Natural Language Inference as a Judge: Detecting Factuality and Causality Issues in Language Model Self-Reasoning for Financial Analysis

Yilin Wu[†], **Han Yuan**[†], **Li Zhang, Zheng Ma**[✉]
Global Decision Science, American Express
{Yilin.Wu, Han.Yuan1, Li.Zhang1, Zheng.Ma2}@aexp.com

Abstract

Language models (LMs) have revolutionized financial analysis by demonstrating expert-level versatility. Recent advances in self-reasoning have further improved LMs' performance on complex tasks. However, LMs are known to hallucinate facts and generate non-causal reasoning paths, which compromise their output quality, lead to erroneous conclusions, and pose risks of monetary losses. Therefore, detecting factual and causal errors in LMs' reasoning is essential for risk management and responsible application of LMs in finance. In this study, we adopt natural language inference (NLI) as a paradigm for detecting factual and causal errors in LMs' reasoning. We evaluate this approach by constructing a dataset comprising financial reasoning points generated by LMs, along with annotations by domain experts. Our findings demonstrate that NLI, powered by backbones of either pre-trained encoders or LMs, exhibits statistically significant capability in detecting factual and causal issues. Also, we show that, although LMs achieve improved performance with increasing parameters, they underperform encoders and exhibit self-evaluation bias. Fine-tuning effectively mitigates this type of bias and enhances both backbones' detection capability.

1 Introduction

Language models (LMs) have transformed financial natural language processing (NLP) through their expert-level comprehension of financial information and versatile problem solving capabilities according to users' instructional prompts (Li et al., 2023; Kong et al., 2024; Hu et al., 2025). Recent advancements in self-reasoning (Liu et al., 2024a) have further enhanced LMs' ability to tackle complex jobs that cannot be resolved through direct question answering. However, LMs are known to hallucinating facts or producing non-causal statements during the reasoning process, which can lead to erroneous conclusions and compromise the quality of their outputs (Manakul et al., 2023; Laban et al., 2023; Li et al., 2024; Paul et al., 2024; Chandler et al., 2024; Chen et al., 2025). Such issues pose significant risks in financial applications, where inaccuracies result in monetary losses (Chatwal et al., 2025; Shukla et al., 2025). Even when the final outcome is correct, flawed reasoning steps may mislead users who interpret these steps as justifications for LMs' decision and indicate that the outcome was reached by chance rather than logic (Wu et al., 2024; Wang, 2024; Chu et al., 2025; Bao et al., 2025). Sole outcome evaluation risks overlooking deficiencies in the underlying reasoning and potentially leading to monetary losses.

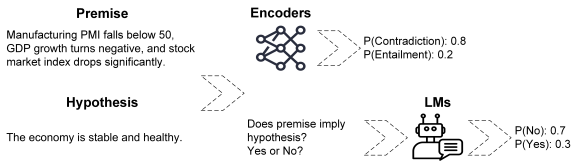

Figure 1: NLI can detect factual and causal errors in LMs' self-reasoning for financial analysis

Therefore, detecting factual and causal errors in LMs' reasoning is crucial for mitigating potential risks in financial decisions and for supporting effective regulation and compliance (Chatwal et al., 2025). In this study, we adopt a classic and computationally affordable paradigm, natural language inference (NLI), in identifying factual and causal errors in LMs' reasoning (Lattimer et al., 2023). To address the absence of well-annotated datasets aligned with our objectives, we construct a specialized dataset by employing LMs to generate final decisions and reasoning process on a public financial

[†] These authors contributed equally to this work.
[✉] Correspondence: Zheng Ma, Singapore Decision Science Center of Excellence, American Express, 1 Marina Boulevard, 018989, Singapore.

Proceedings of The 10th Workshop on Financial Technology and Natural Language. EMNLP-2025, Suzhou, China

dataset. Domain experts then manually annotate the factuality and causality of each reasoning point. After that, we test pre-trained encoders and LMs as NLI backbones to derive the probability of factual or causal issues. Finally, We perform rigorous statistical analyses to evaluate the feasibility of NLI as a paradigm for detecting factuality and causality, compare the performance of pre-trained and fine-tuned encoders and LMs, and investigate biases when LMs assess their proprietary reasoning.

As a pilot study on LMs' self-reasoning in finance, our work contributes in four aspects. First, we provide an annotated dataset with labels of factuality and causality on LMs' reasoning points. Second, we demonstrate the effectiveness of the classic NLI as a detection paradigm for factual and causal errors, using encoders and LMs as backbones. Third, we perform referable statistical analyses to illustrate limitations of LMs in this task: their inferior accuracy compared to encoders and potential biases when assessing proprietary reasoning in certain scenarios. Last, we demonstrate the necessity of fine-tuning, which not only enhances the detection ability of both backbones but also mitigates LMs' self-evaluation bias. Relevant dataset and notebooks are open-accessed on GitHub[1].

2 Related work

As a fundamental NLP task, NLI determines the logical relationship between a given pair of sentences: a premise and a hypothesis. Typically, transformer encoder-based NLI models (Devlin et al., 2019) output three probabilities: entailment, contradiction, and neutrality (Gubelmann et al., 2024; Guo and Yang, 2024; Magomere et al., 2025). Specifically, entailment indicates that the hypothesis logically follows from the premise, contradiction signifies that the hypothesis is false given the premise, and neutrality implies that the premise is insufficient to determine the truth of the hypothesis.

NLI plays a crucial role in tasks involving causality, and its capabilities have significantly improved with the evolution of foundational models from pre-trained encoders to LMs (Rozanova et al., 2023; Guo and Yang, 2024). For example, Ionescu et al. (2020) employ five pre-trained encoders to examine causality in financial documents. Pre-trained encoders, in addition to being used for post hoc causality detection, can also be integrated in real-time content generation. ConCoRD (Mitchell et al., 2022) is a framework that enhances LMs' output quality by selecting optimal sentences that maintain causal consistency throughout the generation process. With LMs' advancement, they outperform specialized pre-trained encoders in some tasks.

