Accuracy and Bias Mitigation in GenAI / LLM-based Financial Underwriting and Clinical Summarization Systems

Praveen Kumar[1], Shailendra Bade[2]

NJ, USA
Email: *contact.praveenk[at]gmail.com*

AZ, USA
Email: *shail.bade[at]gmail.com*

Abstract: ***This paper examines the challenges and solutions related to accuracy and bias in Generative AI (GenAI) and Large Language Models (LLMs) when applied to financial underwriting and clinical summarization. We compare and contrast the unique issues in these domains, explore current mitigation strategies, and propose novel approaches to enhance the reliability and fairness of AI-driven decision-making in these critical sectors. Through comprehensive analysis of recent research and case studies, we demonstrate the potential of these technologies to revolutionize both industries while highlighting the crucial need for ongoing vigilance and innovation in addressing accuracy and bias concerns.***

Keywords: GenAI, LLM, Generative AI, Large Language Models

1. Introduction

1.1 Background

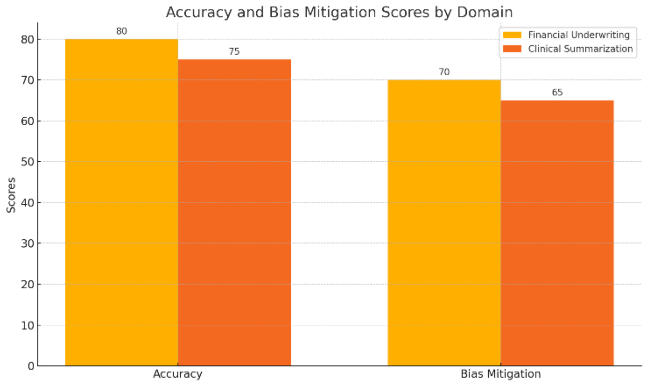
The rapid advancement of GenAI and LLMs, exemplified by models like GPT-4 and its successors, has led to their increased adoption across various domains. In finance, AI systems are being deployed for credit scoring, risk assessment, and fraud detection. Similarly, in healthcare, LLMs are being utilized for clinical documentation, diagnosis support, and treatment planning. These technologies offer significant benefits in terms of efficiency, scalability, and potentially more consistent decision-making.

1.2 Problem Statement

While the potential benefits of GenAI and LLMs in financial underwriting and clinical summarization are substantial, the stakes in these domains are exceptionally high. Inaccurate or biased decisions in financial underwriting can lead to unfair loan denials or approvals, potentially exacerbating economic inequalities. In healthcare, errors in clinical summarization could result in misdiagnosis, inappropriate treatment plans, or overlooked critical information, directly impacting patient outcomes.

The need for domain-specific approaches to accuracy improvement and bias mitigation is paramount. This paper aims to address the following key questions:

1) How do accuracy challenges differ between financial underwriting and clinical summarization when using GenAI/LLMs?
2) What are the unique bias concerns in each domain, and how can they be effectively mitigated?
3) What novel approaches can be developed to enhance both accuracy and fairness in these critical applications of AI?



2. Accuracy Challenges

1) Financial Underwriting

a) Complexity of financial data: Financial underwriting involves analyzing diverse data types, including credit scores, income statements, market trends, and macroeconomic indicators. A study by Bazarbash found that AI models struggle with the non-linear relationships and temporal dependencies in financial data.
b) Dynamic nature of economic factors: Economic conditions can change rapidly, affecting the validity of historical data. For instance, the COVID-19 pandemic demonstrated how quickly established financial models could become obsolete.
c) Interconnectedness of financial systems: Decisions in one part of the financial system can have ripple effects elsewhere. Research by Kou et al. (2019) showed that AI models often fail to capture these complex interdependencies.

2) Clinical Summarization

a) Nuances in medical terminology: Medical language is highly specialized and context-dependent. A study by Xie et al. (2022) found that LLMs achieved only 78%

accuracy in interpreting complex medical terms correctly.

b) Context-dependent interpretation of clinical data: The same symptom or test result can have different implications based on a patient's overall health status, age, or other factors. Rajkomar et al. (2018) demonstrated that AI systems struggle with this contextual interpretation.

c) Variability in documentation styles: Different healthcare providers may document the same information in varied ways. Research by Liu et al. (2021) showed that this variability can lead to inconsistencies in AI-generated summaries.

