KULFi Framework: Knowledge Utilization for Optimizing Large Language Models for Financial Causal Reasoning

Neelesh K Shukla, Sandeep Singh, Prabhat Prabhakar, Sakthivel Thangaraj, Weiyi Sun, C Prasanna Venkatesan, Viji Krishnamurthy
OCI Generative AI Services, Oracle Corporation
Correspondence: neelesh.kumar.shukla@oracle.com

Abstract

This paper presents our contribution to the Financial Document Causality Detection (FinCausal) task 2025. The FinCausal challenge centers on the extraction of cause-and-effect relationships from financial texts written in both English and Spanish. We introduce KULFi, a novel Knowledge Utilization framework designed to augment the capabilities of Large Language Models (LLMs) by leveraging the expertise of more advanced reasoning models. Through the utilization of Teacher LLMs to generate task-specific instructions, KULFi optimizes the performance of Student LLMs via automated prompt optimization. We evaluate the efficacy of KULFi on the Financial Document Causality Detection Task, where Student LLM achieves a similarity score comparable to human-guided prompt optimization for the same LLM, demonstrating significant improvements in causal reasoning performance. Our results demonstrate that KULFi enables effective knowledge transfer from more robust models to less capable ones, as well as efficient learning from training data, minimizing the need for human input in prompt design and enabling more precise causal analysis in financial contexts. Our system attained SAS and Exact Match scores of 0.92 and 0.35 on the English dataset, and 0.92 and 0.09 on the Spanish dataset, respectively. This framework has far-reaching implications, with potential applications in enhancing decision-making across complex financial environments.

1 Introduction

The Financial Document Causality Detection Task (Moreno-Sandoval et al., 2025) focuses on determining the causes of changes in the financial environment to generate concise financial narrative summaries. It evaluates how events or chains of events lead to transformations in financial objects within specific contexts. Participants were tasked with identifying either the cause or effect for particular segments of text. The task consists of two subtasks, one in English and one in Spanish, using datasets from UK and Spanish financial annual reports to test the performance of multilingual models. Different from earlier editions (Moreno-Sandoval et al., 2023; Mariko et al., 2022) that used extractive methods, the 2025 task redefines the challenge as a generative AI problem, where systems generate cause-effect responses, assessed through exact match and similarity metrics.

Recently, the potential of LLMs to identify causal relationships and perform reasoning within natural language contexts has garnered significant attention (Section 2). Existing work (LYU et al., 2022) analyzes the approach of distinguishing between causal relationships ($X \rightarrow Y$) and their reverse ($Y \rightarrow X$) by framing an input-output learning task between the two variables. While this approach is effective for many task-specific models trained on input-output pairs, continued task-specific training may be impractical or prohibitively expensive for these general-purpose LLMs. In the era of Large Language Models (LLMs), Knowledge Distillation (KD) (Xu et al., 2024) is pivotal for transferring advanced capabilities from powerful models to weaker models on specific domains or tasks. This process mimics a skilled teacher imparting knowledge to a student, enhancing the performance of weaker models through the expertise of stronger ones.

In this work, we present Knowledge Utilization framework, **KULFi**, where a model with limited reasoning ability learns from a more capable reasoning model, specifically targeting Financial Causal Reasoning. Although not yet evaluated, this framework has the potential to be generalized to a wide range of tasks where prompt optimization or knowledge transfer is required to enhance performance.

Proceedings of the Joint Workshop of the 9th FinNLP, the 6th FNP, and the 1st LLMFinLegal, pages 222–229
January 19–20, 2025.

2 Related Works

2.1 Causal Reasoning with LLM

Recent studies have investigated the causal reasoning capabilities of LLMs. (Shukla et al., 2023) conducted an investigation of LLMs on FinCausal-2023 task using RAG based Few-Shot learning approach. (LYU et al., 2022) conducted a post-hoc analysis using natural language prompts to describe various potential causal narratives behind X-Y pairs. Despite the advancements, some studies (Zečević et al., 2023) argue that LLMs often function as "causal parrots," reiterating embedded causal knowledge without deep causal understanding. Overall, while numerous studies (Gao et al., 2023; Kiciman et al., 2024; Jin et al., 2024; Chen et al., 2024) acknowledge the strengths of LLMs in causal reasoning tasks, they also emphasize persistent limitations in reliably discerning causal relationships.