Beyond its original purpose of causality, NLI has also proven effective in tasks concerning factuality. Similar to causality, both pre-trained encoders (Kryscinski et al., 2020; Goyal and Durrett, 2020; Sathe and Park, 2021; Fabbri et al., 2022; Utama et al., 2022; Ni et al., 2024; Yang et al., 2024) and LMs (Fatahi Bayat et al., 2023; Lattimer et al., 2023; Li et al., 2024) have been employed for factuality detection. SummaC (Laban et al., 2022) is a comprehensive benchmark for evaluating the performance of NLI encoders in factuality detection. It demonstrates that NLI encoders based on classic architectures, such as BERT (Devlin et al., 2019), can achieve a balanced accuracy of nearly 0.75. A recent comprehensive framework, SelfCheckGPT (Manakul et al., 2023), integrates both pre-trained encoders and LMs to perform NLI for assessing the factuality of LMs' generated Wikipedia content.

3 NLI as a Judge

NLI evaluates whether a hypothesis follows from a premise, producing probabilities for three possible outcomes: entailment, neutrality, and contradiction (Yu et al., 2024). Our study adopts NLI as the framework for detecting factual and causal issues in LMs' self-reasoning for financial analysis.

Formally, D denotes the input dataset and D_i refers to a specific case within D. Each D_i contains J sentences, denoted as $D_{i,j}$ $(j = 1, 2, ..., J)$, which provide various input details for financial classification. Given D_i as input, a LM generates a response O_i, comprising K sentences of $O_{i,k}$. The first sentence, $O_{i,1}$, states the classification outcome for D_i. The subsequent sentences, $O_{i,k}$ $(k = 2, ..., K)$, outline the reasoning points underlying this classification and the primary focus of this study is to detect factual and causal errors in $O_{i,k}$ through the paradigm of NLI.

Specifically, NLI takes a premise S_p and a hypothesis S_h as input. Then it outputs probabilities of three possible outcomes: entailment, neutrality, and contradiction. For factuality detection (Utama et al., 2022), the premise S_p corresponds to the input information D_i and the hypothesis S_h is each reasoning statement $O_{i,k}$ $(k = 2, ..., K)$.

[1] https://github.com/Han-Yuan-Med/nli-as-a-judge

For causality detection, the S_p is the reasoning statement $O_{i,k}$ $(k = 2, ..., K)$ and S_h is the classification outcome $O_{i,1}$. Following Manakul et al. (2023), we omit the neutral class and focus only on the probability of entailment $P_e(S_p, S_h)$ and contradiction $P_c(S_p, S_h)$. With this simplification, the output becomes binary and is further normalized to ensure the entailment probability $P'_e = P_e/(P_e + P_c)$ to be bounded within $[0, 1]$. For both factuality and causality detection of $O_{i,k}$, a reasoning point is classified as containing factual or causal errors if P'_e is less than 0.5. We adopt a threshold of 0.5 because our dataset is relatively small, reserving a separate validation set for threshold optimization would further reduce the effective training data and increase the overfitting risk, and this choice is consistent with established practice (Kazemi et al., 2023; Chicco and Jurman, 2023).

For the comprehensiveness, both classic encoders and LMs are used as backbones for calculating P'_e. For encoders pre-trained on NLI, the output has been shaped into probability suiting the formulation. For LMs, we follow the design in Lattimer et al. (2023) and prompt the LMs with the following template: *"S_p Question: does this imply S_h? Yes or No?"*. The logits corresponding to *"Yes"* and *"No"* are extracted as $P_e(S_p, S_h)$ and $P_c(S_p, S_h)$, respectively. The final probability P'_e is then computed as illustrated above. The simple prompt design is adopted to enhance computational efficiency, eliminate variance introduced by prompt optimization, facilitate domain-agnostic assessment without the need for adaptation, enable the evaluation of long test cases by employing short prompts with fewer tokens (Laban et al., 2023), and eliminate hallucination introduced by techniques such as In-Context Learning (ICL) and Chain-of-Thought (CoT) (Gao et al., 2023; Paul et al., 2024; Zhang et al., 2024; Turpin et al., 2023).

4 Dataset

We conduct our experiments using a refined version (Yuan et al., 2025) of the public German credit dataset (Hofmann, 1994) with increased signal-to-noise ratio and better alignment with modern LMs' training context. Since no standard annotations of factuality and causality in LMs' generated reasoning points are available for this dataset, we construct our own data through a two-step process: (1) collecting LMs' responses, including both classification outcomes and reasoning points; and (2) manually annotating the reasoning sentences for factual and causal issues.

Following Zhang et al. (2024), we utilize the processed data, formatted as text input, to prompt three LMs, Llama-3.2-3B (Touvron et al., 2023), Gemma-2-2B (Mesnard et al., 2024), and Phi-3.5-mini (3.8B) (Abdin et al., 2024), to generate both classification outcomes and the underlying reasoning points behind their decisions. Specifically, the three LMs generate 862, 495, and 515 reasoning points, respectively, for 50 positive and 50 negative cases. This suggests that Llama, on average, employs more reasoning points than the other two.

After that, two authors annotate the reasoning points using a two-step workflow. Each point is first assessed for factuality issues, defined as the involvement of non-factual information. If no factual errors are found, the reasoning point is further evaluated for causality issues, also referred to as logical inconsistencies. A causality issue is identified when a negatively framed reasoning point is incorrectly presented as supporting a positive classification, or vice versa (Yuan et al., 2025). After independent annotation, the two annotators summarize conflicting cases and consult the senior authors to resolve discrepancies and reach consensuses. Among the 1,872 annotated reasoning points, 72 (3.9%) were labeled as factually inaccurate, and 329 (17.6%) were identified as causally erroneous.

