a span containing either a cause or an effect, participants had to extract the corresponding span that triggered it or was triggered by it. This edition also introduced the Spanish subtask. Evaluation relied on Exact Match (EM) at span level, and weighted F1, precision and recall at token level.

In 2025, the task shifted to an Extractive Question-Answering (EQA) task (Moreno-Sandoval et al., 2025). Given a context (C) and an abstractively formulated question (Q), participants had to extract the verbatim answer span (A) from the context. Because questions are abstractive while answers are extractive, this setup is more challenging for encoder-only systems and requires deeper contextual understanding. To better support generative models, Semantic Answer Similarity (SAS) (Risch et al., 2021) was introduced alongside EM to evaluate answers.

The 2026 edition builds on the 2025 framework and preserves the EQA formulation. This year, we prioritize *explanatory causality*, i.e., causes that lead to measurable effects, while reducing instances of *justificatory causality*, where text provides motives rather than direct triggers. In addition, EM and SAS were considered insufficient to capture all relevant response-quality nuances, and an LLM-as-a-judge framework was adopted. The main motivation is that many system outputs are semantically correct but lexically different from the reference, so a judge model can score semantic adequacy and causal grounding more faithfully than strict overlap-based metrics.

The 2026 design also increases task difficulty. We reviewed previous datasets and removed ambiguous or overly simple cases. We expanded the corpus with more than 500 new fragments per language, emphasizing multi-step causal chains with three or more elements. Moreover, 10% of the abstractive questions were rephrased to reduce lexical matching shortcuts and encourage deeper reasoning. Finally, training and test partitions were randomly re-split to distribute these changes evenly across the dataset.

2. Dataset

Building upon the 2025 dataset (Carbajo-Coronado et al., 2025), the current 2026 dataset (Moreno-Sandoval et al., 2026) transitions to a more complex EQA framework, distancing even more from the

Proceedings of the 7th Financial Narrative Processing Workshop @ LREC 2026, pages 114–124

Context	Question	Answer
<effect_2>Amadeus' non-air bookings declined by 1.5% in 2018 versus the previous year**</effect_2>** as a consequence of **<nested_cause_2><effect_1>**a decline in rail bookings**</effect_1>**, mostly driven by **<cause>**strikes impacting a key customer, which more than offset the double-digit increase in Amadeus' hotel bookings**</cause></nested_cause_2>**.	**<effect_2>**What factor caused Amadeus' 1.5% drop of non-air booking in 2018?**</effect_2>**	**<nested_cause_2>**a decline in rail bookings, mostly driven by strikes impacting a key customer, which more than offset the double-digit increase in Amadeus' hotel bookings**</nested_cause_2>**

Table 1: Sample for the English subtask marked with XML tags. Cause_1 corresponds to effect_1. They form nested_cause_2, corresponding to the effect_2. Effect_2 is used in the question (*Q*) to obtain the answer (*A*) from nested_cause_2.

2023 dataset (Moreno-Sandoval et al., 2023). The dataset is organized into two language-specific partitions, one for English and one for Spanish. Each example consists of a unique (*ID*), the context (*C*), the abstractive question (*Q*), and the gold-standard extractive answer (*A*). The distribution of the samples for the training and test partitions is detailed in Table 2.

The dataset for the Spanish subtask is sourced from the FinT-esp (Moreno Sandoval et al., 2020) corpus, which is composed of financial annual reports from Spanish IBEX 35 companies spanning 2014 to 2018. The official English translations of the 2018 reports were included and aligned, resulting in a bilingual ES-EN parallel corpus. For the English subtask, these English versions were combined with additional reports from the Lancaster UCREL research team corpus (El-Haj et al., 2019). All texts were manually annotated by linguists under expert supervision. The dataset is balanced across both subtasks to facilitate the development and evaluation of multilingual models. An example of the dataset is shown in Table 1. Additional information about the dataset and the competition is available on the official website.[1]

Subtask	Training Set	Test Set	Total
English	2,000	500	2,500
Spanish	2,000	503	2,503

Table 2: Distribution of samples for the FinCausal 2026 Shared Task.

3. Participants

A total of 20 teams registered for the task, 9 of which uploaded official submissions. All 9 participating teams submitted a corresponding system description paper for FinCausal. Among these teams, 7 of them competed in both the English and Spanish subtasks, while 2 of them focused exclusively on the English subtask. An additional team, while not submitting official test runs, provided a technical description of their proposed system, which is detailed in Subsection 5.4.

