

Examining LLM Hallucinations in Medicine and Finance

Contents

Literature review	3
Paper 1: Hallucination in artificial-intelligence-assisted diagnosis of arthritis: a case report	4
Paper 2: MedHalu: Hallucinations in Responses to Healthcare Queries by Large Language Models.....	4
Paper 3: Can Hallucinations Help? Boosting LLMs for Drug Discovery	5
Paper 4: Deficiency of Large Language Models in Finance: An Empirical Examination of Hallucination	5
Paper 5 :FAITH: A Framework for Assessing Intrinsic Tabular Hallucinations in finance.....	6
Paper 6: Journey of Hallucination-minimized Generative AI Solutions for Financial Decision Makers.....	6
Critical Analysis.....	7
Connections	11
Conflicts between the findings.....	11
References	13

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.

Paper 1: Hallucination in artificial-intelligence-assisted diagnosis of arthritis: a case report

- Link: <https://pubmed.ncbi.nlm.nih.gov/41161297/>

Summary

This 2025 paper describes a case where **ChatGPT-4o** was used to help diagnose a **51-year-old woman with hand pain (arthralgia)**. The goal was to show why doctors must be careful when using generative AI in real medical situations. The study focused on this single case. The main finding was that the AI **hallucinated**—it made up or gave inaccurate information when interpreting the patient's **radiology and lab results**, which then led to **incorrect treatment suggestions**. The authors conclude that **strong verification steps and expert human oversight** are needed to keep patients safe.

Paper 2: MedHalu: Hallucinations in Responses to Healthcare Queries by Large Language Models

- Link: <https://arxiv.org/abs/2409.19492>

Summary

This 2024 study conducts a "pioneering study" on LLM hallucinations in a different context: responses to *real-world healthcare queries from laypeople*. The objective was to move beyond standardized medical exams, which do not capture the ambiguity of real patient interactions, and to create a new benchmark, *MedHalu*, for this purpose. The methodology involved collecting queries from public, expert-curated healthcare datasets (e.g., HealthQA, LiveQA). Hallucinated answers were then generated by prompting GPT-3.5 with specific instructions to create *input-conflicting*, *context-conflicting*, and *fact-conflicting* responses. A 500-pair subset of this benchmark was validated by 30 medical experts. The study's key findings are stark: it reveals a significant gap in detection ability between medical experts (0.70 macro-F1 score) and laypeople (0.57 macro-F1). Critically, even advanced LLMs like GPT-4 performed similarly to laypeople (0.56 macro-F1) and *significantly underperformed experts*. The paper also proposed an "expert-in-the-loop" mitigation strategy, which

improved GPT-4's detection by 6.3%.

Paper 3: Can Hallucinations Help? Boosting LLMs for Drug Discovery

- Link: <https://arxiv.org/abs/2501.13824>

Summary

This 2025 paper questions the common view that LLM hallucinations are always negative. It tests the idea that these errors might have creative value and could improve performance in predicting molecular properties, an important step in early drug discovery.

Researchers used seven different LLMs and asked them to write natural language descriptions from molecular SMILES strings. These descriptions often contained hallucinations, but when added as extra input features in later classification tasks, prediction accuracy improved for some models. Hallucinations produced by GPT-4o gave the best results. The researchers identified more than 18,000 useful hallucinations, with the most effective type being “structural misdescriptions,” or false statements about molecular structure that appeared to increase model confidence.

Paper 4: Deficiency of Large Language Models in Finance: An Empirical Examination of Hallucination

- Link: <https://arxiv.org/abs/2311.15548>

Summary

This 2023 paper takes a broad look at how LLMs hallucinate in different financial tasks, aiming to measure their basic weaknesses in a field that demands high accuracy. The researchers tested several models, including Llama2, GPT-3.5, GPT-4, and the finance-focused FinMA-7B, on three tasks: recognizing financial abbreviations, explaining financial terms, and retrieving historical stock prices. They measured performance using accuracy, FactScore for explanations, and Mean Absolute Error for stock price predictions. The study found that general-purpose LLMs show serious hallucination problems in these financial tasks, highlighting the need for more specialized, domain-specific solutions.

Paper 5 :FAITH: A Framework for Assessing Intrinsic Tabular Hallucinations in finance

- Link: <https://arxiv.org/abs/2508.05201>

Summary

This 2024 paper provides a more advanced analysis of hallucinations in finance. Its goal was to create FAITH, a framework for measuring intrinsic hallucinations—cases where a model fails to accurately use the context, especially financial tables.

