Empowering Small Language Models with Factual Hallucination-Aware Reasoning for Financial Classification

Han Yuan Yilin Wu Li Zhang Zheng Ma

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

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

Motivation

- Small language models (SLMs) are increasingly used for financial classification due to their inference speed and local deployability.
- However, compared with large language models, SLMs are more prone to factual hallucinations in reasoning and exhibit weaker classification performance. This raises a natural question: Can mitigating factual hallucinations improve SLMs' financial classification?
- We propose a three-step pipeline named AAAI (Association identification, Automated detection, and Adaptive Inference).
- Compared with prior studies on model reflection, our work introduces statistical analyses to quantify the relationship between erroneous reasoning and misclassifications and to validate the discriminative power of automated detectors in the context of SLMs for finance.

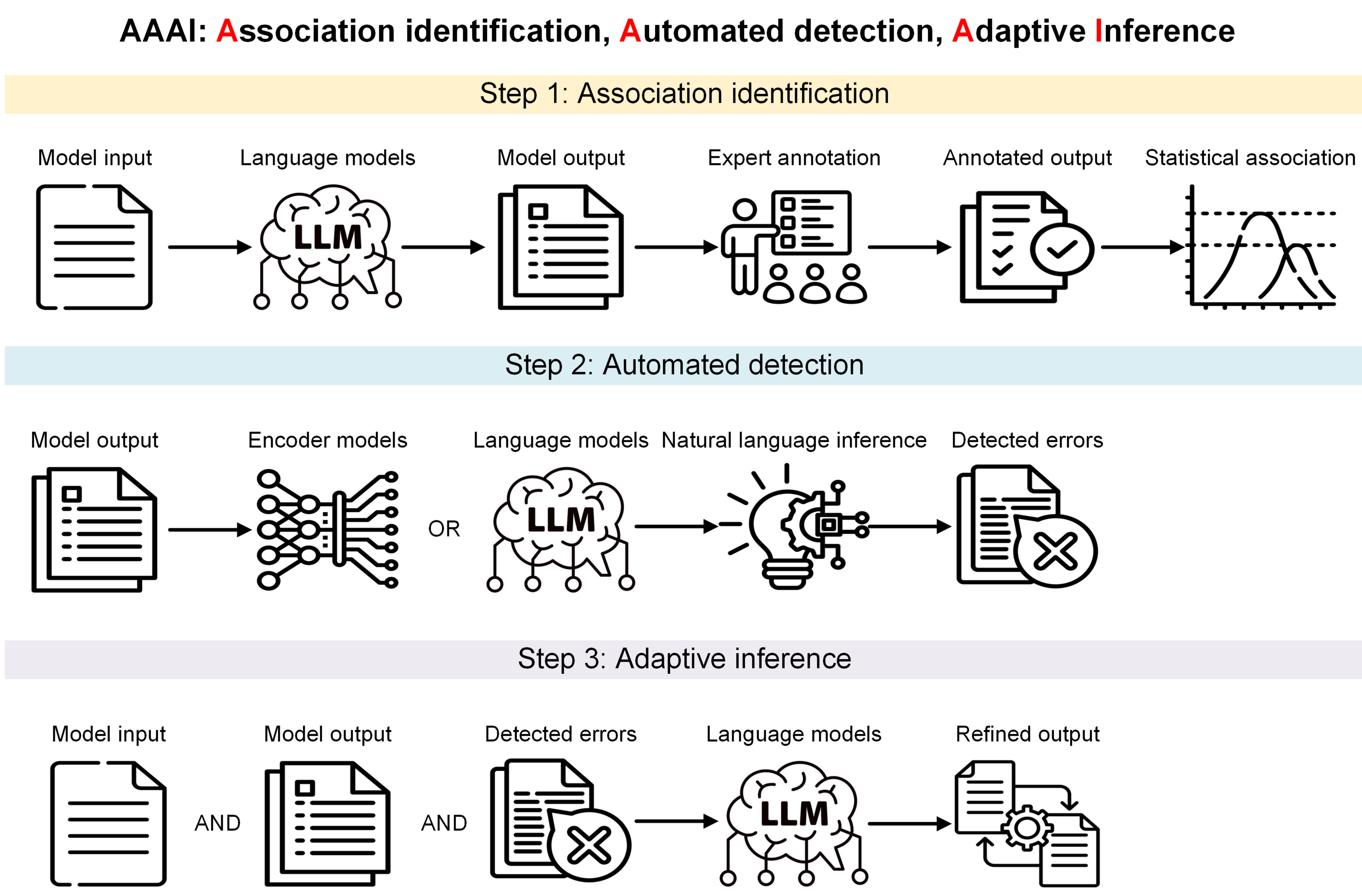

Figure 1: The pipeline for factual error-aware reasoning

Association identification

- Pearson correlation coefficients show the positive relationship between factual hallucinations and misclassifications across SLMs.
- Positive risk differences demonstrate that the risk of misclassification is higher in cases with factual errors than in those without across SLMs.

Automated detection

- Encoder-based architectures of DeBERTa-v3-large, RoBERTa-large, and BART-large are adopted as verifiers for factual errors in SLMs' reasoning.
- Wilcoxon rank-sum test is used to validate verifiers' discriminability. Except for RoBERTa-large on Phi, all p-values are below 0.01.