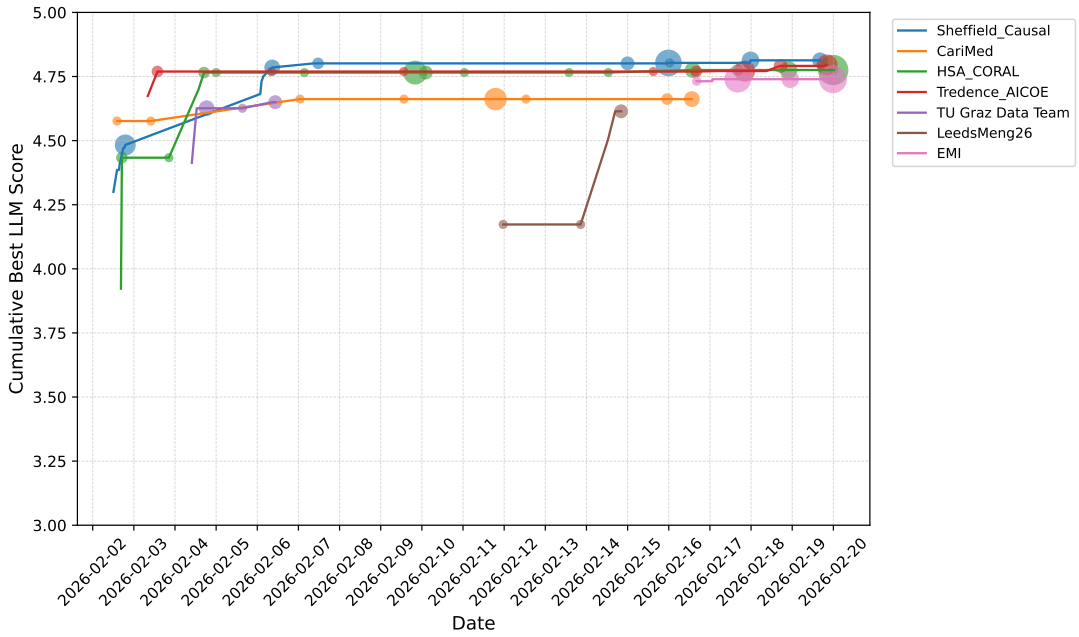
4. Competition Dynamics

In this competition, teams can make submissions that are scored using an internal metric not available to participants. Our goal in this section is to analyze the competition progress through the submission timeline and draw conclusions about competitive dynamics (e.g., participation pace, incremental improvements, stagnation periods, and convergence/divergence across teams). All data used for this analysis can be found in the competition leaderboard[2], which records each submission with the team name, score, and timestamp.

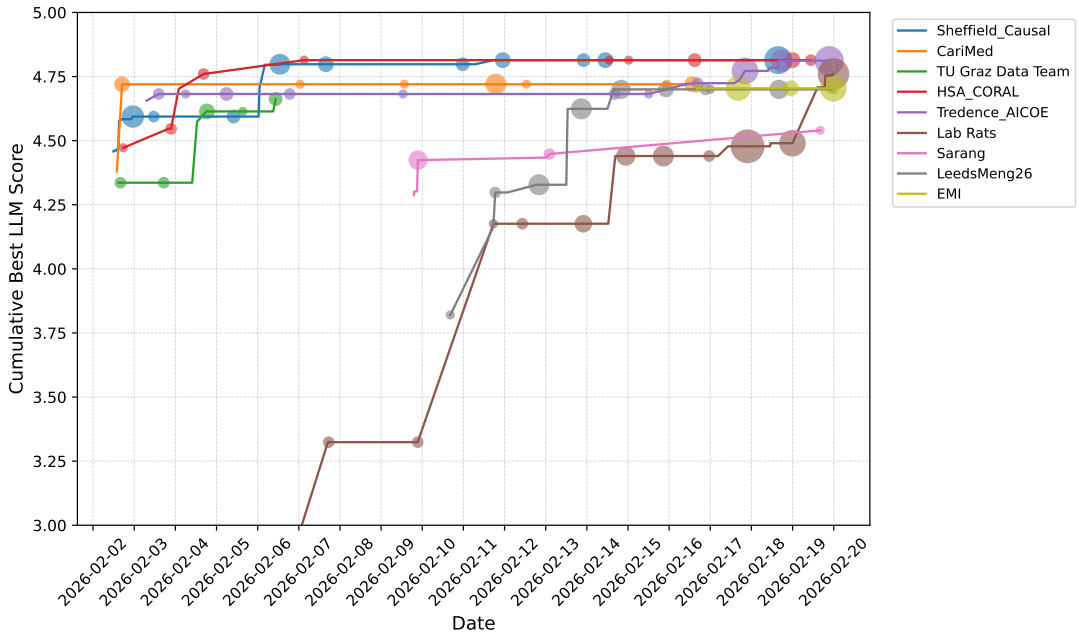
We build a time-ordered *timeline* that allows us to observe: the number of submissions per unit of time (activity), the evolution of each team's *best-so-far* (improvement curves), the evolution of the *leader frontier* (best overall score at each time), and the score distribution over time (convergence). Figure 1 shows two time-series line charts (Spanish on the left, English on the right) built from leaderboard records. Each colored line corresponds to one team and traces that team's cumulative best score over time, so upward steps indicate genuine improvements and flat segments indicate no improvement. Colors are used only to distinguish teams consistently across dates. The circles mark individual submissions at their submission timestamp and score. Circle size is proportional to submis-