The methodology was novel: the authors used a "context-aware masked span prediction" task. They took tables and text from the "Item 7" section of 10-K reports, hid numerical values in the text, and asked LLMs to recover those numbers using only the table. The paper also introduced a taxonomy of four reasoning types of increasing difficulty: Direct Lookup, Comparative Calculation, Bivariate Calculation, and Multivariate Calculation.

Key findings showed that even the most advanced models make errors 10-20% of the time on multi-step numerical tasks. One major failure type was "scale error," where the model gets the number right but misses the correct magnitude, for example "\$150" instead of "\$150 million."

Paper 6: Journey of Hallucination-minimized Generative AI Solutions for Financial Decision Makers

- Link: <https://arxiv.org/abs/2311.10961>

Summary

This 2024 paper offers a practical solution to the hallucination problems identified in earlier studies. It proposes a three-stage approach—Prototyping, Scaling, and Evolution with human feedback—for creating LLM systems designed to reduce hallucinations in financial decision-making.

The method uses a Langchain-based framework with four custom modules: Question

Intention Classification, Data Chunk Generation and Filtering (turning tables into text), Custom Prompt Generation (similar to RAG), and Response Quality Scoring to check outputs for hallucinations. The paper argues that this modular prompting approach can reduce errors, especially those that occur when combining numerical facts with language context.

Critical Analysis

Across medical and financial research, one main idea stands out: **LLMs can sound confident but still make things up**, and this is especially risky in fields where mistakes have serious consequences.

A major strength of recent studies is that they look at **real questions and real tasks**, not just artificial tests. This makes it easier to see how hallucinations actually appear in practice—whether in answering patient questions, reading lab results, or interpreting financial reports. Researchers are also getting better at **classifying different types of errors**, which helps us understand whether a model failed because it lacked context, misunderstood numbers, or handled tables badly.

There is also an important insight: hallucinations are **not always bad**. In creative areas, like early-stage drug discovery, unexpected ideas can sometimes help. But in precise jobs—like diagnosing a patient or reporting stock prices—hallucinations are dangerous.

Despite progress, the research has clear weaknesses. Many studies use small or biased datasets, rely on LLMs to check other LLMs, or propose solutions without actually testing them. As a result, we still don't have reliable proof that current models can safely handle high-stakes decisions.

Overall, the big message is simple: **LLMs are improving, but they are not yet safe or accurate enough for medical or financial decision-making**. To fix this, the field needs better testing, more expert oversight, and tools built specifically to handle numbers and structured data—things general-purpose models still struggle with.

Synthesis Matrix for Literature Review:

Publication Year	Author(s)	Methodology	Key Findings	Weaknesses/ Limitations
2025	Yuan, S. et al. ²	Empirical ML Experiment: Prompted 7 LLMs to generate text from SMILES strings; used this text as a feature in downstream molecule property prediction tasks.	Conflict Identified: Hallucinations ("structural misdescriptions") <i>significantly improved</i> predictive accuracy in drug discovery. GPT-4o's hallucinations were most beneficial.	Limited Generalizability: Findings are likely specific to <i>exploratory/predictive</i> tasks (drug discovery) and not <i>precision</i> tasks (diagnosis). Causal mechanism is not fully proven.
2025	(Authors of ⁵)	Case Report: Observational analysis of a single diagnostic attempt (hand arthralgia) using ChatGPT-4o in a clinical setting.	Baseline Risk: The AI "hallucinated" by <i>fabricating</i> unsupported radiologic and laboratory findings, leading to incorrect treatment recommendations.	Not Generalizable : As a single case report, findings cannot be generalized. ⁵ Lacks methodological details (e.g., specific prompts used).

2024	(Authors of ³)	Benchmark Creation & Evaluation: Created <i>MedHalu</i> benchmark from real-world patient queries. Generated hallucinations (fact-conflicting, etc.) and had 30 experts validate a subset.	LLMs (0.56 F1) perform as poorly as laypeople (0.57 F1) and worse than experts (0.70 F1) at detecting medical hallucinations. ³	Benchmark Bias: Hallucinations were <i>generated</i> by GPT-3.5, which may not represent all hallucination types. ³ Only 25% of the benchmark was human-validated.
2024	Roychowdhury, S. ¹	Systems Design Proposal: Proposed a 3-stage (Prototyping, Scaling, Evolution) journey and a 4-module (Intent, Chunking, Prompting, Scoring) Langchain framework.	A modular, RAG-like architecture can <i>minimize</i> hallucinations that arise from combining <i>mathematical facts with language context</i> . ¹	Lacks Empirical Validation: The paper <i>proposes</i> a framework but provides no quantitative results or benchmarks to <i>prove</i> its effectiveness. ¹
2024	Zhang, M. et al. ¹¹	Benchmark Creation & Evaluation:	Numerical Failure: Even top LLMs have 10-20% error	Unsolved Problem: The 10-20% failure rate in <i>frontier</i>

