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

Han Yuan[*]**, Yilin Wu**[*]**, Li Zhang, Zheng Ma**[†]

Global Decision Science, American Express

{Han.Yuan1, yilin.wu, Li.Zhang1, Zheng.Ma2}@aexp.com

Abstract

Small language models (SLMs) are increasingly used for financial classification due to their fast inference 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?* To address this, we propose a three-step pipeline named AAAI (**A**ssociation Identification, **A**utomated Detection, and **A**daptive **I**nference). Experiments on three representative SLMs reveal that: (1) factual hallucinations are positively correlated with misclassifications; (2) encoder-based verifiers effectively detect factual hallucinations; and (3) incorporating feedback on factual errors enables SLMs' adaptive inference that enhances classification performance. We hope this pipeline contributes to trustworthy and effective applications of SLMs in finance.

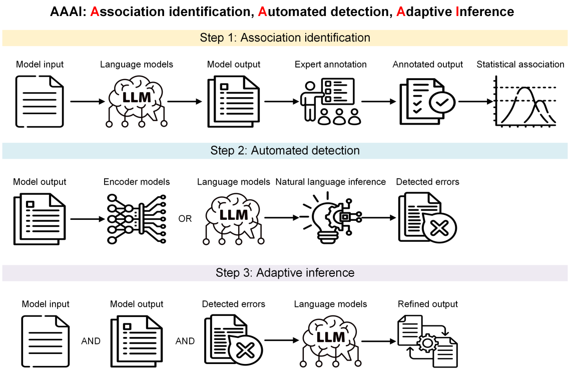

Figure 1: The pipeline for factual error-aware reasoning

Introduction

Language models (LMs) are increasingly being deployed for financial classification (Guo, Xu, and Yang 2023; Li et al. 2023b; Chen et al. 2024; Hu et al. 2025). Two main development paths have emerged: one focuses on large language models (LLMs) with superior performance, while the other targets small language models (SLMs) suitable for local deployability (Cheng et al. 2024). Although SLMs offer advantages in fast inference and privacy protection, they are prone to factual hallucinations (Li et al. 2023a). A reasoning path containing factual errors undermines both the trustworthiness of an SLM's output and the quality of downstream classification (Lin et al. 2024). Therefore, enabling SLMs to recognize factual hallucinations in their reasoning potentially enhances the quality of their overall generation.

To demonstrate the practical applicability of this approach in finance, we implement a three-step analytical pipeline, abbreviated as AAAI (**A**ssociation identification, **A**utomated detection, and **A**daptive **I**nference), and Figure 1 visualizes each step of this pipeline. First, statistical analysis is conducted to identify the positive association between factual hallucinations in SLMs' reasoning and the accuracy of classifications derived from that reasoning in the initial inference round. This step provides a rationale for improving classification by guiding SLMs to recognize and correct factual errors in their adaptive inference. Second, various methods are used to automatically detect factual errors in the reasoning of the initial inference. This step includes statistical analysis to demonstrate the discriminative ability of the detection models, supporting the scalability of factual error identification. Third, the detected errors are used as feedback for SLMs, prompting adaptive refinement of their inference. Quantitative metrics show that feedback from some methods enhances SLMs' financial classification, while feedback from others degrades it, highlighting the importance of feedback quality in factual error-aware reasoning (Huang et al. 2024).

We emphasize that, unlike studies that use classification correctness to guide SLMs (Shinn et al. 2024; Kim, Baldi, and McAleer 2023), all feedback in this work focuses on factual issues in LMs' reasoning, which aligns with real-world scenarios where users lack access to the ground truth (Lightman et al. 2024; Wang et al. 2024). If the answer were already known, collaboration with LMs would be unnecessary (Huang et al. 2024). However, users can readily identify errors in LMs' reasoning, particularly factual hallucinations that contradict the given context in self-contained analyses where external information is not required (Uesato et al. 2022; Chakraborty, Ornik, and Driggs-Campbell 2025).

[*] These authors contributed equally.

[†] Correspondence, Singapore Decision Science Center of Excellence, American Express, 1 Marina Boulevard, 018989, Singapore.