5 Experiments

We evaluate both pre-trained encoders and LMs, along with their fine-tuned versions, on the annotated data for factuality and causality detection. Pre-trained models are used in their original form as released on HuggingFace. For fine-tuning, we explore full-parameter fine-tuning (FPFT) and two parameter-efficient fine-tuning (PEFT) methods (Appendix A) of Last Layer fine-tuning (LLFT) and Low-Rank Adaptation (LoRA) (Hu et al., 2022). Specifically, we apply three-fold cross-validation for all fine-tuning experiments, using one fold for testing and the remaining two for training in each run. A consistent training setup is adopted for both encoders and LMs, using the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 1e-5 and and default settings for other hyperparameters over five epochs. For encoders, the input consists of premise–hypothesis pairs and the output is a binary classification label (either entailment or contradiction). For LMs, the premise

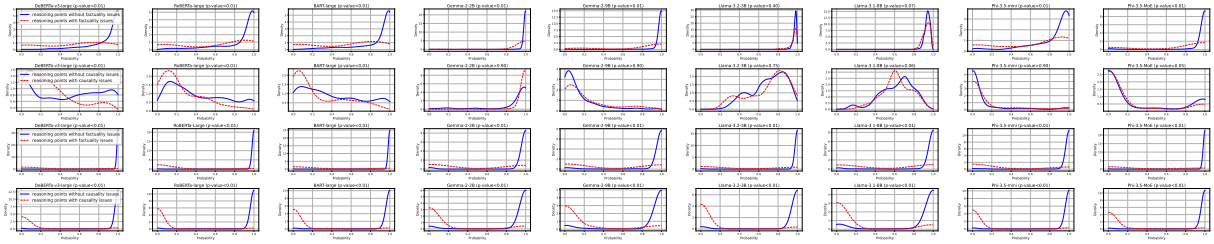

Figure 2: Entailment probability distributions for statements with and without factual or causal errors

and hypothesis are concatenated into a coherent instruction, and the model is trained to generate a target token, either *Yes* or *No*, reflecting the relationship between premise and hypothesis. We acknowledge that additional hyperparameter tuning and training techniques (e.g., warm-up schedules) may further enhance model performance. However, the primary objective of fine-tuning is to demonstrate its advantages over pre-trained models, rather than to achieve the upper-bound performance of fine-tuning, which is reserved for future work.

For backbones based on transformer encoders, we select DeBERTa-v3-large (He et al., 2021), RoBERTa-large (Liu et al., 2019), and BART-large (Lewis et al., 2020). For open-access (OA) LMs (Lasheras and Pinheiro, 2025), we utilize the same three families for dataset construction: Llama (Llama-3.2-3B and Llama-3.1-8B), Gemma (Gemma-2-2B and Gemma-2-9B), and Phi (Phi-3.5-mini and Phi-3.5-MoE). All OA models, except for Phi-3.5-MoE, have fewer than 10 billion parameters, aligning with the constraints of our computational resources. Although Phi-3.5-MoE contains a total of 60.8 billion parameters, only 6.6 billion parameters are active during any single inference due to its mixture-of-experts (MoE) architecture, thereby keeping computation within our budget. Section 3 details the process of obtaining the normalized entailment probability P'_e, which is used for performance comparison and statistical tests in the following sections. In addition to OA LMs, we include the proprietary GPT-4o (OpenAI, 2024) as a state-of-the-art (SOTA) backbone. It should be noted that GPT-4o is evaluated only under the pre-trained setting, as the internal training procedures and fine-tuning methodologies used by OpenAI are not publicly disclosed (OpenAI). To estimate P'_e, we use the same instruction prompt as for the OA LMs and query the API ten times, calculating the proportion of *"Yes"* as a proxy.

First, we assess the effectiveness of NLI as a detection paradigm for pre-trained backbones. Tables

Model	Mode	F1	BA	AUPRC	AUROC
DeBERTa-v3-large	Pre-trained	0.28	0.67	0.30	0.84
	FPFT	0.82	0.88	0.92	0.99
BART-large	Pre-trained	0.23	0.66	0.35	0.84
	FPFT	0.77	0.85	0.80	0.96
RoBERTa-large	Pre-trained	0.19	0.62	0.29	0.77
	FPFT	0.84	0.92	0.88	0.99
Llama-3.2-3B	Pre-trained	0.00	0.50	0.10	0.51
	FPFT	0.74	0.82	0.67	0.85
Llama-3.1-8B	Pre-trained	0.00	0.50	0.07	0.55
	FPFT	0.38	0.66	0.37	0.77
Gemma-2-2B	Pre-trained	0.09	0.53	0.12	0.71
	FPFT	0.44	0.70	0.40	0.77
Gemma-2-9B	Pre-trained	0.28	0.60	0.15	0.64
	FPFT	0.48	0.70	0.41	0.79
Phi-3.5-mini	Pre-trained	0.17	0.63	0.20	0.65
	FPFT	0.73	0.82	0.68	0.93
Phi-3.5-MoE	Pre-trained	0.22	0.60	0.21	0.62
	FPFT	0.84	0.89	0.86	0.95
GPT-4o	Pre-trained	0.32	0.76	0.28	0.80

Table 1: Factuality detection results of pre-trained and FPFT encoders and LMs under NLI paradigm

Model	Mode	F1	BA	AUPRC	AUROC
DeBERTa-v3-large	Pre-trained	0.37	0.62	0.21	0.59
	FPFT	0.92	0.95	0.92	0.98
BART-large	Pre-trained	0.34	0.52	0.28	0.64
	FPFT	0.91	0.96	0.92	0.98
RoBERTa-large	Pre-trained	0.36	0.61	0.36	0.67
	FPFT	0.92	0.96	0.94	0.99
Llama-3.2-3B	Pre-trained	0.19	0.51	0.24	0.49
	FPFT	0.86	0.92	0.91	0.97
Llama-3.1-8B	Pre-trained	0.18	0.48	0.19	0.53
	FPFT	0.88	0.92	0.85	0.95
Gemma-2-2B	Pre-trained	0.03	0.46	0.14	0.39
	FPFT	0.86	0.93	0.88	0.97
Gemma-2-9B	Pre-trained	0.28	0.46	0.16	0.42
	FPFT	0.74	0.89	0.82	0.95
Phi-3.5-mini	Pre-trained	0.31	0.50	0.14	0.39
	FPFT	0.91	0.95	0.92	0.98
Phi-3.5-MoE	Pre-trained	0.32	0.54	0.18	0.53
	FPFT	0.91	0.95	0.89	0.98
GPT-4o	Pre-trained	0.31	0.51	0.19	0.52

Table 2: Causality detection results of pre-trained and FPFT encoders and LMs under NLI paradigm