[1] https://www.lllf.uam.es/wordpress/fincausal-26/

[2] https://leptis.lllf.uam.es/fincausal2026/



(a) Spanish Subtask



(b) English Subtask

Figure 1: Competition leaderboard dynamics for the Spanish (a) and English (b) subtasks. Each colored line represents one team and traces its cumulative best LLM score over time (step increases indicate improvements; flat segments indicate no improvement). Circles mark individual submissions at their timestamps and scores, and circle size is proportional to local submission density, so larger circles denote periods with more concentrated submission activity.

sion density at that point (i.e., larger circles indicate more submissions concentrated at that score/time region), helping visualize activity bursts and periods of low participation.

Temporal analysis reveals a consistent pattern across both subtasks: teams achieve rapid gains in the first days, followed by longer plateau phases with smaller marginal improvements. This behavior helps characterize iteration strategies (many submissions with small gains versus fewer submissions with larger jumps) and highlights *sprint* periods near the deadline. In the Spanish track, top systems

Team Name	Subtask	Core Methodology	Primary Model(s)
CariMed	Both	Multi-agent pipeline (RaR + RSA)	N/A
EMI	Both	Structural SFT (prompt repetition)	GPT-4.1-nano
HSA CORAL	Both	Prompt optimization, dynamic few-shot selection, and SFT	GPT-4.1-mini
Lab Rats	English	intra-context TF-IDF retrieval	Qwen3-4B-Instruct
LeedsMeng26	Both	Two-stage pipeline (candidate + verifier)	Qwen2.5 + Gemini
Sarang	English	Prompt optimization (DSPy/MIPROv2)	Gemma3-12B
Sheffield Causal	Both	Hybrid RAG (Dense) + bilingual SFT	GPT-4.1-mini
Tredence AICOE	Both	Voting-based ensamble (EN) | augmented SFT (ES)	GPT-5.1 / 4.1-mini
TU Graz Data Team	Both	SpanDiffusion (flow matching)	DeBERTa-v3
YT*	English	Difficulty-aware SFT + span anchoring	Llama-3.1-8B

Table 3: Comparison of participant methodologies. Teams are sorted alphabetically. *The YT team did not submit official results for the test evaluation phase.

converge early within a narrow score band (around 4.75–4.81), with limited late rank changes, suggesting faster stabilization. By contrast, the English track shows greater volatility, including later jumps from mid-ranked teams and stronger rank pressure near the top. Overall, the larger circles near the end of both timelines indicate denser submission activity, consistent with intensified last-minute experimentation.

5. Results

Table 4 reports, for each subtask, the best submission obtained by every participating team under the LLM-based evaluation protocol. For each entry, the table includes the final rank, team name, and best LLM score; it also reports summary statistics of score distribution, including the average score, to support direct comparison across language tracks. The Spanish and English rankings are presented side by side, making visible both top-level ties and score gaps across teams.

Overall, the results show strong and closely matched performance across teams in both subtasks. The English leaderboard is particularly competitive at the top, including a tie for first place. The Spanish leaderboard shows a similarly narrow gap among the leading systems. This distribution suggests that current systems are reaching a mature performance regime in this competition.

5.1. Spanish Subtask

The results for the Spanish subtask, as detailed in Table 4, reflect a high degree of technical proficiency. While the top three contenders remained consistent with the English track, the internal hierarchy shifted slightly. Sheffield Causal emerged as the winner with a score of 4.813, closely followed by Tredence AICOE and HSA CORAL. The overall performance was notably homogeneous; the gap between the leading submission and the bottom of the leaderboard did not exceed 0.2 points, with all participants scoring above the 4.6 threshold. This narrow variance suggests that the complexity of the Spanish financial corpus was effectively addressed by the participants' multilingual strategies.

5.2. English Subtask

The performance across the English subtask reveals an exceptionally competitive landscape. A remarkable tie for the first position was achieved by HSA CORAL and Sheffield Causal, both reaching a near-perfect score of 4.814 out of 5. The margin for the top positions was minimal; Tredence AICOE secured the third spot, trailing by a mere 0.002 points. This tight clustering of results extends throughout the ranking, where even the lowest-scoring submissions maintained a high standard of causal extraction, demonstrating the robustness of current LLM-based architectures in processing English financial narratives.

5.3. System Descriptions

To contextualize the ranking results, we briefly summarize the system-design choices reported by participants. Table 3 provides an overview of the systems employed by each team.