Related work

Reasoning and classification

Recent advancements such as DeepSeek (Liu et al. 2024a) show the potential of reasoning to enhance model classification performance without explicit user instructions. Although overall performance gains have been observed, some studies further explore the role of reasoning in enhancing generation by LMs (Zhao et al. 2023; Yuan, Zhang, and Ma 2025). Lampinen et al. (2022) examine the effect of providing a few in-context examples to LMs' prompts and concluded that explanations improved model performance in general domains. Ye and Durrett (2024) investigate the use of triplets comprising a question, classification, and explanation in few-shot examples, proving LMs tend to generate nonfactual explanations when making wrong predictions. Turpin et al. (2023) also reveal a close relationship between reasoning and decision, showing that even when the decision is incorrect, LMs tend to adjust their explanations to justify it. Our study advances this line by statistically quantifying the relationship between reasoning and classification, providing evidence that factually erroneous reasoning is correlated with misclassification.

Factuality verification

Verifying factual statements and detecting hallucinations are essential for ensuring the trustworthiness and safety of LMs (Tang et al. 2024; Chen et al. 2025). For LMs' outputs that include both reasoning and classification, hallucination detection can be broadly divided into two types: outcome detection and process detection (Welleck et al. 2023; Lightman et al. 2024). As discussed previously, outcome detection assumes prior knowledge of whether the outcome is correct, which is often impractical in real-world settings. Former studies have trained process verifiers and demonstrated their effectiveness as feedback to LMs. Lightman et al. (2024) fine-tune GPT-4 to predict the correctness of each reasoning step in mathematical problems. Wang et al. (2024) further extend it by using relatively smaller LMs to provide feedback on each reasoning step and demonstrating that model-generated feedback can guide SLMs to produce better outputs through proximal policy optimization. In this study, we also focus on process verification, specifically examining factual hallucinations within each reasoning step. For backbone models, we adopt transformer encoders for their efficiency and effectiveness as step verifiers of LMs' reasoning (Li et al. 2023c; Tang et al. 2023; Sun et al. 2024).

Adaptive inference

Adaptive inference has gained popularity in recent research as a means to enhance LMs' performance through feedback from diverse sources. The previous section reviewed studies that use externally fine-tuned models to verify the factuality of LMs' reasoning, representing one source of feedback. Beyond this, three common sources are knowledge databases, human experts, and other LMs. Feedback from experts serves as the ground truth for assessing the factuality of LMs' reasoning and represents human-in-the-loop practices (Yuan et al. 2024). Knowledge databases are excluded because we focus on self-contained scenarios where all factual context required for correct inference is provided. Feedback from LMs can be categorized into two types: LLM-as-a-judge (Jang, Lee, and Kim 2022; Koutcheme et al. 2024; Ye et al. 2025) and self-reflection (Dou et al. 2024; Gupta et al. 2024; Liu et al. 2024b; Zhao et al. 2024; Li et al. 2024). Both approaches use LMs to identify errors and guide LMs to iteratively refine earlier outputs. The key difference is that LLM-as-a-judge typically relies on other, often more capable, models, whereas self-reflection uses the same one for initial inference, feedback, and iterative improvement. Compared with general feedback investigated in most former studies, our work focuses specifically on factual hallucinations. Some research also examines adaptive inference based on factuality of LMs' initial outputs. For example, Ji et al. (2023) employ a customized scorer to assess knowledge generated by LMs for open-ended medical tasks. In contrast, our study targets closed-end financial tasks.

Experiments

To illustrate the potential for improving SLMs' financial classification through reflection on factual hallucinations, we implement a three-step AAAI pipeline[1] (Figure 1).

Dataset

We use the German credit financial classification dataset (Hofmann 1994), a widely recognized benchmark in financial natural language processing (NLP). In this context, $L = 1$ indicates a good profile and $L = 0$ represents a bad profile. Prior studies (Xie et al. 2023; Bhatia et al. 2024) on this dataset overlook the critical role of data processing in enhancing the signal-to-noise ratio and revealing the true capabilities of SLMs. Specifically, the original dataset includes outdated information and pre-existing bias (Zehlike et al. 2017) that pose challenges for SLMs. For instance, certain features are denominated in Deutsche Marks, a currency that has been obsolete for over two decades. Also, SLMs often exhibit limited sensitivity to numeric reasoning (Mishra et al. 2022). To address these issues, we exclude features misaligned with contemporary SLMs development contexts and converted numeric features into percentile representations. All processing steps were conducted ad hoc and did not involve any operations related to the labels, ensuring that performance was not affected by information leakage. Also, we conduct all experiments using all minority cases paired with an equal number of majority cases to avoid the adverse impact of data imbalance on the analyses.