1 and 2 present the performance of pre-trained and FPFT backbones in terms of F1 score, balanced accuracy (BA) (Utama et al., 2022), the area under the precision-recall curve (AUPRC), and the area under the receiver operating characteristic curve (AUROC), ensuring a robust comparison in the scenario

of class imbalance (Yuan et al., 2022). Additionally, we adopt statistical tests to demonstrate the effectiveness of NLI as a detection paradigm. We collect P'_e for reasoning points with and without factual or causal errors and apply the Wilcoxon rank-sum test (Wilcoxon, 1947). The null hypothesis assumes no difference in P'_e between the two groups, while the alternative hypothesis asserts that sentences containing errors exhibit lower P'_e. A p-value below 0.05 rejects the null hypothesis and adopts the alternative hypothesis, indicating that NLI, powered by a certain backbone, has statistically significant distinguishability at the 95% confidence level. Figure 2 shows the entailment probability distribution of pre-trained and FPFT models across the two tasks. The red lines represent reasoning points containing errors, while the blue lines denote those without errors. Each subplot displays results for a specific backbone, with corresponding p-values shown at the top. The first two rows present results from pre-trained models on factuality and causality detection tasks, respectively and the bottom two rows show results from FPFT models on the two tasks. The statistically significant p-values demonstrate that **NLI is an effective paradigm for distinguishing sentences containing factual or causal errors** in both pre-trained and fine-tuning settings.

Second, we aim to compare the discriminability of different backbone models. We employ the same rank-sum test for this comparison, conducting separate tests on sentences with and without errors. For sentences containing factual or causal errors, the null hypothesis assumes no difference in P'_e between two backbones, B_1 and B_2, while the alternative hypothesis posits that P'_e from B_1 is lower than that from B_2. A p-value below 0.05 supports the alternative hypothesis, indicating that B_1 outperforms B_2 in identifying erroneous sentences. For sentences without factual or causal errors, the null hypothesis again assumes no difference in P'_e between B_1 and B_2, while the alternative hypothesis asserts that P'_e from B_1 is greater than that from B_2. A p-value below 0.05 supports the alternative hypothesis, demonstrating that B_1 excels in classifying error-free reasoning sentences.

Due to space constraints, we present capability comparisons in factuality detection on reasoning points containing factual errors under both pre-trained and fine-tuning settings, and comprehensive details are available in Appendix C. The color indicates the p-value from a pairwise comparison between the model in the column and the model in the row. A significant p-value illustrates that the column model significantly outperforms the row model. Figure 3 reveals that encoders outperform LMs in 17 out of 21 cases under the pre-trained setting. Figure 4 shows that encoders outperform LMs in 15 out of 18 cases under the FPFT setting. **Under the pre-trained setting, GPT-4o demonstrates consistent superiority over both encoders and other LMs in factual error detection**, aligning with its status as the SOTA model. However, its superiority does not extend to causal error detection (Appendix C). These results suggest that, despite the general superiority and widespread adoption across NLP tasks, **LMs achieve performance inferior to that of encoders in certain scenarios**. Laban et al. (2023) and Jin et al. (2024) report simi-

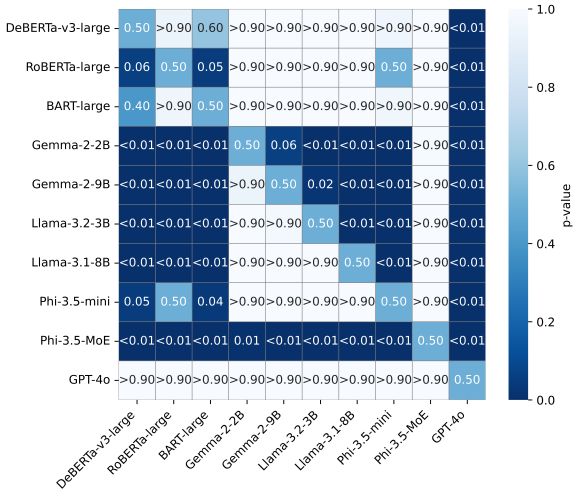

Figure 3: Pairwise comparison of factuality detection on erroneous reasoning points in the pre-trained setting

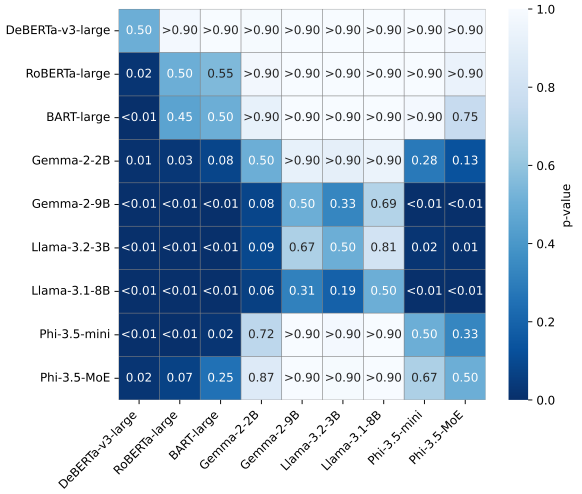

Figure 4: Pairwise comparison of factuality detection on erroneous reasoning points in the FPFT setting

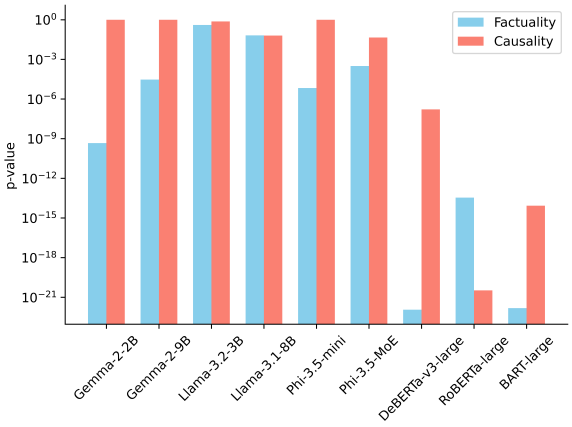

Figure 5: P-value difference in detection capability of pre-trained and FPFT models

