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

Singapore Decision Science Center of Excellence, Global Decision Science, American Express

{Yilin.Wu, Han.Yuan1, Li.Zhang1, Zheng.Ma2}@aexp.com

Introduction

- Language models (LMs) have revolutionized financial analysis by demonstrating expert-level self-reasoning versatility.
- LMs are known to hallucinate facts and generate non-causal reasoning paths, which pose risks of monetary losses.
- Detecting factual and causal errors in LMs' reasoning is essential for risk management and responsible application of LMs in finance.

Contributions

- Examine fine-grained labels of factuality and causality on LMs' reasoning.
- Demonstrate the effectiveness of the classic NLI as a detection paradigm for factual and causal errors, using encoders and LMs as backbones.
- 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.
- Demonstrate the necessity of fine-tuning, which not only enhances the detection of both backbones but also mitigates LMs' self-evaluation bias.

Method

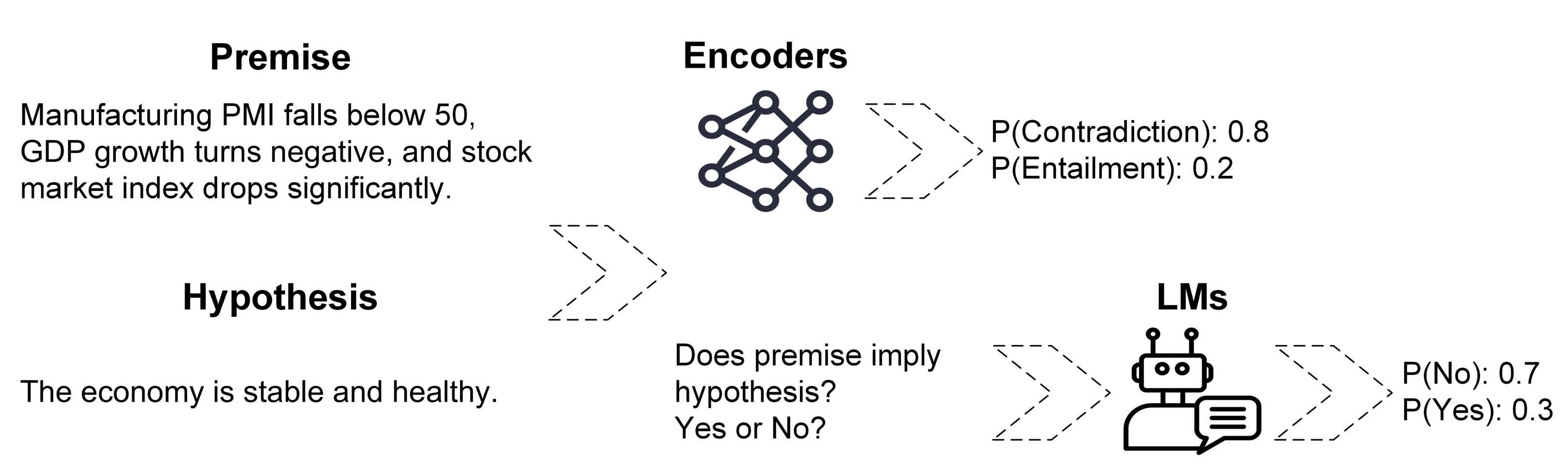

Figure 1: NLI can detect factual and causal errors in LMs' self-reasoning in finance.

- NLI takes a premise S_p and a hypothesis S_h as input. Then it outputs probabilities of entailment, neutrality, and contradiction.
- A LM response O_i comprises 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.
- For factuality detection, the premise S_p corresponds to the input information D_i and the hypothesis S_h is each reasoning statement $O_{i,k}$ $(k = 2, ..., K)$. For causality detection, the S_p is the reasoning statement $O_{i,k}$ $(k = 2, ..., K)$ and S_h is the classification outcome $O_{i,1}$.
- 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.

Results

- NLI is an effective paradigm for distinguishing sentences containing factual or causal errors. LMs can achieve performance inferior to that of encoders.
- Wilcoxon rank-sum test is used to demonstrate the effectiveness of NLI as a detection paradigm. A p-value < 0.05 indicates that NLI, powered by a certain backbone, has statistically significant distinguishability.

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 of pre-trained and FPFT models under NLI

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 of pre-trained and FPFT models under NLI