2.2 Knowledge Distillation

(Gu et al., 2024) introduced MINILLM, a novel approach using reverse KL divergence to help student models focus on key distribution modes, improving generative tasks' reliability. (Latif et al., 2024) demonstrated KD's effectiveness in educational tasks by distilling BERT-based models for automatic scoring, showing compact models' performance parity with larger ones in resource-constrained environments. (Xu et al., 2024) surveyed KD's role in compressing and self-improving LLMs, noting techniques like data augmentation to enhance training and make distilled models more cost-effective. These studies underscore KD's pivotal role in making LLMs more deployable while maintaining performance. We employed teacher-student learning to optimize prompts, enhancing overall results.

3 Definition of Causality and Task Dataset

3.1 Causality

The task defines causality as a relationship where a cause triggers an effect. Causes may involve agents or facts, while effects must be factual and not based on expectations or projections. Causes can be categorized as:

- *Justification of a statement.* (e.g., This is my final report since I have been succeeded as President of the Commission as of January 24, 2019).
- *The reason explaining a result.* (e.g., In Spain, revenue grew by 10.8% to 224.9 million euros due to increased cement volume and moderate price hikes).

3.2 Dataset Description

The dataset consists of three parts: context, question, and answer:

- *Context:* The original paragraph from the annual reports.
- *Question:* It is formulated to find the other part of the relationship, either the cause or the effect. It will always be abstractive, meaning it should reflect the content of the cause or effect being asked about, but not exactly match the provided context. For example:
  - Why did X (effect) happen?
  - What is the consequence (effect) of X (cause)?
- *Answer:* The answer will be the cause or effect previously questioned, extracted verbatim from the text, making it extractive. If a complex relationship appears (such as a causal chain of three or more elements or a complex relationship that is not a causal chain), a maximum of two questions will be asked.

The English dataset is drawn from various 2017 UK financial annual reports provided by the UCREL corpus at Lancaster University. The Spanish dataset is compiled from Spanish financial annual reports spanning 2014 to 2018. These datasets are aligned in both languages to facilitate multilingual model testing.

4 Initial Approach

4.1 Baseline: Default Prompt

The default prompt includes the definitions of causality and dataset, as specified in sections 3.1 and 3.2[1]. Additionally, it incorporates the Persona and Task outlined below.

Persona: *You are an expert in identifying causal relationships in financial reports.*