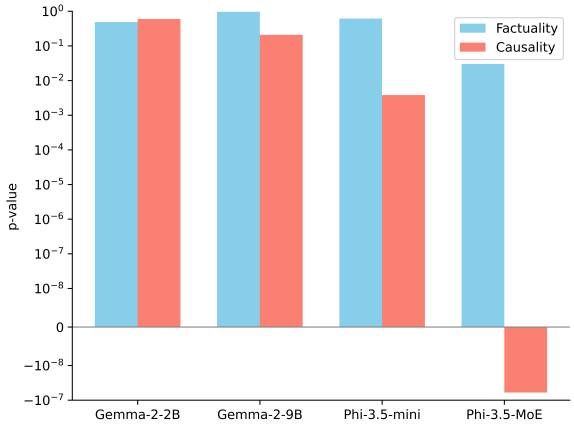

Figure 6: P-value difference in self-evaluation bias between FPFT and pre-trained models

lar findings that LMs, despite having several orders of magnitude more parameters than pre-trained encoders, achieve comparable performance across multiple benchmarks. Gao et al. (2023) demonstrate that ChatGPT, despite being one of the most well-aligned LMs, performs poorly in causal reasoning due to bias introduced during its upgrading training stages. Although increasing the size of OA LMs generally leads to improved performance, as reported by Laban et al. (2023), we do not observe emergent detection capabilities in our experiments. A potential explanation is that such abilities tend to emerge only in models exceeding 100 billion parameters from the same family (Paul et al., 2024; Kojima et al., 2022). Due to computational constraints, we did not test OA LMs of this scale. In addition, the SOTA LMs like GPT-4 exhibit relatively weak performance on causal understanding compared to other natural language understanding tasks (Wang et al., 2023; Romanou et al., 2023; Paul et al., 2024; Liu et al., 2024b), and do not significantly outperform encoders.

Third, we perform fine-tuning to compare the performance of pre-trained versus fine-tuned backbones. We use the same rank-sum test and Figure 5 reports the p-value differences between pre-trained and FPFT models. Smaller p-values indicate better discriminability; therefore, a positive difference, where the p-value of the pre-trained model is higher than that of the FPFT model, suggests that fine-tuning enhances the model's detection capability. The results show that all models exhibit reduced p-values after FPFT, confirming **the effectiveness of fine-tuning** in improving encoders' and LMs' detection performance of factual and causal issues.

Lastly, we investigate whether fine-tuning can mitigate the self-evaluation bias exhibited by LMs, whereby they tend to classify their own erroneous reasoning or that of models within the same family but with different parameter sizes as correct (Zheng et al., 2023). To quantify this bias, we apply the rank-sum test to compare P'_e assigned to erroneous proprietary reasoning versus erroneous reasoning from other models. The null hypothesis posits no difference of P'_e between the two groups, while the alternative hypothesis suggests that P'_e assigned to erroneous proprietary reasoning is higher than that for reasoning generated by other LMs, revealing that LMs are less capable of detecting errors in their own or closely related outputs compared to those from other LMs. Based on the computed p-values from pre-trained and FPFT LMs, Figure 6 presents the p-value differences between FPFT models and their pre-trained counterparts. A positive difference implies that the p-value of self-evaluation bias is higher in the FPFT model than in the pre-trained model, suggesting that **fine-tuning effectively mitigates LMs' self-evaluation bias**. The Llama family is excluded from this comparison due to their pre-trained versions' near-zero discriminability.

6 Conclusions

Our study investigates factual and causal error detection in financial analysis by LMs. We adopt NLI as the detection paradigm supported by both encoders and LMs. Our experiments show that while LMs outperform encoders in many financial NLP tasks, users should realize their potential disadvantages relative to encoders as well as their susceptibility to biases when evaluating proprietary

reasoning. Also, practitioners are advised, although pre-trained models show certain ability, to fine-tune models when resources permit, as it enhances discriminability and mitigate self-evaluation bias.

Limitations

First, we generated 1,872 reasoning points from responses of 3 LMs to 50 positive and 50 negative cases in a public dataset. To further validate our findings, future experiments should extend to additional tasks, a wider range of LMs, and diverse NLI backbone models. Second, the results indicate that LMs exhibit relatively weak performance compared to pre-trained encoders in certain scenarios, likely due to the absence of prompt engineering and reliance solely on the strategy of comparing response probabilities of *"Yes"* and *"No"* (Lattimer et al., 2023). Future work will explore prompt engineering to improve detection accuracy of pre-trained LMs (Shukla et al., 2025). Third, we do not conduct a thorough evaluation against top professionals, but for time-sensitive applications, AI models hold a clear advantage since human experts cannot process thousands of pieces of information within seconds. Last, we do not examine detection methods such as keyword-based approaches, and future work will evaluate whether NLI offers meaningful improvements over these simpler methods.

Ethics statement

This study investigates the factual and causal errors in the reasoning process of LMs within the financial domain. We demonstrate that NLI is a computationally efficient detection paradigm. Our results indicate that its current performance, including leveraging LMs as backbones, remains suboptimal. This aligns with findings by Lasheras and Pinheiro (2025) that even advanced models such as GPT-4o exhibit limited capability in causal reasoning. Additionally, most existing benchmarks for factuality and causality detection are built on English tasks and datasets, often overlooking the pragmatic differences and cultural nuances inherent in other languages (Lasheras and Pinheiro, 2025). Therefore, users are recommended to conduct thorough evaluations before deploying NLI-based detection backbones in real-world applications.

Disclaimer

This paper is provided solely for informational purposes as an academic contribution by the authors to the research community and does not represent, reflect, or constitute the views, policies, positions, or practices of American Express or its affiliates. Nothing in this paper should be cited or relied upon as evidence of, or support for, the business views, policies, positions, or practices of American Express or its affiliates.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone.

Guangsheng Bao, Hongbo Zhang, Cunxiang Wang, et al. 2025. How likely do LLMs with CoT mimic human reasoning? In *Proceedings of the International Conference on Computational Linguistics*.

Alex Chandler, Devesh Surve, and Hui Su. 2024. Detecting errors through ensembling prompts (DEEP): An end-to-end LLM framework for detecting factual errors. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

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 Financial Technology and Natural Language Processing, the Financial Narrative Processing, and the Workshop on Large Language Models for Finance and Legal*.

Yingjian Chen, Haoran Liu, Yinhong Liu, Jinxiang Xie, Rui Yang, Han Yuan, Yanran Fu, Peng Yuan Zhou, Qingyu Chen, James Caverlee, and Irene Li. 2025. GraphCheck: Breaking long-term text barriers with extracted knowledge graph-powered fact-checking. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14976–14995.

Davide Chicco and Giuseppe Jurman. 2023. The matthews correlation coefficient (mcc) should replace the roc auc as the standard metric for assessing binary classification. *BioData Mining*, 16(1):4.