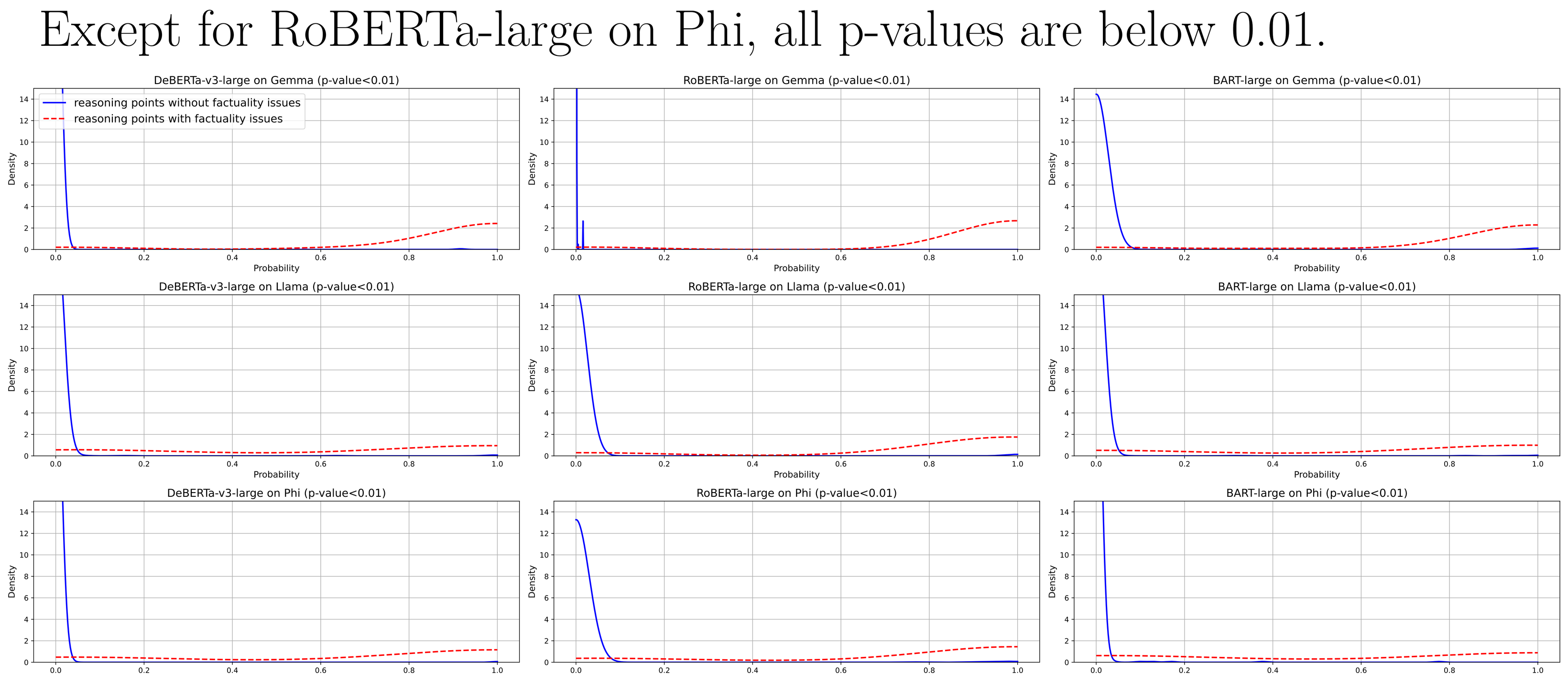

Figure 2: Probability density distribution of verifiers on reasoning w/wo factual errors

Adaptive inference

- Factual hallucinations is incorporated in the SLMs' reasoning, detected by diverse methods, as feedback to prompt SLMs to refine answers through a tandem round of hallucination-aware reasoning.
- The importance of feedback quality is underscored for adaptive inference of SLMs. Oracle feedback from human experts consistently enhances, or at least does not reduce, SLMs' performance.
- Compared with self-reflection, verifiers yield better performance in Llama and Gemma, highlighting the caution against overreliance on LMs.
- Self-reflection improves Gemma's performance, demonstrating the potential of SLMs to correct their own generations without external feedback.
- Phi exhibits the lowest steerability (the likelihood of adjusting its output behavior in response to external instructions), as feedback from either sources does not induce any change from its initial decision.

SLMs	Verifiers	Mode	AUPRC↑	BA↑
Llama	DeBERTa	Pre-trained	34.04	72.66
		FPFT	82.62	80.69
	RoBERTa	Pre-trained	55.71	74.91
		FPFT	76.33	92.39
	BART	Pre-trained	59.72	78.36
		FPFT	76.12	83.07