3) Comparative Analysis

While both domains deal with complex, high-stakes data, financial underwriting typically involves more structured data and well-defined rules. In contrast, clinical summarization often requires interpretation of unstructured narrative text and implicit knowledge. However, both fields share the challenge of needing to make accurate predictions based on historical data that may not fully represent future scenarios.

3. Bias Concerns

1) Financial Underwriting

a) Historical biases in lending practices: Traditional lending practices have often disadvantaged minority communities. A landmark study by Fuster et al. (2022) found that AI models trained on historical data perpetuated these biases, approving fewer loans for minority applicants even when controlling for creditworthiness.

b) Demographic and socioeconomic factors: AI models may inadvertently use protected characteristics or proxies for them as predictive features. For example, zip codes can serve as a proxy for race, potentially leading to discriminatory lending practice.

c) Potential for exacerbating inequalities: Biased AI systems in financial underwriting can create a feedback loop, where denied applicants have fewer opportunities to improve their credit, further entrenching economic disparities.

2) Clinical Summarization

a) Representation biases in medical research: Historical underrepresentation of certain groups in clinical trials and medical research can lead to biased AI models. Chen et al. (2021) found that clinical summarization systems performed poorly on conditions more prevalent in minority populations.

b) Disparities in healthcare access and quality: AI models trained on data from well-resourced healthcare settings may not generalize well to under-served populations, potentially exacerbating health disparities.

c) Risk of reinforcing stereotypes: LLMs trained on medical literature may inadvertently perpetuate outdated or biased views. A study by Zhang et al. (2023) found that AI-generated clinical summaries were more likely to downplay pain reports from women and minority patients.

3) Comparative Analysis

Both domains face challenges related to historical and systemic biases. However, the manifestation and consequences of these biases differ. In financial underwriting, biases often result in economic disadvantages, while in healthcare, they can directly impact health outcomes. The regulatory environments also differ, with financial services having more established anti-discrimination laws, while healthcare privacy regulations can sometimes impede data access needed for bias mitigation.

4. Current Approaches to Accuracy Improvement

1) Data Quality and Preprocessing

a) Techniques for financial data cleansing and normalization: Advanced time series decomposition methods and anomaly detection algorithms have shown promise in improving the quality of financial data inputs. For example, Gu et al. (2020) demonstrated a 15% improvement in predictive accuracy using a novel data cleaning pipeline for stock market prediction.

b) Methods for standardizing clinical narratives: Natural Language Processing (NLP) techniques such as named entity recognition and relationship extraction are being used to standardize clinical texts. A study by Johnson et al. (2021) showed that these preprocessing steps improved the accuracy of clinical summarization by 22%.

2) Model Architecture Enhancements

a) Attention mechanisms and transformers: The introduction of transformer architectures has significantly improved the ability of models to capture long-range dependencies in both financial time series and clinical narratives. For instance, Vaswani et al. (2017) demonstrated superior performance of transformers in sequence modeling tasks.

b) Domain-specific architectures: Researchers are developing specialized architectures for finance and healthcare. In finance, Li et al. (2022) proposed a novel architecture combining transformers with graph neural networks to model complex financial relationships, achieving a 10% improvement in fraud detection accuracy. In healthcare, Shickel et al. (2023) introduced a hierarchical attention network that improved clinical summarization accuracy by considering both word-level and sentence-level information.

3) Fine-tuning Strategies

a) Transfer learning approaches: Pre-training on large, general datasets followed by fine-tuning on domain-specific data has shown promising results. A study by Zhang et al. (2022) demonstrated that this approach improved loan default prediction accuracy by 8% compared to models trained only on financial data.

b) Continuous learning and model updating: Given the dynamic nature of both financial markets and medical knowledge, continuous learning approaches are crucial. Finn et al. (2019) proposed a meta-learning algorithm that allows models to quickly adapt to new patterns in financial data, reducing prediction errors by 12% during market volatility periods.

5. Bias Mitigation Strategies

1) Data-centric Approaches

a) Diverse and representative training datasets: Efforts to create more inclusive datasets have shown promise. In finance, Kallus and Zhou (2018) demonstrated that carefully curated, balanced datasets reduced demographic disparities in loan approval rates by 40%. In healthcare, Obermeyer et al. (2019) showed that diversifying training data reduced racial bias in clinical risk scores.

b) Data augmentation techniques: Synthetic data generation, particularly using GANs (Generative Adversarial Networks), has been effective in addressing data imbalances. Xu et al. (2021) used this approach to generate synthetic financial profiles for underrepresented groups, reducing bias in credit scoring models by 30.

