# Exploring the Reliability of Self-explanation and its Relationship with Classification in Language Model-driven Financial Analysis

Han Yuan  Li Zhang  Yilin Wu  Zheng Ma

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

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

## Introduction

- Language models (LMs) have exhibited exceptional versatility in reasoning and analyzing financial information through their proprietary information processing capabilities.
- Previous research has focused on evaluating classification performance while often overlooking explainability or implicitly assuming that improved explainability corresponds to higher classification accuracy, thereby using explanation refinement as a means to enhance classification.

## Contributions

- Conducted detailed annotations of self-explanations and illustrated the statistically significant relationship between the accuracy of classifications and the factuality or causality of self-explanations.
- Proved the feasibility of using pre-trained encoders for natural language inference to identify issues of factuality and causality in self-explanations.

## Experiments

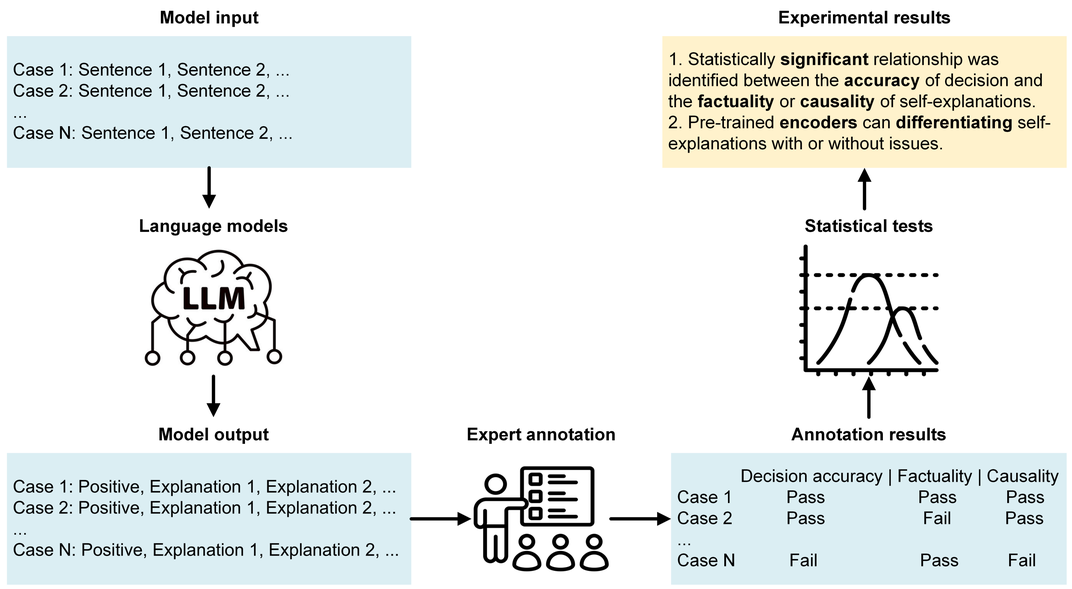


**Figure 1:** Schematic plot of our experimental pipeline

- Prompted LMs to generate both classification outcomes and the underlying explanations behind their classification decisions.
- Annotated the self-explanation sentences by authors.
- Conducted Chi-squared test to evaluate the independence of factuality or causality issues from the accuracy of model classifications.
- Leveraged pre-trained encoders to output probabilities of entailment and contradiction.
- Applied the Wilcoxon rank-sum test to validate the discriminability of pre-trained encoders in sentences with and without errors.