Gemma	DeBERTa	Pre-trained	46.44	69.98
		FPFT	96.97	96.05
	RoBERTa	Pre-trained	25.56	59.84
		FPFT	100.00	96.15
	BART	Pre-trained	29.19	63.36
		FPFT	90.66	93.80
Phi	DeBERTa	Pre-trained	26.82	58.63
		FPFT	91.51	83.90
	RoBERTa	Pre-trained	14.78	53.06
		FPFT	87.29	87.39
	BART	Pre-trained	22.20	56.86
		FPFT	73.61	77.90

Table 1: Verifiers' performance on SLMs' reasoning w/wo factual hallucinations

SLMs	Feedback	F1 score↑	Weighted cost↓
Llama	No feedback	76.42	41
	Oracle	80.67	31
	Verifier-DeBERTa	79.66	36
	Verifier-RoBERTa	80.67	31
	Verifier-BART	78.99	37
	Self-reflection	76.42	41
Gemma	No feedback	67.11	49
	Oracle	68.49	46
	Verifier-DeBERTa	68.97	45
	Verifier-RoBERTa	68.97	45
	Verifier-BART	69.44	44
	Self-reflection	67.57	48
Phi	No feedback	67.11	49
	Oracle	67.11	49
	Verifier-DeBERTa	67.11	49
	Verifier-RoBERTa	67.11	49
	Verifier-BART	67.11	49
	Self-reflection	67.11	49

Table 2: SLMs' performance w/wo factual hallucination-aware reasoning

Additional rounds

- Additional rounds of self-reflection and adaptive inference do not always improve SLMs' performance compared with the initial generation without feedback. SLMs overcriticize prior reasoning when its quality is high, but provide constructive criticism when its quality is low.