Zheng Chu, Jingchang Chen, Zhongjie Wang, et al. 2025. Towards faithful multi-step reasoning through fine-grained causal-aware attribution reasoning distillation. In *Proceedings of the International Conference on Computational Linguistics*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, et al. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, et al. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

Farima Fatahi Bayat, Kun Qian, Benjamin Han, et al. 2023. FLEEK: Factual error detection and correction with evidence retrieved from external knowledge. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Jinglong Gao, Xiao Ding, Bing Qin, et al. 2023. Is ChatGPT a good causal reasoner? a comprehensive evaluation. In *Findings of the Association for Computational Linguistics*.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics*.

Reto Gubelmann, Ioannis Katis, Christina Niklaus, et al. 2024. Capturing the varieties of natural language inference: A systematic survey of existing datasets and two novel benchmarks. *Journal of Logic, Language and Information*, 33(1):21–48.

Yue Guo and Yi Yang. 2024. EconNLI: Evaluating large language models on economics reasoning. In *Findings of the Association for Computational Linguistics*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, et al. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *Proceedings of the International Conference on Learning Representations*.

Hans Hofmann. 1994. Statlog (German Credit Data). UCI Machine Learning Repository.

Bo Hu, Han Yuan, Vlad Pandelea, Wuqiong Luo, Yingzhu Zhao, and Zheng Ma. 2025. Extract, match, and score: An evaluation paradigm for long question-context-answer triplets in financial analysis. In *ICLR 2025 Workshop on Advances in Financial AI: Opportunities, Innovations and Responsible AI*.

Edward J Hu, yelong shen, Phillip Wallis, et al. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations*.

Marius Ionescu, Andrei-Marius Avram, George-Andrei Dima, et al. 2020. UPB at FinCausal-2020, tasks 1 & 2: Causality analysis in financial documents using pretrained language models. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*.

Zhijing Jin, Jiarui Liu, Zhiheng LYU, et al. 2024. Can large language models infer causation from correlation? In *Proceedings of the International Conference on Learning Representations*.

Hamid Reza Kazemi, Kaveh Khalili-Damghani, and Soheil Sadi-Nezhad. 2023. Estimation of optimum thresholds for binary classification using genetic algorithm: An application to solve a credit scoring problem. *Expert Systems*, 40(3):e13203.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, et al. 2022. Large language models are zero-shot reasoners. In *Proceedings of the International Conference on Neural Information Processing Systems*.

Yaxuan Kong, Yuqi Nie, Xiaowen Dong, et al. 2024. Large language models for financial and investment management: Models, opportunities, and challenges. *Journal of Portfolio Management*, 51(2).

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, et al. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Philippe Laban, Wojciech Kryscinski, Divyansh Agarwal, et al. 2023. SummEdits: Measuring LLM ability at factual reasoning through the lens of summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, et al. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10.

Uriel Anderson Lasheras and Vládia Pinheiro. 2025. Calquest. pt: Towards the collection and evaluation of natural causal ladder questions in portuguese for ai agents. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*.

Barrett Lattimer, Patrick H. Chen, Xinyuan Zhang, et al. 2023. Fast and accurate factual inconsistency detection over long documents. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Mike Lewis, Yinhan Liu, Naman Goyal, et al. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Junyi Li, Jie Chen, Ruiyang Ren, et al. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Yinheng Li, Shaofei Wang, Han Ding, et al. 2023. Large language models in finance: A survey. In *Proceedings of the ACM International Conference on AI in Finance*.

Aixin Liu, Bei Feng, Bing Xue, et al. 2024a. Deepseek-v3 technical report. *arXiv*.

Xiao Liu, Zirui Wu, Xueqing Wu, et al. 2024b. Are LLMs capable of data-based statistical and causal reasoning? benchmarking advanced quantitative reasoning with data. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Yinhan Liu, Myle Ott, Naman Goyal, et al. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the International Conference on Machine Learning*.

Jabez Magomere, Elena Kochkina, Samuel Mensah, et al. 2025. FinNLI: Novel dataset for multi-genre financial natural language inference benchmarking. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4545–4568.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, et al. 2024. Gemma: Open models based on gemini research and technology.

Eric Mitchell, Joseph Noh, Siyan Li, et al. 2022. Enhancing self-consistency and performance of pretrained language models through natural language inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Jingwei Ni, Minjing Shi, Dominik Stammbach, et al. 2024. AFaCTA: Assisting the annotation of factual claim detection with reliable LLM annotators. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

OpenAI. Fine-tuning now available for gpt-4o. https://openai.com/index/gpt-4o-fine-tuning/.

OpenAI. 2024. Gpt-4 technical report.

Debjit Paul, Robert West, Antoine Bosselut, et al. 2024. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics*.

Angelika Romanou, Syrielle Montariol, Debjit Paul, et al. 2023. CRAB: Assessing the strength of causal relationships between real-world events. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Julia Rozanova, Marco Valentino, Lucas Cordeiro, et al. 2023. Interventional probing in high dimensions: An NLI case study. In *Findings of the Association for Computational Linguistics*.

Aalok Sathe and Joonsuk Park. 2021. Automatic fact-checking with document-level annotations using BERT and multiple instance learning. In *Proceedings of the Workshop on Fact Extraction and VERification*.

Neelesh Kumar Shukla, Sandeep Singh, Prabhat Kumar Prabhakar, et al. 2025. KULFi framework: Knowledge utilization for optimizing large language models for financial causal reasoning. In *Proceedings of the Joint Workshop of the Financial Technology and Natural Language Processing, the Financial Narrative Processing, and the Workshop on Large Language Models for Finance and Legal*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, et al. 2023. Llama: Open and efficient foundation language models.

Miles Turpin, Julian Michael, Ethan Perez, et al. 2023. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. In *Proceedings of the International Conference on Neural Information Processing Systems*.

Prasetya Utama, Joshua Bambrick, Nafise Moosavi, et al. 2022. Falsesum: Generating document-level NLI examples for recognizing factual inconsistency in summarization. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

Jiaan Wang, Yunlong Liang, Fandong Meng, et al. 2023. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*.