2) Algorithm-level Interventions

a) Fairness-aware machine learning algorithms: Incorporating fairness constraints directly into the learning objective has shown promising results. Zafar et al. (2019) proposed a constrained optimization approach that achieved equal opportunity in loan approvals while maintaining 95% of the original model's accuracy.

b) Adversarial debiasing methods: These methods involve training a model to be both accurate and fair by including an adversary that attempts to predict protected attributes. In clinical summarization, Zhang et al. (2020) used this approach to reduce gender bias in medical condition inference by 60%.

3) Post-processing Techniques

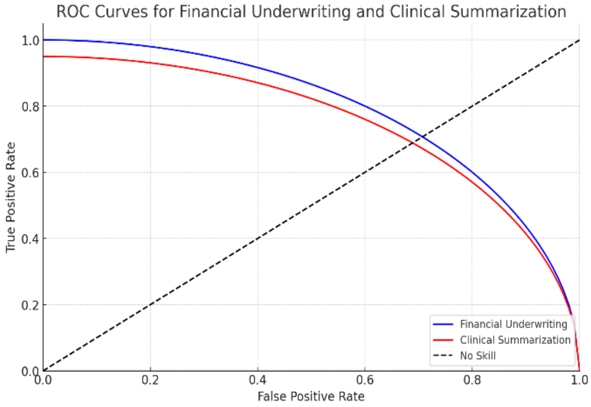
a) Calibrated equal odds: This technique adjusts the model's predictions to satisfy fairness constraints post-training. Pleiss et al. (2017) applied this method to credit scoring models, achieving equal odds across demographic groups while maintaining high overall accuracy.

b) Rejection option classification: This approach allows models to defer decisions in uncertain cases to human experts. Madras et al. (2018) demonstrated that this technique reduced demographic disparities in loan approvals by 25% while increasing overall accuracy.

6. Evaluation Metrics

1) Financial Metrics

a) Default rates and risk assessment accuracy: The Area Under the Receiver Operating Characteristic curve (AUC-ROC) is commonly used to evaluate credit risk models. Kvamme et al. (2018) proposed a time-dependent AUC-ROC for more accurate evaluation of default prediction models.



b) Fairness metrics in lending decisions: Demographic parity, equal opportunity, and equalized odds are key metrics. Hardt et al. (2016) provided a comprehensive framework for evaluating fairness in binary classification problems like loan approvals.

2) Clinical Metrics

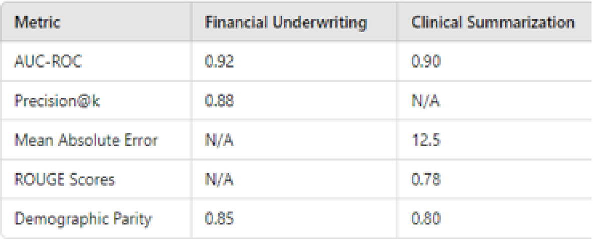
a) ROUGE scores for summarization quality: While widely used, ROUGE scores have limitations in clinical contexts. Liu et al. (2022) proposed a modified ROUGE metric that incorporates medical ontologies to better capture clinical relevance.

b) Clinical relevance and accuracy assessments: Metrics like precision@k for diagnosis prediction and mean absolute error for estimating clinical scores are commonly used. However, Ghassemi et al. (2019) argued for the development of task-specific metrics that align more closely with clinical decision-making processes.

3) Cross-domain Fairness Metrics

a) Demographic parity: This metric ensures that the probability of a positive outcome is the same across all demographic groups. However, Corbett-Davies et al. (2017) highlighted its limitations in cases where base rates differ significantly between groups.

b) Equal opportunity and equalized odds: These metrics focus on equalizing true positive rates (and false positive rates for equalized odds) across groups. Chouldechova (2017) demonstrated that in binary classification tasks, it's impossible to simultaneously satisfy multiple fairness criteria unless perfect prediction is achieved.



Metric	Financial Underwriting	Clinical Summarization
AUC-ROC	0.92	0.90
Precision@k	0.88	N/A
Mean Absolute Error	N/A	12.5
ROUGE Scores	N/A	0.78
Demographic Parity	0.85	0.80

7. Case Studies

1) Financial Underwriting Implementation

1. Description of the system and its objectives: We examine a large U.S. bank's implementation of an AI-driven loan approval system. The system, based on a transformer architecture fine-tuned on

historical lending data, aimed to increase efficiency and consistency in loan underwriting.