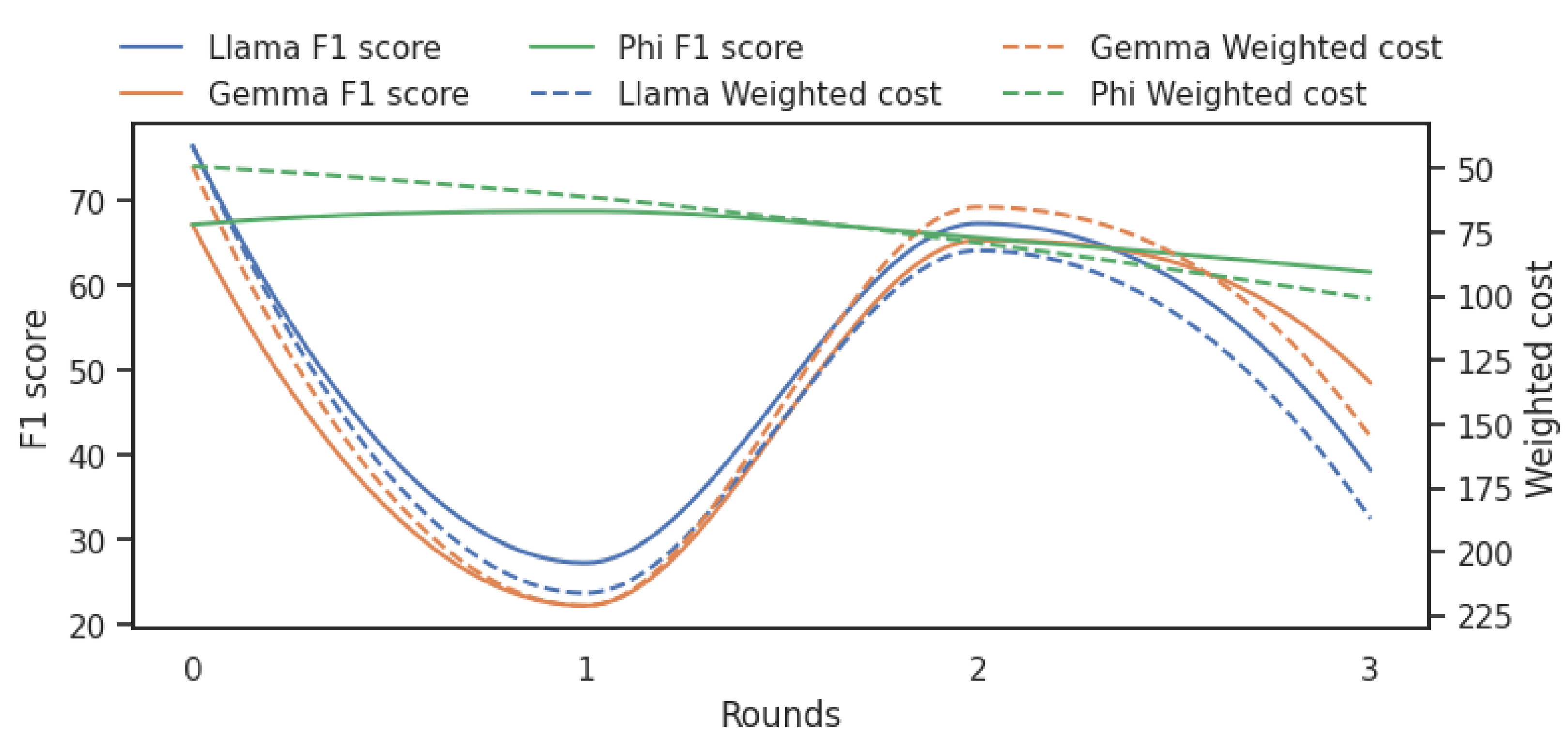

Figure 3: Performance comparison of SLMs across different reasoning rounds

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.