[1]https://www.lllf.uam.es/wordpress/fincausal-25/fnp-2025/

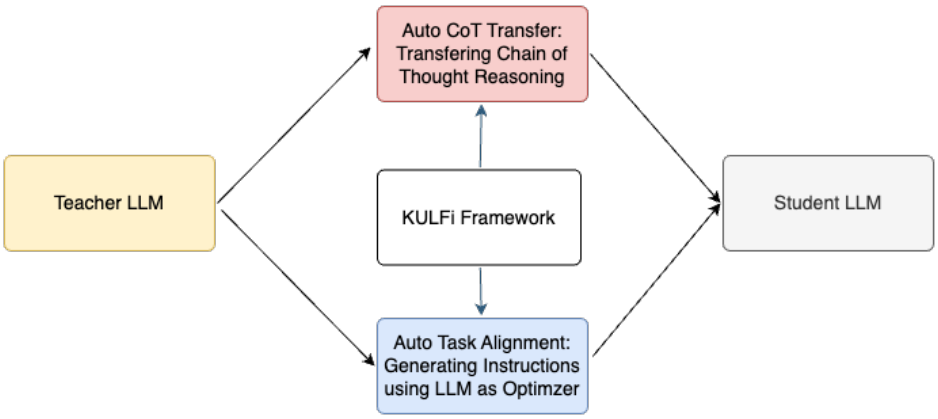

Figure 1: KULFi Framework

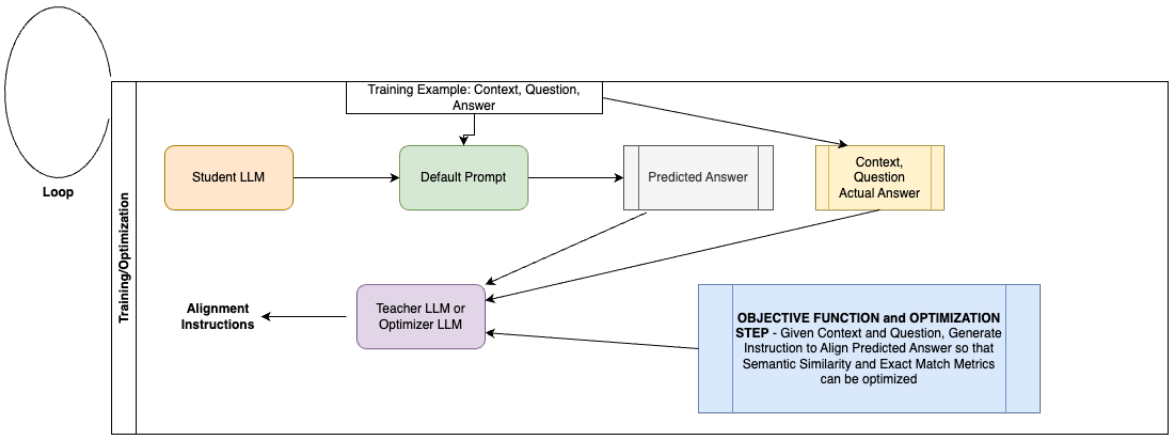

Figure 2: Auto Task Alignment: LLM as Optimizer

Task: *You will be provided an original paragraph from the annual reports as 'CONTEXT' and 'QUESTION' which is formulated to find the other part of the relationship, either the cause or the effect.*

Input
CONTEXT: %s
QUESTION: %s
ANSWER:

4.2 Data Analysis and Human-Guided Alignment Prompt

A manual review of the dataset confirmed that the ground truth answers were extractive. While the LLM-generated answers were similar to the ground truth, they were not extractive in nature. To better align the answers, we incorporated additional manual instructions to make the task explicitly extractive and review the answer post generation.

Additional Instruction: *Your task is to extract an 'ANSWER' directly from the provided CONTEXT. The 'ANSWER' must be a verbatim excerpt from the CONTEXT, meaning it should not be paraphrased or altered in any way. This is an extractive task. After extraction, review the 'ANSWER' to ensure it exactly matches the wording in the original text, without any modifications.*

5 KULFi Framework

While human-guided prompt engineering improves LLM performance, it requires domain-specific expertise, making it labor-intensive, dataset-specific. Fine-tuning LLMs on the given training data requires substantial computational resources, which can be a significant barrier for smaller teams and limited budgets. Fine-tuned models also risk limited adaptability to new information and may suffer from catastrophic forgetting (Luo et al., 2024).

An alternative approach could be automatic prompt optimization using training data, which reduces both cost of training LLM and human involvement in designing prompts. Our preliminary analysis shows that some LLMs possess inherently stronger reasoning abilities than others. We

