QRAFT: QLoRA Retrieval-Augmented Fine-Tuning for Causal Span Extraction in Financial Documents

Bavya Sarda[1], **Pulkit Chatwal**[2], **Sonal Dabra**[3]
[1]Galgotias University, Greater Noida, India
[2]Rajiv Gandhi Institute of Petroleum Technology, India
[3]Sony Research, India
{bhavyasarda19, pulkitchatwal, sonaldabral26}@gmail.com

Abstract

Understanding *why* financial outcomes occur is as important as knowing *what* they are. Annual reports and regulatory filings are rich with causal reasoning, yet extracting that reasoning automatically remains a difficult problem — one that sits at the intersection of domain expertise, linguistic nuance, and machine comprehension. In this paper, we describe our participation in the English subtask of the Financial Document Causality Detection shared task, FinCausal 2026, where systems are asked to identify verbatim causal spans from financial paragraphs in response to abstractive causal questions.Our approach is grounded in the intuition that a small, well-adapted model with the right inductive biases can outperform a larger but unfocused one. We fine-tune Qwen3-4B-Instruct-2507 on 2,000 domain-annotated instances using QLoRA, a parameter-efficient technique that enables meaningful adaptation under modest computational resources. Before training, we reformat all instances into the Qwen ChatML instruction template to align the model's generation behaviour with the verbatim extraction requirement of the task. At inference time, we further guide the model by retrieving the most causally relevant sentence from the context using TF-IDF cosine similarity, providing an explicit local signal before generation. Outputs are produced via greedy decoding to ensure deterministic, source-grounded predictions.Under the official LLM-as-a-judge evaluation framework — which scores responses on a 1–5 adequacy scale based on semantic correctness rather than lexical overlap — our system achieves a score of **4.76 out of 5**, placing **4th out of nine teams** on the English leaderboard. Our results suggest that combining instruction-tuned fine-tuning with lightweight retrieval is a practical and effective strategy for causal reasoning in specialised financial text.

Keywords: causal question answering, financial NLP, QLoRA, parameter-efficient fine-tuning, TF-IDF retrieval, span extraction, LLM-as-a-judge

1. Introduction

Financial documents such as earnings reports, annual filings, and regulatory disclosures do more than report numbers — they explain *why* those numbers changed. Understanding the causal reasoning embedded in these texts is fundamental to building systems that support financial analysis, risk assessment, and automated report generation. However, causality in financial language is rarely straightforward: causal relationships often span multiple sentences, involve compound contributing factors, and are expressed without explicit connectives such as *because* or *therefore* (Cheng et al., 2024; Girju, 2003).

The Financial Document Causality Detection shared task (FinCausal) (Mariko et al., 2020, 2022; Moreno-Sandoval et al., 2023, 2025) directly addresses this challenge by evaluating systems on their ability to identify cause-and-effect relationships in financial text across English and Spanish. The 2026 edition introduces three significant advances over prior years: a revised dataset with richer and more complex causal annotations, the inclusion of multi-hop causal chains involving three or more events, and a new evaluation framework in which an LLM judge scores system responses on a 1–5 adequacy scale (Zheng et al., 2023), replacing the earlier exact-match and similarity-based metrics.

We participate in the English subtask and frame it as a **Causal Question Answering (CQA)** problem, where a system must extract the verbatim causal span from a financial paragraph in response to an abstractive question. Our approach consists of three components: (1) reformatting the training data into the Qwen ChatML instruction format, (2) fine-tuning Qwen3-4B-Instruct-2507 using QLoRA for parameter-efficient domain adaptation, and (3) an inference pipeline that combines intra-context TF-IDF retrieval with constrained greedy decoding to enforce verbatim extraction and reduce hallucination.

Our system achieves a score of **4.76 out of 5** on the official test set, demonstrating that a carefully fine-tuned compact language model, paired with a lightweight retrieval signal, can effectively identify causal relationships in complex financial text.