5.3.1. English and Spanish Systems

Team **Carimed** (Jay et al., 2026) introduced the VERSA system to tackle the task. It is a five-stage multi-agent pipeline in a zero-shot scenario that relies on API-connected models. It utilizes

Rank	Team Name	LLM Score					Average
		1	2	3	4	5	
1	Sheffield Causal	1	2	16	52	432	4.813
2	Tredence AICOE	0	2	28	41	432	4.795
3	HSA CORAL	4	1	21	52	425	4.775
4	EMI	2	6	28	49	418	4.739
5	CariMed	14	7	18	57	407	4.662
6	TU Graz Data Team	3	7	25	93	375	4.650
7	LeedsMeng26	8	7	24	83	379	4.614
	Total	**32**	**32**	**160**	**427**	**2868**	

(a) Spanish Subtask

Rank	Team Name	LLM Score					Average
		1	2	3	4	5	
1	HSA CORAL	3	2	21	33	441	4.814
1	Sheffield Causal	4	2	18	35	441	4.814
3	Tredence AICOE	1	3	24	33	439	4.812
4	Lab Rats	3	9	24	33	431	4.760
5	CariMed	5	6	28	46	415	4.720
6	EMI	6	6	40	26	422	4.704
7	LeedsMeng26	3	11	26	53	407	4.700
8	TU Graz Data Team	8	9	58	55	370	4.662
9	Sarang	1	3	24	33	439	4.540
	Total	**34**	**51**	**263**	**347**	**3805**	

(b) English Subtask

Table 4: LLM-based ranking of each team's best submission for the Spanish (a) and English (b) subtasks. Each row reports the team rank, team name, score distribution and the LLM score (avg.) of their best submitted system.

two main techniques: (1) Rephrase-and-Respond (RaR), where the model optimizes the prompt; and (2) Recursive Self-aggregation (RSA), which provides candidate answers and recursively adds subsets to obtain consensus. Stage 1 functions as a causal analyst, stage 2 applies the rephaser (RaR), stage 3 extracts the relevant information, stage 4 aggregates candidates (RSA), and stage 5 validates the candidates before providing the final answer. This approach obtained a score of 4.6620 in Spanish (5th) and 4.720 in English (5th).

Team **HSA CORAL** (Gautam et al., 2026) compares three approaches to address the EQA: (1) encoder-only for token classification using BERT-based models, (2) encoder-decoder for sequence-to-sequence generation using BART-like models, and (3) decoder-only models for generation, enforcing extraction with prompt optimization, few-shot selection, and fine-tuning. They retrieve the most similar cosine-similarity-based C and Q pairs from the training set to those pairs from the test set to include them as shots. They found that fine-tuned generative models outperform encoder-only and encoder-decoder models. Also, 20-shot prompting enables compact models to outperform bigger models in a zero-shot scenario. Their best system consists of a bilingually fine-tuned GPT-4.1 mini with 20 shots similar to the test sample, achieving a score of 4.7753 in Spanish (3rd) and 4.814 in English (1st).

Team **Tredence AICOE** (Chopra et al., 2026) presented a multilingual financial causality extraction system for English and Spanish based on few-shot prompting, supervised fine-tuning, data augmentation, and ensemble arbitration. They experimented with models such as GPT-5, Gemini 3.0 Pro, Qwen-14B, Gemma-12B, GPT-4.1 mini, GPT-OSS 20B, and GPT-5.1, and explored techniques including joint EN+ES training, LoRA-style efficient fine-tuning, bidirectional translation, synthetic data generation, distilled chain-of-thought reasoning, confidence-based selection, semantic consensus, and voting-based ensembles. Overall, the best results come from multilingual fine-tuning plus selective ensembling, with GPT-5.1-arbitrated voting performing best in English and synthetically augmented GPT-4.1 mini performing best in Spanish. In Spanish, the best score was 4.795 (2nd), achieved by GPT-4.1 mini fine-tuned with synthetic data augmentation. In English, the best result was

obtained with a voting-based ensemble arbitrated by GPT-5.1, which achieved 4.812 (3rd), outperforming all standalone prompting and fine-tuned systems.

Team **EMI** (Attak, 2026) presented a system based on supervised fine-tuning of instruction-following language models, with particular emphasis on prompt repetition as a training strategy to reinforce the relationship between the question format and the expected extractive answer. The authors evaluate both open-weight (Qwen2.5-7B, Qwen2.5-14B) and proprietary models (GPT-4.1-Nano), and show that this simple intervention can improve extraction fidelity, especially for open models, by reducing over-generation and helping the model stay closer to the source span. Their best systems (GPT-4.1-nano) obtained 4.7396 in Spanish (4th) and 4.704 in English (6th).