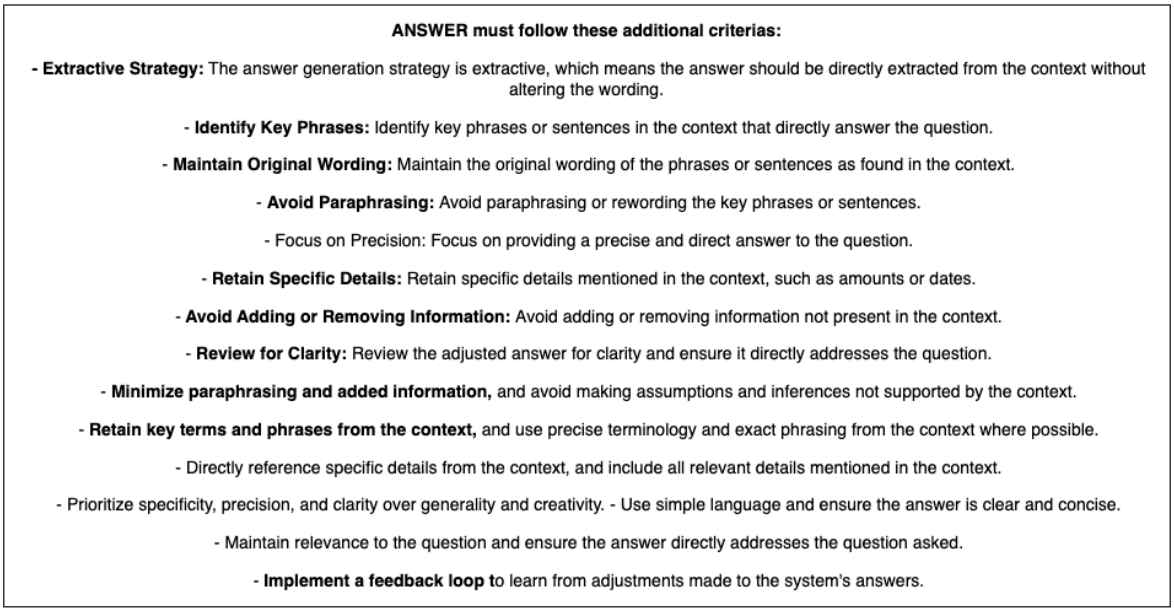

Figure 3: Alignment Instructions Generated by LLM as Optimizer

present KULFi—Knowledge Utilization for Optimizing LLMs, an automated framework (Figure 1) that employs Prompt Optimization using a Teacher-Student model. The Teacher refines prompts based on the Student's performance, iteratively enhancing output quality. Its functions as follows:

- The Student LLM harnesses the reasoning abilities of the Teacher LLM via Chain-of-Thought (CoT) generation, (Auto CoT Transfer).
- The Teacher LLM generates task-specific instructions, functioning as an optimizer to align the Student LLM with task requirements (Auto Task Alignment).

Optimized prompt instructions were generated as outlined in the following sections and added to the default prompt (Section 4.1) for the Student LLM.

5.1 Auto CoT Transfer

Chain-of-thought (CoT) prompting enables complex reasoning through intermediate steps (Wei et al., 2023). The Teacher LLM was provided with training examples <Context, Question, Answer> and default prompt (Section 4.1), with added instructions to generate and then summarize CoT for each example

Prompt Instruction for generation of CoT: *Please explain your chain of thought to reach to the answer. We want to convert that to a framework which can help improve weaker LLMs.*
CHAIN OF THOUGHT:

Chain of Thought Instructions Generated by Teacher LLM:

Follow step-by-step approach that involves:

1. Identifying key elements: Recognize the key elements in the context, such as the cause and effect.

2. Determining the question type: Determine whether the question is asking for a cause or an effect.

3. Locating the causal relationship: Find the sentence or phrase that describes the causal relationship between the cause and effect.

4. Extracting the answer: Extract the relevant information from the context that answers the question, ensuring it is a verbatim excerpt.

5. Verifying the answer: Review the extracted answer to ensure it matches the original text and logically answers the question.

5.2 Auto Task Alignment using LLM as Optimizer

We propose leveraging LLMs as optimizers (Figure 2) ,with the optimization task described in natural language, similar to the approach of (Yang et al., 2024). In each iteration, the Student LLM is given training examples in the form <Context, Question, Answer> and generates an answer using the default prompt. The Teacher LLM then evaluates the generated answer against the ground truth based

Model	Approach	SAS	EM	ROUGE-L	Dataset
Command R+	Default Prompt	0.765	0.009	0.515	EN-Practice
Command R+	Default Prompt + Human Alignment	0.887	0.218	0.814	EN-Practice
Command R+	Default Prompt + KULFi Framework	**0.880**	0.079	0.766	EN-Practice
Command R+	Default Prompt	0.767	0.009	0.422	ES-Practice
Command R+	Default Prompt + Human Alignment	0.859	0.079	0.778	ES-Practice
Command R+	Default Prompt + KULFi Framework	**0.845**	0.04	0.700	ES-Practice
Command R+	Default Prompt	0.766	0.002	0.477	EN-Test
Command R+	Default Prompt + Human Alignment	0.885	0.174	0.814	EN-Test
Command R+	Default Prompt + KULFi Framework	**0.878**	0.072	0.771	EN-Test
Command R+	Default Prompt	0.770	0.004	0.466	ES-Test
Command R+	Default Prompt + Human Alignment	0.895	0.094	0.810	ES-Test
Command R+	Default Prompt + KULFi Framework	**0.885**	0.048	0.736	ES-Test
Command R+	Default Prompt	0.754	0.002	NA	EN-Eval
Command R+	Default Prompt + Human Alignment	0.876	0.144	NA	EN-Eval
Command R+	Default Prompt + KULFi	**0.853**	0.064	NA	EN-Eval
Command R+	Default Prompt	0.772	0.002	NA	ES-Eval
Command R+	Default Prompt + Human Alignment	0.899	0.059	NA	ES-Eval
Command R+	Default Prompt + KULFi Framework	**0.879**	0.044	NA	ES-Eval

Table 1: Results of Command R+ (Student LLM) on English (EN) and Spanish (ES) datasets, where the KULFi framework achieves performance comparable to human-guided prompts.

on the objective function and provides alignment instructions. These prompt instructions serve as pseudo-weights, which the Teacher LLM optimizes in each iteration to optimize the objective function.