2. Related Work

2.1. Causality Detection in NLP

Early approaches to causality detection relied on lexical cues such as *because*, *therefore*, and *as a result* (Girju, 2003; Blanco et al., 2008), but

Proceedings of the 7th Financial Narrative Processing Workshop @ LREC 2026, pages 175–180
May 16, 2026. ©2026 European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

struggled with implicit causal expressions. Subsequent work introduced machine learning methods — SVMs, CNNs, and RNNs — that learned patterns directly from data (Do et al., 2011). More recently, transformer-based models such as BERT have become the dominant approach due to their contextual understanding (Cheng et al., 2024). Recent studies have also explored the use of prompt engineering with large language models to enhance causal relationship detection, particularly in domain-specific settings such as finance (Chatwal et al., 2025). Our work extends this line by fine-tuning the Qwen model on financial causal data.

2.2. Causal Question Answering

Extractive QA benchmarks such as SQuAD require answer spans to be identified directly within a context, but do not emphasise causal reasoning. Tasks like COPA (Gordon et al., 2012) target commonsense causality, while FinCausal (Mariko et al., 2020, 2022; Moreno-Sandoval et al., 2023, 2025) focuses specifically on financial text. FinCausal 2026 raises the difficulty further by combining abstractive questions with extractive answers and introducing multi-hop causal chains, requiring models to reason rather than match.

2.3. Financial NLP

Financial text presents unique challenges for general NLP tools. Loughran and McDonald (2011) demonstrated that standard sentiment lexicons perform poorly on financial language, motivating domain-specific models such as FinBERT (Araci, 2019). Causality extraction in financial reports has gained traction through the FinCausal shared task series, which our work directly builds upon.

2.4. Retrieval-Augmented Generation and Parameter-Efficient Fine-Tuning

Retrieval-Augmented Generation (RAG) (Lewis et al., 2020)(Didwania et al., 2024) has shown that grounding generative models with retrieved context improves factual accuracy. We adopt a lightweight variant of this idea, using TF-IDF cosine similarity to surface the most causally relevant sentences before generation. For efficient adaptation, we employ QLoRA (Dettmers et al., 2023), which enables fine-tuning of large language models under 4-bit quantization with minimal performance degradation. Finally, FinCausal 2026 adopts the LLM-as-a-judge evaluation framework (Zheng et al., 2023), which scores responses on semantic adequacy rather than lexical overlap — better reflecting the quality of causal reasoning.

3. Problem Statement

Financial narratives describe measurable changes in economic indicators and corporate performance, yet numerical values alone rarely explain *why* such changes occur. The Financial Document Causality Detection task, FinCausal 2026 (mor; Uniyal et al., 2021), addresses this gap by targeting **text-internal causal relationships** within financial documents. We formalize this as a **Causal Question Answering (CQA)** problem, where each instance is structured as a triplet (C, Q, A): a financial paragraph $C = \{w_1, w_2, \ldots, w_n\}$ serving as context, an abstractive causal question Q, and an extractive answer span $A \subseteq C$ drawn verbatim from the text.

A causal relation is defined as an ordered pair (e_c, e_e), where e_c is the causal event and e_e is the resulting event, such that $e_c \rightarrow e_e$. The task focuses strictly on how causality is *encoded within the document*, not on the real-world validity of the stated relationships.

3.1. Learning Objective

Given (C, Q), the goal is to learn a function $f_\theta : (C, Q) \rightarrow \hat{A}$, where $\hat{A} = C[i : j]$ is a contiguous extractive span. Model parameters are optimized by minimizing the negative log-likelihood:

$$\mathcal{L}(\theta) = -\sum_{k=1}^{N} \log P_\theta(A_k \mid C_k, Q_k) \tag{1}$$

Despite the extractive constraint, the task is posed in a **generative QA format** to handle complex multi-hop causal chains of the form $e_1 \rightarrow e_2 \rightarrow \cdots \rightarrow e_k$, which are a key feature of the 2026 edition.