Team **LeedsMeng26** (Shahrouri et al., 2026) presented a two-stage extractive question answering system for FinCausal 2026. They cast financial causality detection as a QA problem over English and Spanish financial texts, returning a verbatim span from the context rather than generate a free-form answer. Their system works in two steps. First, a model generates an initial candidate span under strict prompting designed to force extractive behavior. Then a second model acts as a verifier and boundary refiner, checking whether the candidate is correct and adjusting its span boundaries when necessary, while still being constrained to output only a contiguous substring from the source text. The paper says this second stage was introduced to fix typical span errors such as truncation, overrun, or occasional paraphrasing. Their final submitted configuration was Qwen-2.5-1.5B-Instruct + Gemini 2.5-flash achieveving scores of 4.6143 in Spanish (7th) and 4.7000 in English (7th).

Team **TU Graz Data Team** (Niess and Kern, 2026) introduced SpanDiffusion to approach financial causal question answering as an extractive span prediction problem, but replaced standard start–end classification with a continuous denoising approach in a system called SpanDiffusion. The question and context are encoded with DeBERTa-v3-large plus LoRA, and the target answer is represented as two Gaussian masks marking the start and end of the span. A transformer denoiser trained with flow matching reconstructs these masks from noise, and the final span is obtained by taking the peak of each mask. This design aims to model boundary uncertainty while preserving the extractive nature of the task. In the results, the proposed method proves competitive but not superior to a simpler baseline, namely DeBERTa-v3-large + LoRA + a linear span head. The best SpanDiffusion model reaches 83.0 Exact Match, with notable gains from LoRA and from using flow matching instead of DDPM. However, the standard span-classification baseline achieves 85.8 Exact Match, outperforming the diffusion model with much lower complexity. Overall, the paper's contribution is therefore mainly methodological, offering an original alternative to conventional extractive QA rather than a stronger empirical system. Their best systems achieved a score of 4.6501 in Spanish (6th) and 4.662 in English (8th).

Team **Sheffield Causal** (Alqarni et al., 2026) addressed financial causal question answering in English and Spanish as a generative extractive task based on GPT-4.1-mini. The authors combine prompt engineering, retrieval-augmented generation (RAG), and supervised fine-tuning, while enforcing strict verbatim extraction from the input context. Their pipeline has three stages: indexing the training set, retrieving top-k similar examples, and constructing few-shot prompts for either the base or the fine-tuned model. They compare several retrieval strategies, including random example selection, BM25, dense retrieval with text-embedding-3-large and LlamaIndex, a pattern-aware retrieval method that groups questions into CAUSE, EFFECT, or OTHER templates, and a hybrid BM25+dense approach using reciprocal rank fusion. In addition, they evaluate simple, expert, and multilingual prompts, and fine-tune GPT-4.1-mini on bilingual training data formatted as question-context-answer triples. Their system ranked first in both subtasks, reaching a score of 4.813 for Spanish (1st) and 4.814 for English (1st).

5.3.2. English-only Systems

Team **Lab Rats** (Sarda et al., 2026) participation in the competition consists of two main aspects: QLoRA SFT of Qwen3-4B-Instruct on the English dataset, which was adapted to the ChatML instruction format required by the model; and an intra-context TF-IDF retrieval to enrich context for enforcing verbatim span extraction at inference time. The inference pipeline involves four stages: (1) loading the model in 4-bit, (2) intra-context retrieval, comparing each sentence from the given C against the Q, thus filtering the most relevant fragment, (3) constrained prompting, where the model is instructed to extract the span verbatim, and the previously retrieved fragment is also provided, and (4) greedy decoding and deterministic settings aimed at improving extraction accuracy. They achieved a score of 4.760 in English (4th).

Team **Sarang** (Trivedi and Chindukuri, 2026) presented a system based on few-shot prompting and automated prompt optimization for Gemma3-12B. Using DSPy and the MIPROv2 teleprompter, the system optimizes instructions and demonstration examples drawn from the training set, and performs inference locally through Ollama. The paper

compares this setup with RoBERTa and DeBERTa baselines finetuned with other QA datasets and with other few-shot LLM configurations, finding that Gemma3-12B with medium prompt optimization performs best. Their top system obtained an LLM Score of 4.540 in the English shared sub-task (9th), highlighting the effectiveness of prompt optimization for financial causal QA.

5.4. Additional English-only System

Team **YT** utilized a LoRA-finetuned Llama-3.1-8B model using a difficulty-aware training strategy. Their methodology involves a custom labeling scheme that categorizes instances into simple, chain, and abstractive types based on structural and semantic complexity. To ensure strict verbatim extraction, the system employs a three-level span-anchoring mechanism—combining case-insensitive matching, semantic sentence retrieval, and re-prompting—complemented by targeted post-processing rules to truncate redundant clauses. Although the team did not submit official runs for the shared task, their internal evaluations on a partitioned subset of the 2026 dataset demonstrated the effectiveness of combining difficulty-stratified training with post-processing.