We utilize three SLMs to generate structured content, containing both reasoning and decision, as the initial responses (Yuan et al. 2025): Meta's Llama-3.2-3B (Touvron et al. 2023), Google's Gemma-2-2B (Mesnard et al. 2024), and Microsoft's Phi-3.5-3.8B (Abdin et al. 2024). Following Madaan et al. (2023), given an input X, prompt P_{gen} and model M, an initial generation Y is obtained: $Y = M(P_{gen} \bigoplus X)$. Here, P_{gen} is a task-specific prompt for

[1] https://github.com/Han-Yuan-Med/Factuality-Aware-Reasoning

SLMs	Coefficient	Risk difference
Llama	4.47e-2	8.75
Gemma	1.99e-1	1.47
Phi	5.81e-2	1.35

Table 1: Positive relationship between factually hallucinated reasoning and misclassifications measured by Pearson correlation and risk difference

an initial generation: "Assess the creditworthiness of a customer using the following attributes for financial status. Respond with the final decision of either 'good credit' or 'bad credit' in the first line. Respond with the reasoning on the final decision in the second line. And the attributes are as follows: $\{X\}$. Response: ". $\bigoplus$ stands for concatenation, and Y contains two parts of a classification decision Y^{cls} and I supporting reasoning points Y_i^{rsn} $(i = 1, 2, ..., I)$.

For classification metrics, we adopt the standard metrics of F1 score by comparing Y^{cls} with L. In addition, financial classification prioritizes weighted costs, emphasizing the greater consequence of false positive to false negative and a lower cost indicates superior performance. As specified in the original dataset documentation (Hofmann 1994), the cost associated with a false negative is quantified as 5, while that of a false positive is 1. Therefore, the weighted cost is $5 \times Num(Y^{cls} = 0, L = 1) + 1 \times Num(Y^{cls} = 1, L = 1)$. We identify a consistent improvement in weighted cost across all models when using the processed dataset. For F1 score, a substantial enhancement is achieved with Llama, while performance remained comparable for the other two.

Association identification

After evaluating financial classification performance, we conduct an association analysis to examine the relationship between factual hallucinations and misclassifications using Pearson correlation (Pearson 1895). Basically, for each reasoning point Y_i^{rsn}, we annotate whether it contains factual hallucinations, where $H_i^{rsn} = 1$ denotes that Y_i^{rsn} contains factual hallucinations. If any Y_i^{rsn} contains a factual hallucination, the overall reasoning Y^{rsn} is considered factually erroneous in terms of reasoning ($H^{rsn} = 1$). We then examine the correlation between the subgroup of $H^{rsn} = 1$ and the subgroup of $Y^{cls} \neq L$. A positive correlation coefficient indicates that reasoning containing factual errors is more likely to occur with incorrect decisions. Also, we calculate the false decision risk difference: $Prob(Y^{cls} \neq L | H^{rsn} = 1) - Prob(Y^{cls} \neq L | H^{rsn} = 0)$, where $|$ stands for the conditional probability. Similar to the Pearson correlation coefficient, a positive risk difference demonstrates that the risk of misclassification is higher in cases with factual errors than in those without. Table 1 presents the correlation results and risk differences, demonstrating the positive relationship between factual hallucinations and incorrect decisions, thereby supporting our subsequent experiments of improving SLMs' classification by mitigating factual errors.

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 2: Performance of verifiers in classifying reasoning points with or without factual hallucinations