3.2. Evaluation

FinCausal 2026 replaces the earlier SAS + Exact Match scheme with an **LLM-as-a-judge** framework, where a judge model scores each response on a 1–5 adequacy scale. The overall system score is:

$$\text{Score} = \frac{1}{N} \sum_{k=1}^{N} J(A_k,\ \hat{A}_k) \tag{2}$$

This prioritizes **semantic adequacy and reasoning correctness** over strict lexical overlap.

4. Dataset

The FinCausal 2026 English dataset (Moreno-Sandoval et al., 2026) is sourced from UK financial annual reports (2017) compiled by UCREL at Lancaster University, supplemented with excerpts from the 2018 FinT-esp corpus. Relative to prior editions,

the dataset has been substantially revised: ambiguous and trivial instances were removed, and over 500 new fragments featuring complex causal structures — including chains of three or more events — were added. Approximately 10% of questions were rephrased to demand deeper reasoning beyond surface-level lexical matching. Training and test splits were constructed via random partitioning, ensuring uniform distribution of complex examples across both sets. We participated in the **English subtask**; Table 1 reports the split statistics.

Split	Language	Instances
Train	English	2,000
Test	English	500
Total		**2,500**

Table 1: Dataset statistics for the English subtask of FinCausal 2026.

4.1. Data Format

Each instance is a triplet (C, Q, A): a financial paragraph C as context, an abstractive causal question Q probing either the cause or the effect, and an extractive answer span A copied verbatim from C. The dataset is provided in CSV format with four fields: *ID*, *context*, *question*, and *answer* — where the answer field is withheld in the test set.

5. Methodology

Our approach to the FinCausal 2026 English subtask consists of two main stages: supervised fine-tuning of a compact instruction-tuned language model, and a retrieval-augmented inference pipeline designed to enforce verbatim causal span extraction. Algorithm 1 provides a high-level overview of the full system.

5.1. Base Model

We select **Qwen3-4B-Instruct-2507** (Team, 2025) as our base model. This model offers a strong balance between parameter efficiency and instruction-following capability, making it well-suited for a constrained extractive QA task on domain-specific financial text. Its compact size also allows fine-tuning and inference within limited computational budgets using quantization.

5.2. Input Formatting

Before fine-tuning, all training instances are reformatted into the **Qwen ChatML** template, which structures each example as a multi-turn conversation with explicit role markers. Each instance is formatted as follows:

Algorithm 1 FinCausal 2026 System Pipeline

Require: Context C, Question Q, Fine-tuned model f_θ
Ensure: Predicted causal span $\hat{A}$
// — Training Stage —
Reformat all (C, Q, A) instances into Qwen ChatML format
Load base Qwen3-4B-Instruct-2507 in 4-bit quantized mode
Attach QLoRA adapters to attention and feed-forward projections
Optimize θ by minimizing $\mathcal{L}(\theta) = -\sum_{k=1}^{N} \log P_\theta(A_k \mid C_k, Q_k)$
// — Inference Stage —
Load fine-tuned f_θ in 4-bit quantized mode
Tokenize C into sentences $\{s_1, s_2, \ldots, s_m\}$
Compute TF-IDF vectors for each s_i and Q
$s^* \leftarrow \arg\max_{s_i}$ cos(TF-IDF(s_i), TF-IDF(Q))
Construct prompt using full C, retrieved s^*, and Q
Generate $\hat{A} \leftarrow f_\theta(C, s^*, Q)$ using greedy decoding
return $\hat{A}$

Prompt Template

<|im_start|>system
You are a financial question answering assistant. Given a financial text, extract the answer to the causal question directly and **verbatim** from the context. Do not generate any answer that is not present in the text.

<|im_start|>user
Context: $[C_i]$
Question: $[Q_i]$

<|im_start|>assistant
$[A_i]$
<|im_end|>

This formatting aligns the training distribution with the model's pre-trained instruction-following behaviour, ensuring that the model learns to respond within the expected conversational structure rather than treating the task as raw text completion.