6. Evaluation

In FinCausal 2023 (Moreno-Sandoval et al., 2023), the task focused on identifying cause–effect relations linked to events or quantified facts in financial texts. Because it was formulated as an extractive task, system performance was assessed with EM to quantify the proportion of predictions that exactly match the reference span, and token-level precision, token-level recall, and weighted F1 to quantify partial overlap quality for extracted cause and effect spans.

In FinCausal 2025 (Moreno-Sandoval et al., 2025), the extraction task was reformulated as a question-answering task in which questions about causes or effects are posed, and system responses are evaluated with EM and semantic answer similarity metrics. This change accommodates the growing use of generative prompting-based models, many of them based on GPT architectures. For these models, a strict lexical metric such as EM alone is often insufficient, because generative systems can produce answers that are semantically correct but phrased differently from the references. SAS measures semantic similarity between texts rather than exact lexical overlap, making it well suited for abstractive generation tasks. The metric represents texts as vector embeddings using pre-trained models such as BERT (Devlin et al., 2019) or Sentence Transformers (Reimers and Gurevych, 2019), and computes cosine similarity between them. This allows the evaluation to capture cases where two answers express the same meaning despite differences in wording or structure.

In the FinCausal 2026 edition, system submissions are evaluated through an LLM-as-a-judge framework. For each system prediction, the judge model receives a fixed evaluation rubric and scores the response according to a uniform set of criteria. Specifically, each answer is rated on a five-point Likert scale, whose levels capture different degrees of semantic alignment between the predicted answer and the gold-standard reference. The full FinCausal 2026 rubric is provided in Appendix A. In our setup, the judge model is `openai/gpt-oss-20b` (OpenAI, 2025), a 20-billion-parameter open-weight language model from OpenAI's GPT-OSS family released in August 2025. We use the model with a medium reasoning configuration and explicitly instruct it in the prompt to generate concise score justifications, thereby improving the transparency and auditability of the evaluation procedure.

6.1. Error Analysis

The following initial analysis is based on a linguistic review of model outputs graded from 1 to 5. The predictions from the best system of each participant were observed, thus enabling the identification of common error patterns found in the lower and middle scoring ranges. Superficially, no errors were found on the scores of 5, but a more thorough analysis should be performed.

Score 1: Structural failure. At this level, models show major structural problems. While they often identify causal markers (like "due to" or "thanks to"), they extract the wrong information or focus on unrelated events nearby. Another common issue is empty referencing, where the model only extracts the connector (e.g., "Due to the above") without the actual explanation. Additionally, models at this stage sometimes reverse the cause and effect or skip essential steps in a causal chain. These models are also prone to neglecting or hallucinating quantitative financial data; a failure severely penalized by the LLM judge, which assigns the minimum score to responses lacking numerical precision.

Score 2: Incomplete and inaccurate. Errors here involve missing information or technical mistakes. Models often cut the answer too short, leaving out the specific details that provide the actual argument. They also tend to confuse similar financial terms or acronyms (like DVA instead of CVA). We also observed a metric bias: LLM judges sometimes give higher scores to fluent-sounding answers, even if the content is partially incorrect.

This behaviour is typical of metrics based in neural models (Kovacs et al., 2024; Freitag et al., 2024).

Score 3: Partial and selective. These responses overlap with the correct answer but are incomplete. Models often pick only the first factor in a list and ignore the others, providing a one-sided view of the event. They also frequently leave out important time-related details (like specific dates or years). While the core reason is usually captured, these models tend to ignore secondary details or consequences, resulting in an over-simplified version of the reference.

Score 4: Precise with minor noise. A score of 4 represents an accurate answer that includes unnecessary noise. The most common error is marker dragging, where the model includes extra words like connectors or introductory verbs (e.g., "resulting in" or "allowing") that were not part of the target answer. In other cases, the model adds extra context that is true but goes beyond the exact boundaries set by the experts.

The errors progress from total confusion at score 1 to minor noise at score 4. The main challenges for these models are navigating complex financial texts with many variables and staying within the strict boundaries of the required answer. Finally, the metric bias seen in scores 2 and 3 suggests that LLM judges often favor smooth, fluent writing over literal, word-for-word accuracy. Therefore, the differences between scores of 2 and 3 are often blurry, while the scores of 1 and 4 seem reliable.

7. Conclusions

This FinCausal 2026 edition had higher participation than previous editions in terms of effective system-description submissions. Consequently, it enabled us to gain a clearer overview of the current state of causal EQA in the financial domain. Some insights can be drawn from the results obtained by participants.