		Created <i>FAITH</i> benchmark using a "masked span prediction" task on 10-K reports. Developed a 4-level taxonomy of numerical reasoning complexity.	rates on <i>multi-step numerical reasoning</i> . ¹¹ A common failure is "scale error" (e.g., \$150 vs. \$150M).	models highlights this is a "critical barrier" ¹¹ and these tools are not yet reliable for complex financial analysis.
2023	Kang, H. & Liu, X. ¹⁴	Empirical Examination: Evaluated multiple LLMs (Llama2, GPTs) on three financial tasks: abbreviation recognition, term explanation, and historical stock price query.	Baseline Deficiency: Off-the-shelf LLMs exhibit "serious hallucination behaviors" in financial tasks, highlighting a "fundamental deficiency". ¹⁴	Limited Task Scope: The "stock price" task is a <i>retrieval/memory</i> task, not a <i>reasoning</i> task. The "FactScore" metric used a potentially circular (LLM-based) ground truth. ¹⁴

Connections

This analysis shows a clear link between medicine and finance: both high-stakes fields use Retrieval-Augmented Generation (RAG) to reduce the risks of unverified AI output. In these industries, incorrect information can have serious consequences, such as patient harm or financial loss. This shared risk has led researchers to treat LLMs not as fully autonomous experts but as reasoning tools grounded in reliable, domain-specific data.

The financial "Journey" paper illustrates this with its Data Chunk Generation and Custom Prompt Generation modules, which form a RAG-like system. Similarly, the medical MedHalu paper uses an expert-in-the-loop approach, conceptually similar to RAG, where expert reasoning provides the verified information. Other studies confirm that RAG and careful fine-tuning are key strategies for creating reliable medical question-answering systems. This convergence across fields shows a broad agreement: in high-stakes applications, LLM outputs must be anchored to external, verifiable facts.

Conflicts between the findings

1. The Divergent Nature of "Ground Truth"

The biggest difference between domains is the type of factual grounding they need. The benchmarks highlight this clearly.

- **Medical challenge:** Represented by MedHalu, it focuses on semantic and contextual accuracy. The benchmark checks whether an LLM's answer contradicts a large, unstructured body of authoritative biomedical knowledge. Errors are failures of reasoning or comprehension, such as giving advice that conflicts with facts.
- **Financial challenge:** Represented by FAITH, it focuses on numerical and tabular accuracy. The benchmark tests whether an LLM can correctly perform multi-step calculations using structured table data.

This difference explains why hallucinations vary by domain. A critical medical hallucination might be a made-up diagnosis or unsafe advice, while a critical financial hallucination might be a numerical error or a "scale error" (e.g., "\$150" instead of "\$150 million"). This shows that a one-size-fits-all RAG system is not enough: medical RAGs must retrieve and synthesize evidence, while financial RAGs must handle tabular data and accurate calculations.

2. The Conflict: Hallucination as "Bug" versus "Feature"

The second key conflict is about how hallucinations are seen.

- **"Bug" Paradigm:** Most papers agree that hallucination is a serious problem—a

barrier, a deficiency, and a risk that must be minimized. From this view, hallucinations are always harmful failures of factuality.

- **"Feature" Paradigm:** The drug discovery paper finds that hallucinations can improve predictive accuracy and be beneficial. Other research distinguishes "divergent-phase hallucination," which encourages creativity, from "convergent-phase hallucination," which should be controlled with RAG.

This is more than a simple contradiction. The word "hallucination" describes at least two very different phenomena:

1. **Ungrounded Fabrication:** The model invents facts, figures, or calculations unsupported by evidence or context. This is always harmful in precise, high-stakes tasks like medical diagnosis or financial reporting.
2. **Creative Divergence:** The model produces new, plausible connections or descriptions that are not obvious. This can be beneficial in exploratory tasks like drug discovery or scientific research.

Understanding this distinction is crucial. The hallucination dangerous in a hospital is not the same as the one useful in a drug discovery lab.

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/>