Automated detection

Next, we adopt three encoder-based architectures of DeBERTa-v3-large (He et al. 2021), RoBERTa-large (Liu et al. 2019), and BART-large (Lewis et al. 2020) as verifiers V to predict probability of factual errors $Prob_i^v = V(X, Y_i^{rsn})$ in reasoning steps Y_i^{rsn} produced by SLMs (Ji et al. 2023; Wu et al. 2025). To prevent data leakage, full-parameter fine-tuning (FPFT) of V is conducted using a three-fold split, ensuring that $Prob_i^v$ is collected under the fold where X and Y_i^{rsn} serves as test data. Table 2 shows verifiers' performance by comparing $Prob_i^v$ with H_i^{rsn}. Due to the limited sample size, we do not employ an independent validation set and instead use a fixed threshold: If $Prob_i^v \geq 0.5$, the prediction $Pred_i^v = 1$ means that Y_i^{rsn} contains factual hallucinations. Also, given the class imbalance, where reasoning points with factual hallucinations are fewer than those without, we evaluate performance using area under the precision-recall curve (AUPRC) and balanced accuracy (BA). In certain cases, a model may achieve a perfect AUPRC score of 1, while the BA remains below 1. In addition to aggregated metrics such as AUPRC, Figure 2 illustrates the probability density distributions of fine-tuned verifiers, showing that they assign substantially higher probabilities to reasoning points with factual errors ($H_i^{rsn} = 1$) than to those without errors ($H_i^{rsn} = 0$). We further validate this discriminability using Wilcoxon rank-sum tests (Wilcoxon 1947), with p-values reported at the top of each subplot. Except for RoBERTa-large on Phi, all p-values are below 0.01, confirming the verifiers' identification ability.

Adaptive inference

Following Kim, Baldi, and McAleer (2023) and Huang et al. (2024), we incorporate factual hallucinations in the SLMs' reasoning, detected by diverse methods, as feedback F_i^{rsn} to prompt SLMs to refine answers through a tandem round of hallucination-aware reasoning (Madaan et al. 2023).

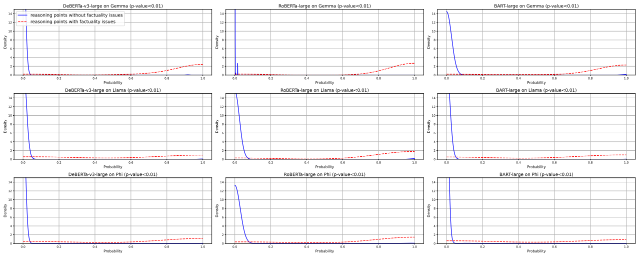

Figure 2: Probability density distribution of fine-tuned verifiers distinguishing reasoning points with and without factual errors

Specifically, we employ three strategies to provide feedback and facilitate a tandem round of classification: oracle, verifier, and self-reflection. The oracle provides gold-standard feedback H_i^{rsn} from human experts who annotate reasoning steps containing factual errors in the SLMs' initial outputs. The verifier uses the three fine-tuned encoders described in the previous section and outputs the feedback of $Pred_i^v$. In contrast, self-reflection (Renze and Guven 2024; Du et al. 2024) relies on the same SLMs to detect factual issues in their own reasoning points without leveraging external feedback (Huang et al. 2024). Mathematically, the feedback $Pred_i^m$ for the reasoning point Y_i^{rsn} is $M(P_{fed} \bigoplus X \bigoplus Y_i^{rsn})$, where P_{fed} refers to the feedback instruction: "$\{X\}$. Question: does this imply $\{Y_i^{rsn}\}$? Yes or No? Response: ". Based on diverse feedback F_i^{rsn} of H_i^{rsn}, $Pred_i^v$, and $Pred_i^m$, the refined generation Y' is $M(P_{ref} \bigoplus X \bigoplus Y \bigoplus F_i^{rsn})$, where P_{ref} denotes the regeneration instruction based on the feedback: "Your previous response contains the following factual errors: $\{F_i^{rsn}\}$. These errors does not match the given attributes. Based on the feedback, improve your decision and reasoning. Response: ". Consistent with the initial generation, Y' contains both classification decision $Y^{cls\prime}$ and supporting reasoning, and we evaluate the refined generation $Y^{cls\prime}$ based on L.

Table 3 compares classification performance across various SLMs and feedback sources. Consistent with Huang et al. (2024) and Madaan et al. (2023), our experiments underscore the importance of feedback quality for adaptive inference of SLMs, which can either improve or decline after factual hallucination-aware reasoning. Oracle feedback from human experts consistently enhances, or at least does not reduce, SLMs' performance. Compared with self-reflection, verifiers yield better classification performance in Llama and Gemma, highlighting the need for caution against overreliance on LMs in NLP tasks (Tang et al. 2024).