Bilingual fine-tuning performs consistently better than monolingual fine-tuning, as shown by direct comparisons in the HSA CORAL and Tredence AICOE submissions. Additionally, as highlighted by HSA CORAL, smaller fine-tuned models can outperform larger zero-shot models.

Another clear trend was the use of retrieval-based techniques and semantic matching. Four teams used similarity measures at different stages of their pipelines for different purposes. HSA CORAL applied vector similarity at inference time to retrieve few-shot examples similar to the test C and Q for in-context learning. Sheffield Causal followed a similar approach to HSA CORAL but conducted a broader evaluation of retrieval methods, including hybrid BM25 and dense retrieval. In contrast, Lab Rats performed intra-context filtering by comparing each sentence in the provided C with the Q to isolate the most relevant fragment before generation. Finally, YT integrated semantic similarity at two points: for the initial difficulty-based classification of the dataset and for a span-anchoring mechanism during post-processing.

On a different note, complex pipelines that rely on zero-shot settings still appear insufficient to consistently enforce verbatim extractive answers, as seen in CariMed's VERSA system and Sarang's prompt-optimization strategy. Compared with similarly complex pipelines that include fine-tuning, such as Tredence AICOE's, these results suggest that fine-tuning remains essential for stronger performance.

Regarding the main model choice, GPT-4.1 mini was used by several teams and proved to be a strong option despite its size. Sheffield Causal used it to obtain first place in both subtasks, tying with HSA CORAL in English. HSA CORAL also used GPT-4.1 mini in its best-performing system, and Tredence AICOE used it for their best Spanish run. A smaller variant, GPT-4.1 nano, was used by EMI. These results suggest an advantage for proprietary models in this edition. Given the compact size of the mini and nano variants, they also offer an attractive cost-performance ratio. On the open-source side, Qwen2.5 appeared in the systems of LeedsMeng26, while Qwen3 appeared in Lab Rats'. Gemma 3 was used by Sarang, while TU Graz Data Team used DeBERTa-v3 together with a span diffusion approach; YT used Llama-3.1-8B.

To conclude, despite the difficulty of the task, all participants achieved strong scores. This highlights both the quality of the participants' work and the consistency and coherence of the dataset in both languages. In addition, the shift to an LLM-as-a-judge metric has proved useful in light of the error analysis. However, a deeper future analysis of its decisions and of its alignment with human judgment is still needed. For future FinCausal editions, some important changes should be considered to maintain participant interest; for instance, including an additional subtask formulated as a pure QA task with abstractive answers, while preserving the current EQA approach.

Acknowledgments

We would like to thank FinCausal 2026 participants for their outstanding contributions to this shared taks.

This work is framed under the Spanish National Project GRESEL (PID2023-151280OB-C21). It was also partially funded by grant PTA2023-023812-I (awarded to Yanco Amor Torterolo Orta) through MICIU/AEI/10.13039/501100011033 and the European Social Fund Plus (ESF+).

8. Bibliographical References

Aali Abdullah Alqarni, Mark Stevenson, and Arif Dwi Laksito. 2026. A Comparative Study of RAG Approaches and Fine-Tuning for Causal QA in Financial Text. In *Proceedings of the 7th Financial Narrative Processing Workshop (FNP 2026) at LREC-COLING 2026*, Palma de Mallorca, Spain. European Language Resources Association (ELRA). To appear.

Sanae Attak. 2026. Improving Verbatim Financial Causality Extraction with Supervised Fine-Tuning and Prompt Repetition. In *Proceedings of the 7th Financial Narrative Processing Workshop (FNP 2026) at LREC-COLING 2026*, Palma de Mallorca, Spain. European Language Resources Association (ELRA). To appear.

Ankush Chopra, Shubham Sharma, and Ashmani Kumar. 2026. Extracting Financial Causality: A Multilingual Approach with SLM Fine-Tuning and LLM-Arbitrated Ensembles. In *Proceedings of the 7th Financial Narrative Processing Workshop (FNP 2026) at LREC-COLING 2026*, Palma de Mallorca, Spain. European Language Resources Association (ELRA). To appear.

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.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Akash Kumar Gautam, Serhii Hamotskyi, and Christian Hänig. 2026. Causal Connections: Leveraging Multilingual Fine-Tuning for Financial QA@FinCausal 2026. In *Proceedings of the 7th Financial Narrative Processing Workshop (FNP 2026) at LREC-COLING 2026*, Palma de Mallorca, Spain. European Language Resources Association (ELRA). To appear.

Aldan Jay, Rafael Berlanda, Yoelvis Moreno, and Vincent Santamarta. 2026. VERSA: Verbatim Extraction via Rephrasing and Self-Aggregation for Financial Causality. In *Proceedings of the 7th Financial Narrative Processing Workshop (FNP 2026) at LREC-COLING 2026*, Palma de Mallorca, Spain. European Language Resources Association (ELRA). To appear.