Optimizer Prompt and Objective Function

1. Evaluate both the SYS_ANSWER and ACTUAL_ANSWER based on semantic similarity and exact match metrics.

2. Provide detailed instructions to adjust the SYS_ANSWER to align with the ACTUAL_ANSWER, taking into account the CONTEXT and QUESTION, and ensuring the system's response optimizes these metrics.

We used 100 randomly selected training examples and performed iterations over them. Figure 3 shows the answer alignment instructions generated by the optimizer, or Teacher LLM.

6 Experiment Setup

We utilized the Llama3.1-405B[2] and Cohere Command R+[3] models, available as OCI GenAI Services offerings[4]. For both models, the temperature and frequency penalty were set to 0.0, and the top-p value was set to 0.95, with all other parameters left at their default values. Llama3.1-405B demonstrated superior performance with default prompts (Table 1, 2), and was selected as the Teacher model to guide Command R+ within the KULFi framework. To prepare the dataset, we randomly selected 25% of the training dataset as a test set. The approach was further evaluated on the organizers' practice and evaluation datasets. Metrics included exact matching, semantic similarity (SAS). We also used ROUGE-L (Lin, 2004) for assessing extractiveness using the longest common subsequence (LCS), providing a more suitable alternative to Exact Match.

7 Results Discussion and Error Analysis

Using the KULFi framework, the performance of the Student LLM, Command R+, consistently outperformed the default prompt and matched the performance of human-guided prompts (Table 1). This underscores the effectiveness of KULFi's automated prompt instruction generation approach. The Llama3.1-405B model performed well with the default prompt, and its performance improved further with human-guided prompt engineering (Table 2).

With a similarity score of approximately 92%, the system exhibits robust performance, with errors primarily concentrated in specific cases. A detailed

[2]https://ai.meta.com/blog/meta-llama-3-1/

[3]https://docs.cohere.com/v2/docs/command-r-plus

[4]https://www.oracle.com/in/artificial-intelligence/generative-ai/generative-ai-service/features/#models

Model	Approach	SAS	EM	ROUGE-L	Dataset
Llama 3.1 405B	Default Prompt	0.872	0.039	0.773	EN-Practice
Llama 3.1 405B	Default Prompt + Human Alignment	0.916	0.287	0.870	EN-Practice
Llama 3.1 405B	Default Prompt	0.875	0.03	0.751	ES-Practice
Llama 3.1 405B	Default Prompt + Human Alignment	0.862	0.069	0.797	ES-Practice
Llama 3.1 405B	Default Prompt	0.887	0.010	0.785	EN-Test
Llama 3.1 405B	Default Prompt + Human Alignment	0.924	0.258	0.886	EN-Test
Llama 3.1 405B	Default Prompt	0.891	0.004	0.767	ES-Test
Llama 3.1 405B	Default Prompt + Human Alignment	0.910	0.116	0.859	ES-Test
Llama 3.1 405B	Default Prompt	0.884	0.014	NA	EN-Eval
Llama 3.1 405B	Default Prompt + Human Alignment	0.924	0.353	NA	EN-Eval
Llama 3.1 405B	Default Prompt	0.893	0.008	NA	ES-Eval
Llama 3.1 405B	Default Prompt + Human Alignment	0.922	0.090	NA	ES-Eval

Table 2: Performance of LLama 3.1-405B (Teacher LLM) on Practice, Test, and Evaluation Datasets in English (EN) and Spanish (ES).