Also, we observe that self-reflection improves Gemma's classification performance, demonstrating the potential of SLMs to correct their own generations without external feedback (Wu et al. 2024). Gemma achieves even better classification with feedback from verifiers than from the oracle, indicating that SLMs can produce correct classifications even when feedback is inaccurate, which is reported in Madaan et al. (2023). Among these cases, some contain no factual hallucinations in the reasoning, yet classifications are incorrect. A promising study is to explore how multiple-perspective feedback beyond factuality enhances SLMs' reasoning and classification (Yan et al. 2024).

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 3: Performance comparison of SLMs with and without factual hallucination-aware reasoning

Moreover, we notice varying levels of steerability across SLMs, where steerability refers to a model's likelihood of adjusting its output behavior in response to external instructions such as feedback (Miehling et al. 2025). Table 3 shows that Phi exhibits the lowest steerability, as feedback from oracle, verifiers, or self-reflection does not induce any change from its initial classification, a behavior also observed in Vicuna (Madaan et al. 2023). Although low steerability limits adjustments to LMs' inherent beliefs, it becomes a desirable property when invalid feedback is given during adaptive inference, as robust LMs should defend their reasoning and decision rather than being easily misled (Wang et al. 2023).

Supplementary analyses

In addition to experiments in the former sections, we implement supplementary analyses under alternative settings.

First, we fine-tune SLMs to detect factual errors in their own reasoning and use this feedback to trigger error-aware reasoning. To prevent bias in SLMs' generation ability, the fine-tuned models are solely used for feedback generation, while the foundation SLMs perform the hallucination-aware reasoning and classification. The resulting classification performance is shown in Table 4. For Llama, feedback from the fine-tuned model improves classification performance, whereas for the other two SLMs, fine-tuning has marginal effect. Also, compared with feedback from encoder-based verifiers, feedback from either fine-tuned or foundation SLMs does not yield superior performance, indicating that transformer encoders are effective and efficient for initiating SLMs' factual hallucination-aware reasoning.

Second, our former experiments adopt a granularity at reasoning point level for self-reflection to maintain comparability with the verifiers. Previous studies (Kim, Baldi,

SLMs	Feedback	F1 score↑	Weighted cost↓
Llama	Fine-tuned	78.33	38
	Foundation	76.42	41
Gemma	Fine-tuned	67.57	48
	Foundation	67.57	48
Phi	Fine-tuned	67.11	49
	Foundation	67.11	49

Table 4: Performance comparison of SLMs with feedback from different SLMs' versions

SLMs	Granularity	F1 score↑	Weighted cost↓
Llama	Entire content	27.27	216
	Single point	76.42	41
Gemma	Entire content	22.22	221
	Single point	67.57	48
Phi	Entire content	68.70	61
	Single point	67.11	49

Table 5: Performance comparison of SLMs with factual reasoning at different granularities

and McAleer 2023; Huang et al. 2024) mainly use a coarser granularity, requiring models to reflect on the entire reasoning content containing multiple points. Table 5 compares the performance of these two granularities and their impact on SLMs' classification performance. For Llama and Gemma, reasoning at the single point granularity yields better performance across all metrics, likely due to their superior capability on short-context tasks. For Phi, using the entire reasoning content achieves higher F1 score. A possible explanation for this phenomenon is that Phi exhibits lower steerability than the other two SLMs. Since single point granularity provides weaker instructional signals than entire content granularity, Phi's behavior adaptation is consequently less pronounced.

Third, we explore multiple rounds of self-reflection and adaptive inference. Due to constraints in annotation budget, SLMs' input token length, and GPU memory, we leverage granularity at the entire content level and retain only the latest response and the original context for each additional round of inference. Figure 3 shows the impact of additional inference rounds on SLMs' classification performance. Consistent with findings from Huang et al. (2024) in general domain, additional rounds of adaptive inference do not always improve SLMs' performance compared with the initial generation without feedback in financial contexts. Another notable observation is that when a round of hallucination-aware reasoning yields weak performance, the next round often restores it, whereas when a round achieves strong performance, the following round frequently impairs it. Based on the identified positive association between reasoning and classification, we hypothesize that current LMs tend to over-criticize prior reasoning when its quality is high, but provide constructive criticism when its quality is low. The drastic fluctuation across self-reflection rounds presents an opportunity for future foundation model development.