Geza Kovacs, Daniel Deutsch, and Markus Freitag. 2024. Mitigating metric bias in minimum Bayes risk decoding. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1063–1094, Miami, Florida, USA. Association for Computational Linguistics.

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.

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.

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, Ana Gisbert, and Elena Montoro. 2020. FinT-esp: A corpus of financial reports in Spanish. In Miguel Fuster-Márquez, Carmen Gregori-Signes, and José Santaemilia Ruiz, editors, *Multiperspectives in analysis and corpus design*, pages 89–102. Comares, Granada.

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.

Georg Niess and Roman Kern. 2026. SpanDiffusion: Flow Matching over Continuous Span

Masks for Financial Causal Question Answering. In *Proceedings of the 7th Financial Narrative Processing Workshop (FNP 2026) at LREC-COLING 2026*, Palma de Mallorca, Spain. European Language Resources Association (ELRA). To appear.

OpenAI. 2025. gpt-oss-120b gpt-oss-20b model card.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

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.

Bavya Sarda, Pulkit Chatwal, and Sonal Dabral. 2026. QRAFT: QLoRA Retrieval-Augmented Fine-Tuning for Causal Span Extraction in Financial Documents. In *Proceedings of the 7th Financial Narrative Processing Workshop (FNP 2026) at LREC-COLING 2026*, Palma de Mallorca, Spain. European Language Resources Association (ELRA). To appear.

Zaid Shahrouri, Ayomide Ivienagbor, Idrees Asad, Rijul Shrestha, Yasemin Bal, and Zahaab Nadeem. 2026. LeedsMEng26: Qwen + Gemini for FinCausal 2026 Causality Detection in Financial Narrative Texts. In *Proceedings of the 7th Financial Narrative Processing Workshop (FNP 2026) at LREC-COLING 2026*, Palma de Mallorca, Spain. European Language Resources Association (ELRA). To appear.

Avinash Trivedi and Mallikarjuna Chindukuri. 2026. Financial Causal QA via Instruction and Prompt Tuning of Gemma3-12B. In *Proceedings of the 7th Financial Narrative Processing Workshop (FNP 2026) at LREC-COLING 2026*, Palma de Mallorca, Spain. European Language Resources Association (ELRA). To appear.

A. FinCausal 2026 Rubric

```
You are an expert evaluator. Your task is to rate how good an STUDENT_ANSWER is
for a given QUESTION, a CONTEXT and a REFERENCE_ANSWER according to
the rubric below.

RUBRIC (score from 1 to 5):
- 5: Excellent quality: The prediction is semantically identical or fully equivalent to the gold standard response. Minor formal variations (e.g., punctuation, casing) are tolerated. The content perfectly addresses the causal question, showing no signs of omission or irrelevant inclusion. The answer is deemed fully appropriate and reliable.
- 4: Good quality: The prediction matches the gold-standard answer in full but included small additional content from the context. These additions did not compromise the semantic correctness of the answer but extended it slightly. Therefore, the prediction could be seen to be complete, relevant and informative, although wordy.
- 3: Medium quality: The prediction contains the central idea or a correct causal link, but is either incomplete or diluted with unrelated information. It may capture the start of a causal phrase but miss important qualifiers or follow-up clauses. Alternatively, the prediction might include correct content but extend unnecessarily beyond the relevant span.
- 2: Low quality: The predictions demonstrates only a superficial connection to the question or to the correct answer. Typically, large portions of the expected content are omitted, and irrelevant elements may have been added. The response might contain a partial clue or topic-related phrase, but failed to provide a clear, informative, or accurate answer. This category captures both underinformative and noisy outputs.
- 1: Very poor quality: Predictions in this category failed entirely to answer the question. These responses were irrelevant, incorrect, or confusing, and often exhibited no meaningful overlap with the gold-standard answer. Even if some surface text matched the source, the essential causal content was absent.

First, briefly explain your reasoning.
Then assign a single integer score from 1 to 5.

Return your response as pure JSON with this exact schema:

{{
```

```
  "score": <integer 1-5>,
  "reasoning": "<short explanation>"
}}

CONTEXT:
{context}

QUESTION:
{question}

REFERENCE_ANSWER:
{reference}

STUDENT_ANSWER:
{answer}
```

B. Language Resource References

Blanca Carbajo-Coronado, Antonio Moreno-Sandoval, Yanco Amor Torterolo Orta, and Paula Gozalo. 2025. The Financial Document Causality Detection Shared Task (FinCausal 2025): Dataset.

Mahmoud El-Haj, Steven Young, and Paul Rayson. 2019. Annual reports key sections corpora 2003 to 2017.

Antonio Moreno-Sandoval, Blanca Carbajo-Coronado, and Jordi Porta. 2023. The financial document causality detection shared task (FinCausal 2023): Dataset.

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