Zeyu Wang. 2024. CausalBench: A comprehensive benchmark for evaluating causal reasoning capabilities of large language models. In *Proceedings of the SIGHAN Workshop on Chinese Language Processing*.

Frank Wilcoxon. 1947. Probability tables for individual comparisons by ranking methods. *Biometrics*, 3(3):119–122.

Junda Wu, Tong Yu, Xiang Chen, et al. 2024. DeCoT: Debiasing chain-of-thought for knowledge-intensive tasks in large language models via causal intervention. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Jiuding Yang, Hui Liu, Weidong Guo, et al. 2024. Reassess summary factual inconsistency detection with large language model. In *Proceedings of the Workshop on Towards Knowledgeable Language Models*.

Fei Yu, Hongbo Zhang, Prayag Tiwari, et al. 2024. Natural language reasoning, a survey. *ACM Computing Surveys*, 56(12):1–39.

Han Yuan, Feng Xie, Marcus Eng Hock Ong, Yilin Ning, Marcel Lucas Chee, Seyed Ehsan Saffari, Hairil Rizal Abdullah, Benjamin Alan Goldstein, Bibhas Chakraborty, and Nan Liu. 2022. Autoscore-imbalance: An interpretable machine learning tool for development of clinical scores with rare events data. *Journal of Biomedical Informatics*, 129:104072.

Han Yuan, Li Zhang, and Zheng Ma. 2025. Exploring the reliability of self-explanation and its relationship with classification in language model-driven financial analysis. In *ICLR 2025 Workshop on Advances in Financial AI: Opportunities, Innovations and Responsible AI*.

Muru Zhang, Ofir Press, William Merrill, et al. 2024. How language model hallucinations can snowball. In *Proceedings of the International Conference on Machine Learning*.

Lianmin Zheng, Wei-Lin Chiang, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the International Conference on Neural Information Processing Systems*.

A PEFT results

Tables 3 and 4 show PEFT results for both encoders and LMs in detecting factual and causal issues, respectively. With the exception of Llama-3.2-3B and Gemma-2-9B in factuality and causality, LoRA consistently outperforms LLFT. Notably, these two exceptional models exhibit consistent behavior across both tasks, suggesting that LoRA struggles to identify more effective parameters than those in the final layer in some scenarios.

Model	Mode	F1	BA	AUPRC	AUROC
DeBERTa-v3-large	LoRA	0.81	0.91	0.87	0.98
	LLFT	0.19	0.76	0.32	0.85
BART-large	LoRA	0.79	0.89	0.87	0.98
	LLFT	0.45	0.79	0.59	0.93
RoBERTa-large	LoRA	0.85	0.94	0.89	0.99
	LLFT	0.26	0.79	0.40	0.88
Llama-3.2-3B	LoRA	0.11	0.64	0.25	0.69
	LLFT	0.38	0.82	0.28	0.71
Llama-3.1-8B	LoRA	0.52	0.86	0.42	0.93
	LLFT	0.15	0.66	0.19	0.76
Gemma-2-2B	LoRA	0.46	0.73	0.33	0.88
	LLFT	0.22	0.58	0.21	0.82
Gemma-2-9B	LoRA	0.35	0.75	0.14	0.71
	LLFT	0.53	0.79	0.49	0.82
Phi-3.5-mini	LoRA	0.37	0.83	0.51	0.92
	LLFT	0.16	0.64	0.27	0.66
Phi-3.5-MoE	LoRA	0.30	0.83	0.27	0.83
	LLFT	0.14	0.64	0.18	0.69

Table 3: Factuality detection results of PEFT models

Model	Mode	F1	BA	AUPRC	AUROC
DeBERTa-v3-large	LoRA	0.89	0.95	0.91	0.98
	LLFT	0.37	0.62	0.22	0.60
BART-large	LoRA	0.85	0.93	0.91	0.98
	LLFT	0.57	0.80	0.73	0.90
RoBERTa-large	LoRA	0.89	0.96	0.89	0.99
	LLFT	0.43	0.68	0.42	0.72
Llama-3.2-3B	LoRA	0.30	0.50	0.18	0.49
	LLFT	0.35	0.59	0.28	0.63
Llama-3.1-8B	LoRA	0.56	0.81	0.37	0.77
	LLFT	0.31	0.51	0.22	0.55
Gemma-2-2B	LoRA	0.07	0.49	0.27	0.70
	LLFT	0.04	0.48	0.22	0.61
Gemma-2-9B	LoRA	0.07	0.47	0.21	0.55
	LLFT	0.29	0.55	0.21	0.56
Phi-3.5-mini	LoRA	0.30	0.54	0.17	0.48
	LLFT	0.27	0.47	0.15	0.42
Phi-3.5-MoE	LoRA	0.45	0.68	0.22	0.58
	LLFT	0.32	0.56	0.19	0.55

Table 4: Causality detection results of PEFT models

B Position bias of LMs

Prior studies have shown that LMs exhibit position bias when making inferences involving swapped answer positions (Zheng et al., 2023). In our context, position bias refers to the effect of presenting prompts in the order of *"Yes"* or *"No"* versus *"No"* or *"Yes"*. To assess the position bias, we perform a chi-squared test on the decisions made by pre-trained models across all samples under two prompt orders. The null hypothesis posits no significant difference between the two variants, while the alternative hypothesis suggests a significant difference. The results indicate that, with the exception of Llama-3.2-3B in factuality detection, and Llama-3.1-8B in both factuality and causality detection, all other models exhibit p-values less than 0.01. This provides strong statistical evidence for the presence of position bias. It is also worth noting that the absence of bias in Llama is attributable to its limited capability, as it generates *"No"* for nearly all samples, resulting in no variation.

To address the position bias, the reported results using either average voting or veto voting across these two prompt variants. For the selection of voting methods, we adopt a heuristic approach: when a model's output is highly skewed toward an answer (i.e., 95% of responses favor one option), we apply veto voting to ensure that the minority is better represented and to encourage output diversity. For outputs that do not exceed this threshold, we use average voting to balance positional effects.

C Pairwise comparison

In addition to the 2 comparisons in the main text, we provide heatmaps of pairwise comparisons across the remaining 6 scenarios, defined by the combination of setting (pre-trained vs. FPFT), task (factuality vs. causality), and issue (true vs. false). A cell with a significant p-value indicates that the column model outperforms the row model.