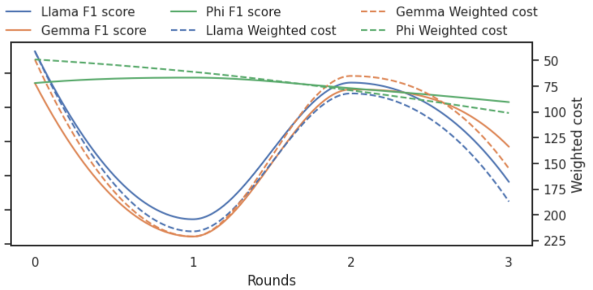

Figure 3: Performance comparison of SLMs across different rounds of reasoning at the entire content granularity

Conclusion and discussion

We present a three-step approach that allows SLMs to enhance their financial classification by realizing factual errors in their reasoning paths. 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. Furthermore, we highlight the importance of pinpointing specific erroneous reasoning steps, which can provide valuable annotation guidance for future SLMs' development (Lightman et al. 2024).

The positive relationship provides an empirical basis for developing a proxy confidence metric for LMs' classification, such as the proportion of factual errors in the reasoning path. In real-world settings, for example classifying the hawkish or dovish stance in Federal Open Market Committee speeches, even experts often struggle to make accurate decisions. As a result, they cannot always judge whether LMs' classifications are trustworthy, whereas hallucinations of reasoning are easier to evaluate. A proxy confidence value enables users to make more informed decisions about adopting LMs' suggestions. We hope that our study can drive further research on LMs' adaptive inference in finance.

Limitation

This work presents a three-step pipeline for studying factual hallucinations in SLMs' reasoning for financial classification and demonstrates the potential to improve SLMs' classification by incorporating factuality into the reasoning. Due to annotation constraints, the experiments are based on 50 positive and 50 negative cases from a public dataset. To validate the generalizability of our findings, future work should include more tasks and SLMs. Beyond the experimental aspect, the technical implementation can also be enhanced. First, the current self-reflection feedback relies on SLMs' zero-shot capability, while providing sufficient few-shot examples may improve their capability. Second, the current work only evaluates factual hallucinations in the initial generation. Although adaptive inference enhances classification, we do not evaluate its impact on reasoning. Further annotation can confirm whether it facilitates self-corrected generation of reasoning sequences (Welleck et al. 2023).

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

Abdin, M.; Aneja, J.; Awadalla, H.; et al. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv:2404.14219*.

Bhatia, G.; Nagoudi, E. M. B.; Cavusoglu, H.; et al. 2024. FinTral: A Family of GPT-4 Level Multimodal Financial Large Language Models. In *Findings of the Association for Computational Linguistics*.

Chakraborty, N.; Ornik, M.; and Driggs-Campbell, K. 2025. Hallucination detection in foundation models for decision-making: A flexible definition and review of the state of the art. *ACM Computing Surveys*, 57(7).

Chen, J.; Zhou, P.; Hua, Y.; et al. 2024. FinTextQA: A Dataset for Long-form Financial Question Answering. In *The Annual Meeting of the Association for Computational Linguistics*.

Chen, Y.; Liu, H.; Liu, Y.; Xie, J.; Yang, R.; Yuan, H.; Fu, Y.; Zhou, P. Y.; Chen, Q.; Caverlee, J.; and Li, I. 2025. GraphCheck: Breaking Long-Term Text Barriers with Extracted Knowledge Graph-Powered Fact-Checking. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Cheng, X.; Li, J.; Zhao, X.; et al. 2024. Small Agent Can Also Rock! Empowering Small Language Models as Hallucination Detector. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *The Conference on Empirical Methods in Natural Language Processing*.

Dou, Z.-Y.; Yang, C.-F.; Wu, X.; et al. 2024. Re-ReST: Reflection-Reinforced Self-Training for Language Agents. In *The Conference on Empirical Methods in Natural Language Processing*.

Du, Y.; Li, S.; Torralba, A.; et al. 2024. Improving factuality and reasoning in language models through multiagent debate. In *The International Conference on Machine Learning*.

Guo, Y.; Xu, Z.; and Yang, Y. 2023. Is ChatGPT a Financial Expert? Evaluating Language Models on Financial Natural Language Processing. In *Findings of the Association for Computational Linguistics*.

Gupta, P.; Kirtania, S.; Singha, A.; et al. 2024. MetaReflection: Learning Instructions for Language Agents using Past Reflections. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *The Conference on Empirical Methods in Natural Language Processing*.