5.3. QLoRA Fine-Tuning

We fine-tune the model using **QLoRA** (Quantized Low-Rank Adaptation) (Dettmers et al., 2023), which combines 4-bit quantization of the base model weights with low-rank adapter layers inserted into the transformer architecture. This significantly reduces GPU memory requirements while preserv-

ing the model's ability to adapt to the target domain.

Low-rank adapters are injected into all major projection layers of the transformer, including the attention projections ($\mathbf{q}$, $\mathbf{k}$, $\mathbf{v}$, $\mathbf{o}$) and the feed-forward projections (`gate_proj`, `up_proj`, `down_proj`). Targeting the feed-forward layers in addition to the attention layers has been shown to substantially improve task-specific adaptation (Dettmers et al., 2023). Table 2 summarises the LoRA configuration used.

Hyperparameter	Value
LoRA rank (r)	32
LoRA alpha (α)	64
LoRA dropout	0.05
Bias	none
Task type	Causal LM
Target modules	q, k, v, o, gate, up, down proj
Quantization	4-bit (NF4)

Table 2: QLoRA fine-tuning configuration.

The rank $r = 32$ provides sufficient adapter capacity for capturing financial domain patterns, while $\alpha = 64$ (set to $2r$ following standard practice) controls the scaling of the adapter updates. A dropout of 0.05 is applied to the adapter layers as a lightweight regularisation measure.

5.4. Inference Pipeline

At inference time, we apply a four-step pipeline designed to minimise hallucination and enforce verbatim extraction from the source context.

Step 1 — Model Loading. The fine-tuned model is loaded in **4-bit quantized mode** using BitsAndBytes, reducing memory footprint while maintaining generation quality.

Step 2 — Intra-Context Retrieval. Rather than passing the full context blindly to the model, we apply an **intra-context retrieval** step. Each sentence in the context C is ranked against the question Q using **TF-IDF cosine similarity**. The top-ranked sentence s^* is selected as the most relevant causal evidence:

$$s^* = \arg\max_{s_i \in C} \cos(\text{TF-IDF}(s_i),\ \text{TF-IDF}(Q)) \quad (3)$$

This retrieved sentence is appended to the prompt alongside the full context, providing the model with an explicit signal about where the causal span is likely to reside.

Step 3 — Constrained Prompt. The model is prompted using a **verbatim extraction format** that explicitly instructs the model to copy the answer word-for-word from the context. Both the full context and the retrieved relevant sentence are included in the prompt, reducing the risk of paraphrased or hallucinated responses.

Step 4 — Greedy Decoding. Generation is performed using **greedy decoding** with temperature set to zero, ensuring fully deterministic outputs. This choice is deliberate: since the task requires precise verbatim spans, stochastic sampling strategies such as top-p or beam search with diversity penalties are counterproductive. Greedy decoding directly maximises the probability of the most likely token at each step, producing stable and reproducible predictions aligned with the source text.

6. Results

6.1. Main Result

We evaluate our system on the FinCausal 2026 English test set comprising 500 instances, using the official LLM-as-a-judge metric. Our system achieves a score of **4.76 out of 5**, ranking **4th** on the official English subtask leaderboard out of nine participating teams. Unlike Exact Match, the LLM judge tolerates minor boundary differences provided the causal meaning is preserved, making it a more faithful measure of reasoning quality in financial text. Table 3 reports the full leaderboard standings.

Rank	Team	LLM Score
1	HSA_CORAL	4.814
1	Sheffield_Causal	4.814
3	Tredence_AICOE	4.812
4	**Lab Rats (Ours)**	**4.760**
5	CariMed	4.720
6	EMI	4.704
7	LeedsMeng26	4.700
8	TU Graz Data Team	4.662
9	Sarang	4.540

Table 3: Official English subtask leaderboard for FinCausal 2026.

6.2. Component Analysis

Three design choices contribute to the strong performance of our system.

QLoRA Fine-Tuning. Training on 2,000 expert-annotated instances adapts the Qwen model to the

vocabulary, syntax, and causal patterns of financial reporting. Without this step, the base model lacks the domain grounding needed to distinguish causal spans from surrounding narrative text.