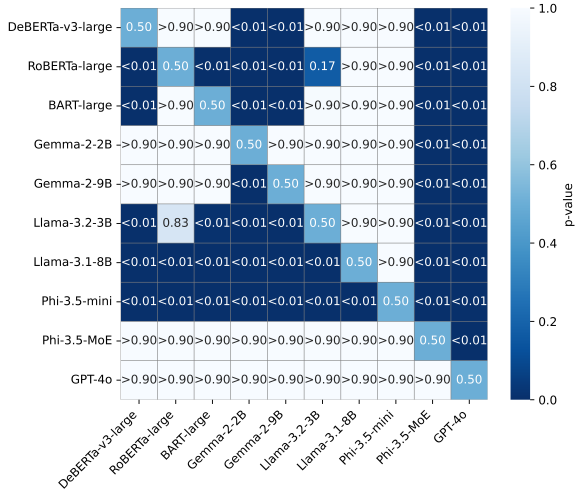

Figure 7: Pairwise comparison of factuality detection on correct reasoning points in the pre-trained setting

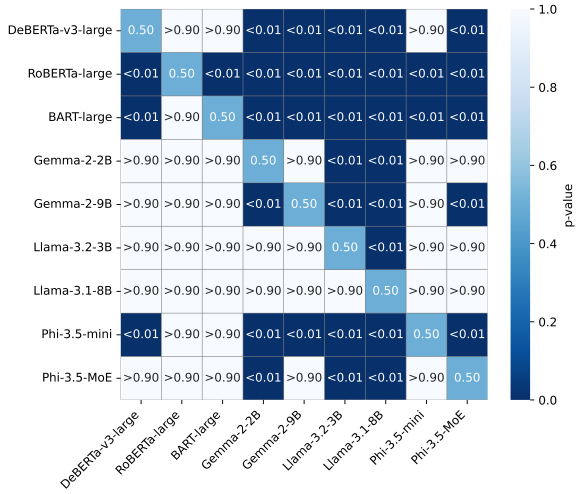

Figure 8: Pairwise comparison of factuality detection on correct reasoning points in the FPFT setting

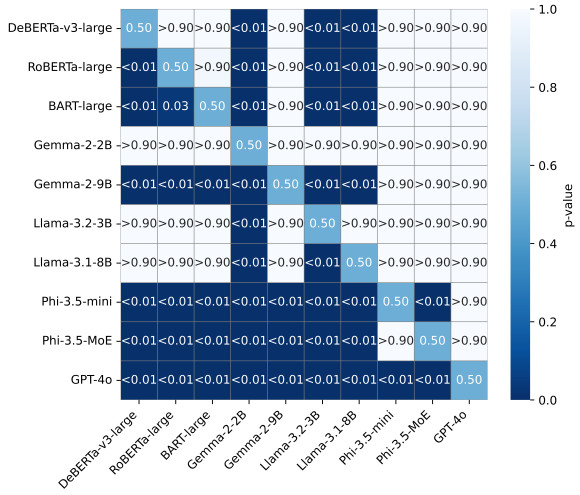

Figure 9: Pairwise comparison of causality detection on correct reasoning points in the pre-trained setting

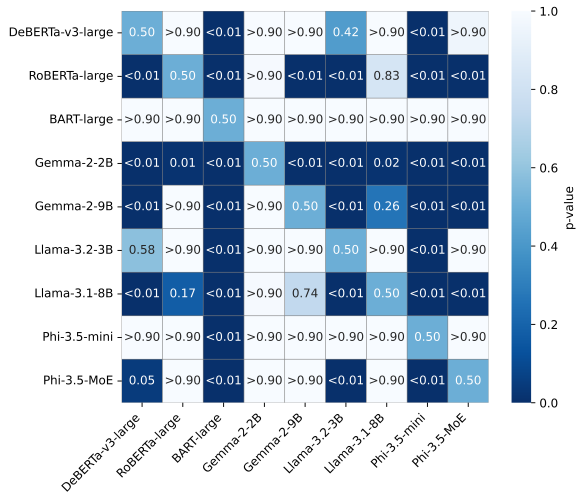

Figure 10: Pairwise comparison of causality detection on correct reasoning points in the FPFT setting

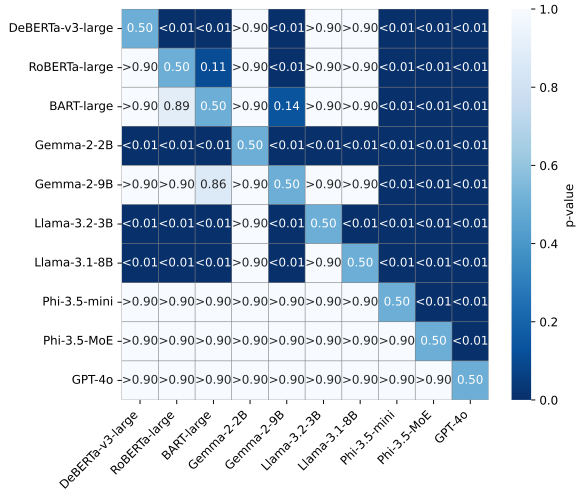

Figure 11: Pairwise comparison of causality detection on erroneous reasoning points in the pre-trained setting

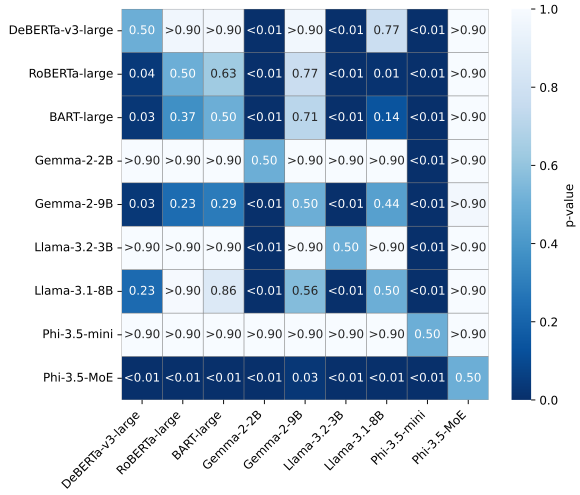

Figure 12: Pairwise comparison of causality detection on erroneous reasoning points in the FPFT setting