He, P.; Liu, X.; Gao, J.; et al. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In *The International Conference on Learning Representations*.

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

Hu, B.; Yuan, H.; Pandelea, V.; Luo, W.; Zhao, Y.; and Ma, Z. 2025. Extract, Match, and Score: An Evaluation Paradigm for Long Question-context-answer Triplets in Financial Analysis. In *CLR 2025 Workshop on Advances in Financial AI: Opportunities, Innovations and Responsible AI*.

Huang, J.; Chen, X.; Mishra, S.; et al. 2024. Large Language Models Cannot Self-Correct Reasoning Yet. In *The International Conference on Learning Representations*.

Jang, Y.; Lee, J.; and Kim, K.-E. 2022. GPT-critic: Offline reinforcement learning for end-to-end task-oriented dialogue systems. In *The International Conference on Learning Representations*.

Ji, Z.; Yu, T.; Xu, Y.; et al. 2023. Towards Mitigating LLM Hallucination via Self Reflection. In *Findings of the Association for Computational Linguistics*.

Kim, G.; Baldi, P.; and McAleer, S. 2023. Language models can solve computer tasks. *The Conference on Neural Information Processing Systems*.

Koutcheme, C.; Dainese, N.; Sarsa, S.; et al. 2024. Open source language models can provide feedback: Evaluating llms' ability to help students using gpt-4-as-a-judge. In *The Innovation and Technology in Computer Science Education*.

Lampinen, A.; Dasgupta, I.; Chan, S.; et al. 2022. Can language models learn from explanations in context? In *Findings of the Association for Computational Linguistics*.

Lewis, M.; et al. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *The Annual Meeting of the Association for Computational Linguistics*.

Li, J.; Cheng, X.; Zhao, X.; et al. 2023a. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. In *The Conference on Empirical Methods in Natural Language Processing*.

Li, M.; Wang, W.; Feng, F.; et al. 2024. Think Twice Before Trusting: Self-Detection for Large Language Models through Comprehensive Answer Reflection. In *Findings of the Association for Computational Linguistics*.

Li, X.; Chan, S.; Zhu, X.; et al. 2023b. Are ChatGPT and GPT-4 General-Purpose Solvers for Financial Text Analytics? A Study on Several Typical Tasks. In *The Conference on Empirical Methods in Natural Language Processing*.

Li, Y.; et al. 2023c. Making Language Models Better Reasoners with Step-Aware Verifier. In *The Annual Meeting of the Association for Computational Linguistics*.

Lightman, H.; Kosaraju, V.; Burda, Y.; et al. 2024. Let's Verify Step by Step. In *The International Conference on Learning Representations*.

Lin, Z.; Guan, S.; Zhang, W.; et al. 2024. Towards trustworthy LLMs: a review on debiasing and dehallucinating in large language models. *Artificial Intelligence Review*.

Liu, A.; Feng, B.; Wang, B.; et al. 2024a. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv:2405.04434*.

Liu, L.; Liu, X.; Wong, D. F.; et al. 2024b. SelectIT: Selective Instruction Tuning for LLMs via Uncertainty-Aware Self-Reflection. In *The Conference on Neural Information Processing Systems*.

Liu, Y.; Ott, M.; Goyal, N.; et al. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692*.

Madaan, A.; Tandon, N.; Gupta, P.; et al. 2023. Self-Refine: Iterative Refinement with Self-Feedback. In *The Conference on Neural Information Processing Systems*.

Mesnard, T.; Hardin, C.; Dadashi, R.; et al. 2024. Gemma: Open Models Based on Gemini Research and Technology. *arXiv:2403.08295*.

Miehling, E.; Desmond, M.; Natesan Ramamurthy, K.; et al. 2025. Evaluating the Prompt Steerability of Large Language Models. In *The Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*.

Mishra, S.; et al. 2022. LILA: A Unified Benchmark for Mathematical Reasoning. In *The Conference on Empirical Methods in Natural Language Processing*.

Pearson, K. 1895. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58(347-352): 240–242.

Renze, M.; and Guven, E. 2024. Self-reflection in large language model agents: Effects on problem-solving performance. In *The International Conference on Foundation and Large Language Models*.

Shinn, N.; Cassano, F.; Gopinath, A.; et al. 2024. Reflexion: language agents with verbal reinforcement learning. In *The Conference on Neural Information Processing Systems*.