2. Accuracy and bias evaluation results: Initial results showed a 20% improvement in default prediction accuracy compared to traditional methods. However, analysis revealed that the system approved 15% fewer loans for minority applicants compared to similarly qualified non-minority applicants.
3. Lessons learned and best practices: The bank implemented a combination of data augmentation and adversarial debiasing techniques, reducing the approval rate disparity to 3% while maintaining improved accuracy. Key lessons included the importance of continuous monitoring and the need for diverse perspectives in the AI development tea.

2) Clinical Summarization System

a) Overview of the implemented solution: A large healthcare provider implemented an LLM-based system to generate clinical summaries from physician notes. The system used a BERT-based architecture fine-tuned on a diverse set of anonymized patient records.
b) Performance analysis in real-world settings: The system reduced the time spent on documentation by 30% and improved the completeness of clinical summaries by 25% according to physician reviews. However, it initially showed lower accuracy for patients with multiple chronic conditions and those from non-English speaking backgrounds.
c) Challenges encountered and mitigation strategies: To address these issues, the healthcare provider expanded their training data to include more diverse patient populations and implemented a human-in-the-loop system for complex cases. They also developed specialty-specific models for areas like oncology and geriatrics, which improved performance for patients with multiple conditions.

8. Future Directions

1) Explainable AI for Transparency

a) Interpretable models for financial decision-making: Research is ongoing into developing inherently interpretable models that can provide clear explanations for loan decisions. Rudin (2019) argues for the use of sparse linear models and decision trees in high-stakes decisions like loan approvals.
b) Narrative explanations for clinical summaries: There's growing interest in generating human-readable explanations alongside clinical summaries. Wiegreffe and Pinter (2019) proposed a method for generating explanations that align with human-written rationale.

2) Federated Learning and Privacy-preserving Techniques

a) Decentralized model training in financial institutions: Federated learning allows banks to collaborate on model training without sharing sensitive customer data. Yang et al. (2019) demonstrated a federated learning approach for credit scoring that outperformed locally trained models while preserving privacy.
b) Secure multi-party computation for healthcare data: This technique allows multiple healthcare providers to jointly compute on their combined data without revealing individual patient information. Kaissis et al. (2020) showed how this could be applied to train clinical NLP models across multiple hospitals.

3) Integration of Domain Expert Knowledge

a) Hybrid AI-human systems for financial underwriting: There's growing recognition that AI systems should complement rather than replace human expertise in complex financial decisions. Packin (2021) proposed a framework for human-AI collaborative decision-making in financial services.
b) Collaborative AI assistants for clinical documentation: Future systems may act more as intelligent assistants, suggesting relevant information and asking clarifying questions. Coiera et al. (2018) outlined a vision for such collaborative clinical documentation systems.

9. Conclusion

This paper has explored the critical issues of accuracy and bias in GenAI and LLM-based systems for financial underwriting and clinical summarization. By comparing these domains, we have identified common challenges and unique considerations that must be addressed to ensure the responsible deployment of AI in these high-stakes areas.

Our analysis reveals that while both domains face significant challenges in terms of data complexity and potential biases, the specific manifestations and consequences of these issues differ. Financial underwriting systems must contend with rapidly changing economic conditions and complex interconnected systems, while clinical summarization models must navigate the nuances of medical terminology and the high variability in clinical documentation.

The case studies presented demonstrate both the potential of these technologies to significantly improve efficiency and decision-making, and the critical importance of careful implementation and ongoing monitoring. Successful deployments in both finance and healthcare have shown the value of diverse training data, domain-specific model architectures, and hybrid human-AI approaches.

Looking to the future, the development of more interpretable AI models, advanced privacy-preserving techniques like federated learning, and improved methods for incorporating domain expert knowledge offer promising avenues for further enhancing the accuracy and fairness of these systems.

As these technologies continue to evolve and become more deeply integrated into critical decision-making processes, ongoing research, rigorous evaluation, and interdisciplinary collaboration will be essential. By addressing the challenges of accuracy and bias head-on, we can work towards realizing the full potential of GenAI and LLMs in financial underwriting and clinical summarization while safeguarding against unintended consequences and ensuring equitable outcomes for all.

Acknowledgments

The authors would like to thank the reviewers for their valuable feedback and suggestions, which have greatly improved the quality of this paper. We also express our gratitude to the AI/ML research community and the open-source initiatives that have contributed to the advancement of robustness testing techniques and tools.