ChatML Instruction Formatting. Structuring each training instance as a ChatML instruction explicitly conditions the model to extract answers verbatim from the context rather than paraphrase or hallucinate plausible-sounding responses.

TF-IDF Intra-Context Retrieval. Ranking sentences by cosine similarity to the question before generation provides the model with an explicit relevance signal. This is particularly effective for instances where the question and causal span share limited lexical overlap, a common characteristic of abstractive causal questions in the 2026 dataset.

7. Limitations

Despite strong overall performance, our system faces two notable limitations. First, predicting long answer spans — particularly those covering multiple clauses — remains challenging, as small boundary errors can reduce semantic fidelity. Second, multi-hop causal chains involving three or more events are difficult to resolve through span extraction alone, as they require discourse-level reasoning that goes beyond identifying a single contiguous passage. Addressing these limitations likely requires models with explicit coreference and discourse structure awareness, rather than relying solely on local retrieval signals.

8. Conclusion

We presented a system for the FinCausal 2026 English subtask that combines parameter-efficient fine-tuning with a lightweight retrieval-augmented inference pipeline. By reformatting training data into the Qwen ChatML instruction format, fine-tuning Qwen3-4B-Instruct-2507 via QLoRA, and augmenting inference with TF-IDF intra-context retrieval and greedy decoding, our system achieves a score of **4.76 out of 5** on the official evaluation.

Our results demonstrate three broader findings. First, compact language models fine-tuned on modest domain-specific datasets can achieve strong performance on financial causal QA when paired with appropriate instruction formatting. Second, lightweight retrieval signals such as TF-IDF remain effective for grounding generation even without dense retrieval infrastructure. Third, the LLM-as-a-judge evaluation framework provides a more meaningful signal than lexical overlap metrics for tasks requiring causal reasoning, rewarding semantic correctness over verbatim matching.

Future work should explore discourse-aware models capable of resolving multi-hop causal chains, as well as dense retrieval methods that better capture semantic similarity between abstractive questions and their corresponding causal spans in financial text.

9. Generative AI Use Disclosure

Generative AI tools were used solely for language editing and paraphrasing during the preparation of this manuscript, including grammar correction and improving the clarity of written expressions. No generative AI tool was used to produce any scientific content, experimental results, analysis, or conclusions presented in this work. All authors are fully responsible and accountable for the content of this paper.

10. Acknowledgements

This research was conducted independently by the authors outside the scope of their professional responsibilities at Sony Research. The views and findings presented in this paper are solely those of the authors and do not reflect the positions or policies of Sony Research.

11. Bibliographical References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Eduardo Blanco, Nuria Castell, and Dan I Moldovan. 2008. Causal relation extraction. In *Lrec*, volume 66, page 74.

Pulkit Chatwal, Amit Agarwal, and Ankush Mittal. 2025. Enhancing causal relationship detection using prompt engineering and large language 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 248–252.

Qing Cheng, Zefan Zeng, Xingchen Hu, Yuehang Si, and Zhong Liu. 2024. A survey of event causality identification: Taxonomy, challenges, assessment, and prospects. *arXiv preprint arXiv:2411.10371*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Krish Didwania, Pratinav Seth, Aditya Kasliwal, and Amit Agarwal. 2024. Agrillm: harnessing transformers for framer queries. In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 179–187.

Quang Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 294–303, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Roxana Girju. 2003. Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering*, pages 76–83.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada. 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, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of finance*, 66(1):35–65.

Dominique Mariko, Hanna Abi-Akl, Estelle Labidurie, Stephane Durfort, Hugues De Mazancourt, and Mahmoud El-Haj. 2020. The financial document causality detection shared task (fincausal 2020). In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 23–32.

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.

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.

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.

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

Qwen Team. 2025. Qwen3 technical report.

Deepak Uniyal, Amit Agarwal, Durga Toshniwal, and Dipanjan Deb. 2021. Dense vector embedding based approach to identify prominent disseminators from twitter data amid covid-19 outbreak. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(3):308–320.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.