Sun, H.; Cai, H.; Wang, B.; et al. 2024. Towards Verifiable Text Generation with Evolving Memory and Self-Reflection. In *The Conference on Empirical Methods in Natural Language Processing*.

Tang, L.; Goyal, T.; Fabbri, A.; et al. 2023. Understanding Factual Errors in Summarization: Errors, Summarizers, Datasets, Error Detectors. In *The Annual Meeting of the Association for Computational Linguistics*.

Tang, L.; et al. 2024. TofuEval: Evaluating Hallucinations of LLMs on Topic-Focused Dialogue Summarization. In *The Conference of the North American Chapter of the Association for Computational Linguistics*.

Touvron, H.; Lavril, T.; Izacard, G.; et al. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv:2302.13971*.

Turpin, M.; Michael, J.; Perez, E.; and Bowman, S. R. 2023. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. In *The Conference on Neural Information Processing Systems*.

Uesato, J.; Kushman, N.; Kumar, R.; et al. 2022. Solving math word problems with process-based and outcome-based feedback. *arXiv:2211.14275*.

Wang, B.; et al. 2023. Can ChatGPT Defend its Belief in Truth? Evaluating LLM Reasoning via Debate. In *Findings of the Association for Computational Linguistics*.

Wang, P.; Li, L.; Shao, Z.; et al. 2024. Math-Shepherd: Verify and Reinforce LLMs Step-by-step without Human Annotations. In *The Annual Meeting of the Association for Computational Linguistics*.

Welleck, S.; Lu, X.; West, P.; et al. 2023. Generating Sequences by Learning to Self-Correct. In *The International Conference on Learning Representations*.

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

Wu, Y.; Yuan, H.; Zhang, L.; and Ma, Z. 2025. Natural Language Inference as a Judge: Detecting Factuality and Causality Issues in Language Model Self-Reasoning for Financial Analysis. In *Proceedings of the Tenth Workshop on Financial Technology and Natural Language Processing*.

Wu, Z.; et al. 2024. Large Language Models Can Self-Correct with Key Condition Verification. In *The Conference on Empirical Methods in Natural Language Processing*.

Xie, Q.; Han, W.; Zhang, X.; et al. 2023. PIXIU: A Comprehensive Benchmark, Instruction Dataset and Large Language Model for Finance. In *The Conference on Neural Information Processing Systems*.

Yan, H.; et al. 2024. Mirror: Multiple-perspective Self-Reflection Method for Knowledge-rich Reasoning. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *The Annual Meeting of the Association for Computational Linguistics*.

Ye, X.; and Durrett, G. 2024. The unreliability of explanations in few-shot prompting for textual reasoning. In *The Conference on Neural Information Processing Systems*.

Ye, Z.; Li, X.; Li, Q.; et al. 2025. Learning LLM-as-a-judge for preference alignment. In *The International Conference on Learning Representations*.

Yuan, H.; Kang, L.; Li, Y.; and Fan, Z. 2024. Human-in-the-loop machine learning for healthcare: current progress and future opportunities in electronic health records. *Medicine Advances*, 2(3): 318–322.

Yuan, H.; Zhang, L.; and Ma, Z. 2025. Exploring the Reliability of Self-explanation and its Relationship with Classification in Language Model-driven Financial Analysis. In *CLR 2025 Workshop on Advances in Financial AI: Opportunities, Innovations and Responsible AI*.

Yuan, H.; Zhao, Y.; Zhang, L.; Luo, W.; and Ma, Z. 2025. Quantifying the Impact of Structured Output Format on Large Language Models through Causal Inference. *arXiv preprint arXiv:2509.21791*.

Zehlike, M.; Bonchi, F.; Castillo, C.; et al. 2017. FA*IR: A Fair Top-k Ranking Algorithm. In *The ACM on Conference on Information and Knowledge Management*.

Zhao, R.; et al. 2023. Verify-and-Edit: A Knowledge-Enhanced Chain-of-Thought Framework. In *The Annual Meeting of the Association for Computational Linguistics*.

Zhao, X.; Zhang, H.; Pan, X.; et al. 2024. Fact-and-Reflection (FaR) Improves Confidence Calibration of Large Language Models. In *Findings of the Association for Computational Linguistics*.