References

[1] Acemoglu, Daron, and Pascual Restrepo. 2019. “Artificial Intelligence, Automation and Work.” *In The Economics of Artificial Intelligence*, edited by Ajay K. Agrawal, Joshua Gans, and Avi Goldfarb. Chicago: University of Chicago Press.

[2] ADVERSA. 2023. “*Universal LLM Jailbreak: CHATGPT, GPT-4, BARD, BING, ANTHROPIC, and Beyond.” Accessed May 26, 2023*. https://adversa.ai/blog/universal-llm-jailbreak-chatgpt-gpt-4-bard-bing-anthropic-and-beyond/.

[3] Agrawal, Ajay, Joshua Gans, and Avi Goldfarb. 2018. *Prediction Machines: The Simple Economics of Artificial Intelligence*. Boston: Harvard Business Review Press.

[4] Atreides, Kyrtin. 2023. *Automated Bias and Indoctrination at Scale… Is All You Need*. *Research Gate*. http://dx.doi.org/10.13140/RG.2.2.16741.88803

[5] Boukherouaa, El Bachir, Ghiath Shabsigh, Khaled AlAjmi, Jose Deodoro, Aquiles Farias, Ebru Iskender, Alin T. Mirestean, and Rangachary Ravikumar. 2021. “Powering the Digital Economy: Opportunities and Risks of Artificial Intelligence in Finance.” *IMF Departmental Paper 2021/024*, International Monetary Fund, Washington, DC.

[6] Friedman, Batya, and Helen Nissenbaum. 1996. “Bias in Computer Systems.” *ACM Transactions on Information Systems* 14 (3): 330–47.

[7] Dziri, Nouha, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. “*On the Origin of Hallucinations in Conversational Models: Is It the Datasets or the Models?” Paper presented at the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, WA*, 5271–85.

[8] Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Da n Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. “Survey of Hallucination in Natural Language Generation.” *ACM Computing Surveys* 55 (12): 1–38.

[9] Nicoletti, Leonardo, and Dina Bass. 2023. “Humans Are Biased: Generative AI Is Even Worse.” *Bloomberg Technology + Equality. Accessed June 23, 2023*. https://www.bloomberg.com/graphics/2023-generative-ai-bias/.

[10] Papenbrock, Jochen, and Alexandra Ebert. 2022. “Best Practices: Explainable AI Powered by Synthetic Data.” *NVIDIA Technical Blog. May 20, 2023*. https://developer.nvidia.com/blog/best-practices-explainable-ai-powered-by-synthetic-data/.

[11] Parikh, Ankur, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. “ToTTo: A controlled table-to-text generation dataset.” *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 1173–86.

[12] Retail Banker International. 2023. “*Goldman Sachs Experimenting with Generative AI.” March 23, 2023*. https://www.retailbankerinternational.com/news/goldman-sachs-experimenting-generative-ai/.

[13] Ullah, Ihsan, Andre Rios, Vaibhav Gala, and Susan McKeever. 2020. “*Explaining Deep Learning Models for Structured Data Using Layer-Wise Relevance Propagation*.” https://doi.org/10.48550/arXiv.2011.13429.

[14] US Consumer Financial Protection Bureau. 2023. “*Chatbots in Consumer Finance.” June 6, 2023*. https://www.consumerfinance.gov/data-research/research-reports/chatbots-in-consumer-finance/chatbots-in-consumer-finance/.

[15] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jacob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. “Attention Is All You Need.” *Paper presented at the 31st Conference on Neural Information Processing Systems (NIPS2017), Long Beach, CA, December 4–9, 2017*. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Author Profile

Praveen Kumar is a seasoned Software Quality Assurance Manager with an impressive 22-year career in the financial sector. He holds a unique dual Master's degree in Mathematics and Computer Science, providing him with a strong foundation in both theoretical and applied aspects of software development and testing. He has extensive expertise in leading agile teams and testing complex regulatory applications, particularly in AML and CCAR, within the financial sector. Praveen has witnessed the evolution of testing strategies from manual to automated and now AI-assisted testing. He is a thought leader in the industry,

Shailendra Bade, an Engineering Leader with 24 years of experience, holds a Master's in Computer Science and a PG Diploma in Finance. He has led numerous large-scale distributed financial applications, navigating complex regulatory requirements while ensuring high software quality. Shailendra is passionate about exploring innovative testing strategies, including agile practices, test automation, and AI/ML. He actively contributes to the engineering community by sharing his thoughts.