Question	Context	Actual Answer	System Answer	SAS	Error Analysis
What helps ensure that the selected candidates bring diverse perspectives?	Non-Executive Directors are appointed to the Board following a formal, rigorous and transparent process, involving external recruitment agencies, to select individuals who have a depth and breadth of relevant experience, thus ensuring that the selected candidates will be capable of making an effective and relevant contribution to the Group.	Non-Executive Directors are appointed to the Board following a formal, rigorous and transparent process, involving external recruitment agencies, to select individuals who have a depth and breadth of relevant experience	a depth and breadth of relevant experience	0.3	The predicted answer is incomplete, providing only part of the sentence. The full answer, which includes details on the appointment process, may be truncated by the system or lacks the subject (Non-Executive Directors) for context
What does the evaluation conducted by the Committee entail?	The main responsibilities of the Committee, in relation to nomination, are: evaluating the current balance of skills, experience, independence and knowledge of the Board and within the senior management team and, in light of this evaluation, preparing a description of the role and capabilities required for particular appointments	preparing a description of the role and capabilities required for particular appointments	evaluating the current balance of skills, experience, independence and knowledge of the Board and within the senior management team	0.55	In this case, we believe the system provides the correct output, including the necessary evaluation components that the ground truth lacks.
What is the reason behind the importance of drawing directors from the widest talent pool?	Board composition I believe that a board sets the tone for the entire business that it governs. This is why it is so important that the directors are drawn from the widest talent pool, best reflecting our society, as well as bringing the right mix of skills, diversity and experience	I believe that a board sets the tone for the entire business that it governs	so that the directors best reflect our society, as well as bring the right mix of skills, diversity and experience	0.45	The system's predicted answer is partially correct, while the ground truth provides fuller reasoning ("sets the tone for the entire company"). This may indicate the system's limited grasp of causal reasoning in case of alternative or supplementary causes.

Table 3: Error Analysis of Examples with Low Similarity Scores

error analysis (Table 3) reveals that errors mainly arise from responses that are either overly detailed or incomplete, often omitting key causal elements in cases with multiple causes and trasitive causes. Additionally, some inconsistencies are attributed to inaccuracies within the ground truth data.

Limitations

The dataset in this study primarily consists of brief contexts, generally limited to 2-3 sentences. Future research could investigate how reasoning performance is affected with longer contexts. We observed that LLMs exhibit limited capability in capturing complex causal reasoning, especially in cases involving transitive causation or multiple causal relationships. Although our optimizer is theoretically expected to surpass few-shot examples in effectiveness, it is unlikely to reach the performance level of supervised fine-tuning (SFT). Given SFT's high computational costs, it was excluded from this study, though it remains a promising direction for future exploration.

Ethical Considerations

This research emphasizes ethical considerations by basing all claims on experimental results, ensuring transparent documentation of methodologies, and sourcing datasets ethically with the necessary permissions.

References

Sirui Chen, Bo Peng, Meiqi Chen, Ruiqi Wang, Mengying Xu, Xingyu Zeng, Rui Zhao, Shengjie Zhao, Yu Qiao, and Chaochao Lu. 2024. Causal evaluation of language models. *Preprint*, arXiv:2405.00622.

Jinglong Gao, Xiao Ding, Bing Qin, and Ting Liu. 2023. Is ChatGPT a good causal reasoner? a comprehensive evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11111–11126, Singapore. Association for Computational Linguistics.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. Minillm: Knowledge distillation of large language models. *Preprint*, arXiv:2306.08543.

Zhijing Jin, Jiarui Liu, Zhiheng LYU, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona T. Diab, and Bernhard Schölkopf. 2024. Can large language models infer causation from correlation? In *The Twelfth International Conference on Learning Representations*.

Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. 2024. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*. Featured Certification.

Ehsan Latif, Luyang Fang, Ping Ma, and Xiaoming Zhai. 2024. Knowledge distillation of llm for automatic scoring of science education assessments. *Preprint*, arXiv:2312.15842.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2024. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *Preprint*, arXiv:2308.08747.

Zhiheng LYU, Zhijing Jin, Rada Mihalcea, Mrinmaya Sachan, and Bernhard Schölkopf. 2022. Can large language models distinguish cause from effect? In *UAI 2022 Workshop on Causal Representation Learning*.

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) - at COLING-2025*.

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.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. 2024. A survey on knowledge

distillation of large language models. *Preprint*, arXiv:2402.13116.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. 2024. Large language models as optimizers. *Preprint*, arXiv:2309.03409.

Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. 2023. Causal parrots: Large language models may talk causality but are not causal. *Transactions on Machine Learning Research*.