## Results

- Chi-squared tests revealed statistically significant associations between the accuracy of classifications and the factuality or causality of self-explanations ($P \leq 0.05$).
- Compared to factuality issues, causality issues were more prevalent in the self-explanations generated by LMs.

| LMs | Case type | Issue | Prevalence | P value |
|---|---|---|---|---|
| Llama-3.2-3B | Positive | factuality | 0.16 | 3.53e-05 |
| Llama-3.2-3B | Negative | factuality | 0.20 | 1.30e-15 |
| Phi-3.5-3.8B | Positive | factuality | 0.26 | 1.63e-14 |
| Phi-3.5-3.8B | Negative | factuality | 0.22 | 1.57e-17 |
| Gemma-2-2B | Positive | factuality | 0.32 | 1.84e-12 |
| Gemma-2-2B | Negative | factuality | 0.16 | 1.00e-20 |
| Llama-3.2-3B | Positive | causality | 0.80 | 6.52e-05 |
| Llama-3.2-3B | Negative | causality | 0.82 | 1.12e-15 |
| Phi-3.5-3.8B | Positive | causality | 0.52 | 4.48e-10 |
| Phi-3.5-3.8B | Negative | causality | 0.68 | 1.37e-13 |
| Gemma-2-2B | Positive | causality | 0.52 | 4.48e-10 |
| Gemma-2-2B | Negative | causality | 0.52 | 7.29e-11 |

**Table 1:** Impact analysis of factuality and causality of self-explanations on the accuracy of classifications

- Wilcoxon rank sum test showed that pre-trained encoders distinguished sentences with factual or causal errors from those without.

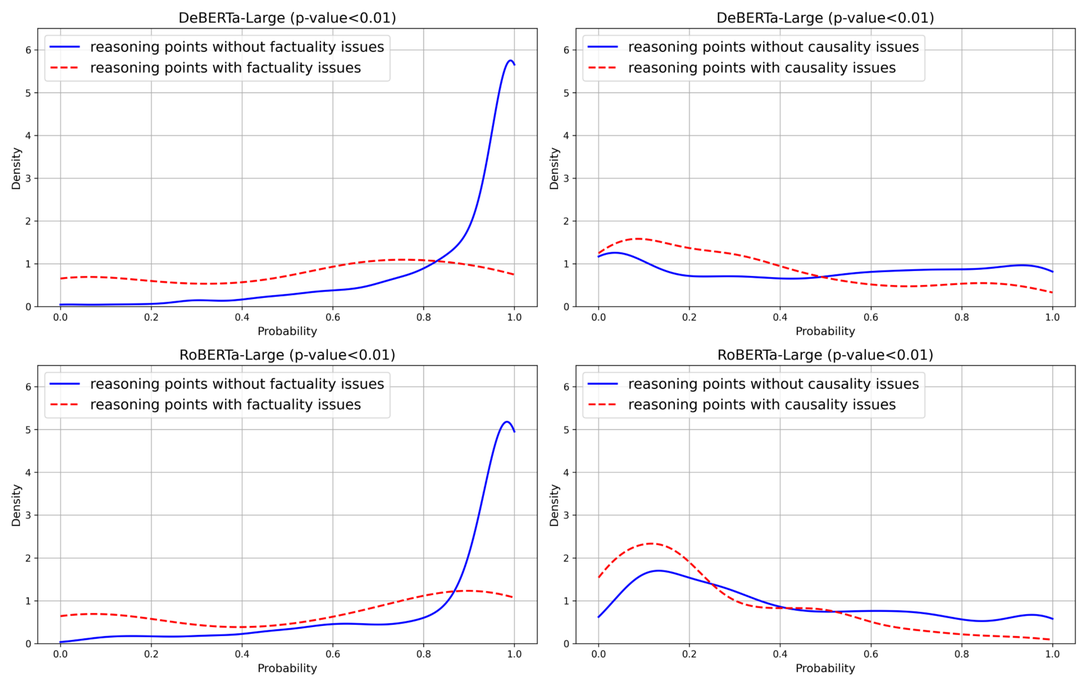


**Figure 2:** Probability distribution of NLI encoders

## Conclusions

- Beyond final classifications, LMs demonstrated the ability to provide human-interpretable explanations.
- While factuality and causality issues exist in self-explanations, recent advancements have mitigated factuality issues.
- With the continued evolution of reasoning capabilities, LMs hold promise for delivering highly accurate and self-explainable financial classifications.

## Disclaimer

This paper is intended solely for informational purposes and is not a product of or intended to constitute any business practice of American Express. The opinions, findings and conclusions of this paper are those of the authors alone.