

Approach for Detecting, Correcting, and Verifying LLM Outputs in High-Stakes Domains

Contents

Problem statement	2
Research Objectives.....	2
Literature review	3
References	3

Problem statement

We used to know in software engineering that there no silver bullet but only trade off and same thing happened with AI (Large Language Models or LLMs) in important areas like finance or medicine, they sometimes make up facts, a problem called "hallucination". This is not acceptable finance or medicine domains while it could be acceptable in others. This is a huge risk in finance because bad information can lead to rule-breaking, losing money, or even bigger market issues.

In finance, everything relies on getting the **numbers, tables, and dates exactly right**. Studies show that the worst errors in finance are usually numerical and relate to specific times, like wrong stock prices on a certain day. This forces us to focus on building systems that handle structured data perfectly.

Research Objectives

We need a new, proven method. Our main goal is to test a specialized system to see if it can stop numerical mistakes in financial AI.

Overall Research Goal: To **design, test, and measure** a new, specialized RAG system that focuses on numbers to stop multi-step numerical hallucinations in financial AI models, aiming to lower the critical error rates.

Supporting Research Objectives:

1. **To design** a special RAG system that keeps numerical calculations separate from the natural language part, the system sticks to the original data in tables.
2. **To create** a test set, based on existing complex benchmarks (like FAITH), focused on counting numerical errors and specifically the "scale errors" in financial documents.
3. **To evaluate** how well our proposed RAG system works against current AI models; by measuring how much it reduces numerical and scale errors in complex table tasks.
4. **To analyze** which specific parts of the new system (like the data handling module) are most effective at reducing hallucinations, providing clear steps for building reliable financial AI.

This research objectives try to fix the lack of calculation ability and the lack of proof; we are designing a special RAG system. The main idea is to **separate the calculations (which must be perfect) from the language generation (which can be creative)**. This way, all high-risk

math is done by a verifiable, external software part.

Literature review

Large Language Models (LLMs) can create fluent, human-like text, but they sometimes “hallucinate”—producing answers that sound right but are actually false or made up. This problem is especially serious in medicine and finance, where wrong information can harm people or cause financial losses.

This review looks at research on these risks. In medicine, hallucinations can lead to wrong diagnoses or unsafe advice. In finance, they can cause non-compliance, financial loss, or market instability.

The report highlights two key ideas: first, each field defines “truth” differently—medicine relies on clinical facts, while finance depends on verified numbers. Second, while most see hallucination as a flaw, some research suggests that in creative areas like drug discovery, it might sometimes be useful.

References

1. Journey of Hallucination-minimized Generative AI Solutions ... - arXiv, accessed November 7, 2025, <https://arxiv.org/pdf/2311.10961>
2. [2501.13824] Can Hallucinations Help? Boosting LLMs for Drug Discovery - arXiv, accessed November 7, 2025, <https://arxiv.org/abs/2501.13824>
3. MedHalu: Hallucinations in Responses to Healthcare ... - arXiv, accessed November 7, 2025, <https://arxiv.org/abs/2409.19492>
4. AI Hallucination: Comparison of the Popular LLMs - Research AIMultiple, accessed November 7, 2025, <https://research.aimultiple.com/ai-hallucination/>
5. Hallucination in artificial-intelligence-assisted diagnosis of arthritis: a ..., accessed November 7, 2025, <https://pubmed.ncbi.nlm.nih.gov/41161297/>

6. Medical Hallucination in Foundation Models and Their Impact on Healthcare - medRxiv, accessed November 7, 2025, <https://www.medrxiv.org/content/10.1101/2025.02.28.25323115v1.full-text>
7. Generative Artificial Intelligence in Finance: Risk Considerations - IMF, accessed November 7, 2025, <https://www.imf.org/-/media/Files/Publications/FTN063/2023/English/FTNEA2023006.ashx>
8. The rise of artificial intelligence: benefits and risks for financial stability, accessed November 7, 2025, https://www.ecb.europa.eu/press/financial-stability-publications/fsr/special/html/ecb.fsrart202405_02~58c3ce5246.en.html
9. [2503.05777] Medical Hallucinations in Foundation Models and Their Impact on Healthcare, accessed November 7, 2025, <https://arxiv.org/abs/2503.05777>
10. Hallucinations and Key Information Extraction in Medical Texts: A Comprehensive Assessment of Open-Source Large Language Models - arXiv, accessed November 7, 2025, <https://arxiv.org/html/2504.19061v2>
11. FAITH: A Framework for Assessing Intrinsic Tabular ... - arXiv, accessed November 7, 2025, <https://arxiv.org/abs/2508.05201>
12. Beyond the Reported Cutoff: Where Large Language Models Fall Short on Financial Knowledge - arXiv, accessed November 7, 2025, <https://arxiv.org/html/2504.00042v2>
13. FinanceQA: A Benchmark for Evaluating Financial Analysis Capabilities of Large Language Models - ResearchGate, accessed November 7, 2025, https://www.researchgate.net/publication/388529404_FinanceQA_A_Benchmark_for_Evaluating_Financial_Analysis_Capabilities_of_Large_Language_Models
14. Deficiency of Large Language Models in Finance: An Empirical ..., accessed November 7, 2025, <https://arxiv.org/abs/2311.15548>
15. assignment1.docx
16. Towards Reliable Medical Question Answering: Techniques and Challenges in Mitigating Hallucinations in Language Models - arXiv, accessed November 7, 2025, <https://arxiv.org/html/2408.13808v1>
17. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools - Daniel E. Ho - Stanford University, accessed November 7, 2025, https://dho.stanford.edu/wp-content/uploads/Legal_RAG_Hallucinations.pdf
18. Financial Style Transfer to Enhance Creativity without Hallucinations in LLMs - arXiv, accessed November 7, 2025, <https://arxiv.org/html/2408.05365v4>
19. Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review - PubMed Central, accessed November 7, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC11480901/>
20. A Call to Address AI “Hallucinations” and How Healthcare Professionals Can Mitigate Their Risks - NIH, accessed November 7, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC10552880/>