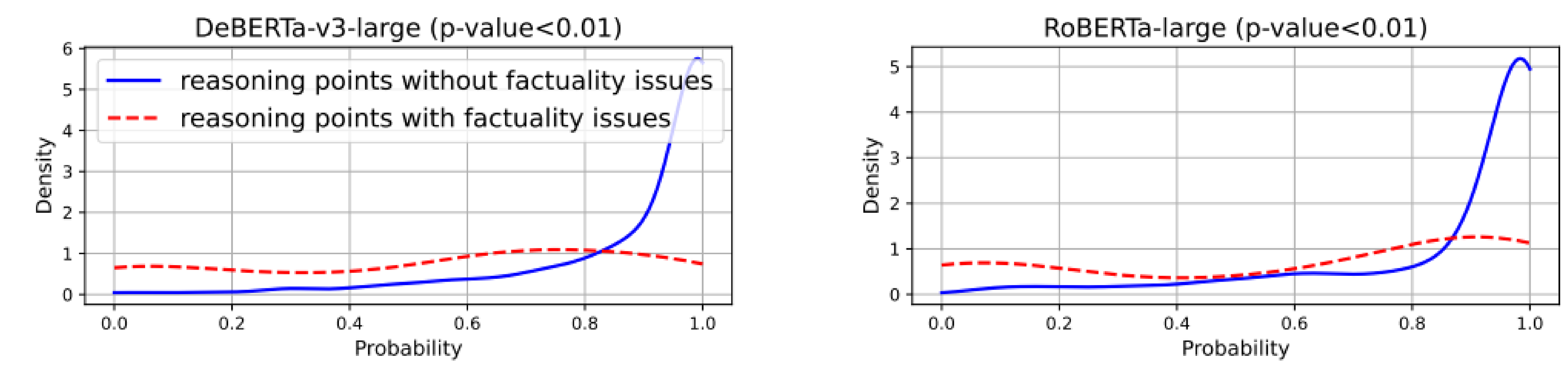

Figure 2: Entailment prob distributions for statements w/wo factual or causal errors.

- In Figure 3, smaller p-values indicate better discriminability; therefore, a positive value implies that the discriminability is better in the FPFT model, suggests that fine-tuning enhances the detection capability.
- In Figure 4, larger p-values indicate less self-evaluation bias; therefore, a positive value implies that the self-evaluation bias is lower in the FPFT model, suggests that fine-tuning mitigates LMs' self-evaluation bias.

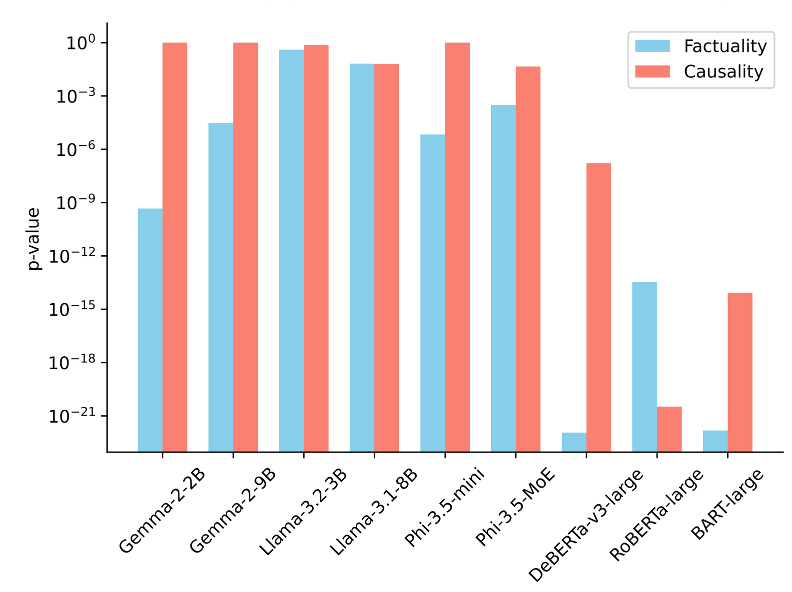

Figure 3: Detection capability comparison

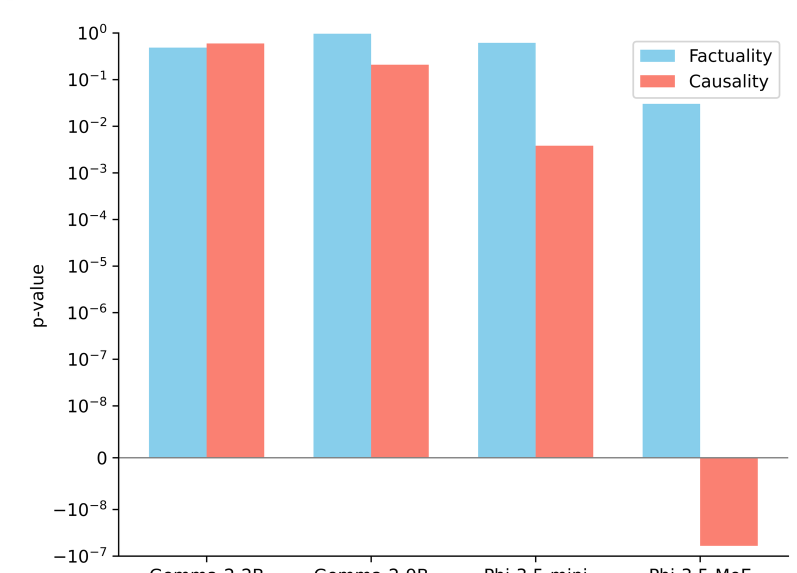

Figure 4: Self-evaluation bias comparison

Disclaimer

Reference

- AFaCTA: Assisting the annotation of factual claim detection with reliable LLM annotators. ACL. 2024.
- Enhancing self-consistency and performance of pre-trained language models through natural language inference. EMNLP. 2022.
- Evaluating the factual consistency of abstractive text summarization. EMNLP. 2020.