Research Statement

Assessing, Explaining, and Rating AI Models for Trust

Kausik Lakkaraju (kausik@email.sc.edu)

The Past and Present

My research has focused on developing causality-grounded evaluation frameworks that extend beyond traditional accuracy metrics to provide a deeper understanding of how models behave in practice. A central outcome of this work is ARC (AI Rating through Causality) [1], a framework that evaluates robustness by quantifying the causal impact of confounders and interventions on AI models' outcomes. ARC allows researchers and practitioners to systematically compare models under diverse input conditions. I have applied this framework to a range of settings:

- Bias in sentiment analysis systems, showing how causal reasoning can identify and quantify systematic disparities in widely used models [2, 3].
- Time-series forecasting with foundation models, identifying when and why these models succeed or fail [4, 5, 6, 7].
- Large language models, evaluating the fairness and sensitivity of these systems, and exploring principled approaches for testing robustness [8].

Alongside this line of work, I have contributed to trustworthy conversational systems. I co-developed SafeChat [9], a framework that integrates fairness and robustness checks into domain-specific chatbots, outperforming general-purpose LLM-based chatbots in critical areas like personal finance [8]. I have also contributed to AI applications in elections, including datasets and evaluation tools for safe deployment in democratic processes [10, 11, 12].

These efforts have resulted in publications at IJCAI, AAAI, AIES, ICAIF, and IEEE venues, as well as three granted U.S. patents. Together, this body of work establishes a foundation for principled methods of assessing, rating, and explaining AI models for trust.

Future Research Directions

Building on this foundation, my future agenda advances in three interconnected directions:

1. **Mechanistic Interpretability of LLMs**: I want to study LLMs through mechanistic interpretability, reverse-engineering their internal algorithms and circuits. My goal is to identify the concrete features and pathways that drive model behavior, so that weaknesses such as instability to small input changes or hidden bias can be traced back to their underlying causes. This direction will let me connect robustness evaluation with a deeper understanding of how LLMs actually work.
2. **Causality-Driven Hypothesis Testing for Explanations**: I want to develop hypothesis-driven causal approaches for providing better explanations to users and giving them more freedom to explore different hypotheses. For example, how specific input features contribute to outputs under controlled interventions. Such methods can provide more reliable and falsifiable explanations than current post-hoc explanation techniques.
3. **Robustness Certification and Verification:** To bridge causal evaluation and formal methods, I will explore techniques for certifying robustness guarantees of machine learning pipelines. This includes combining causal analysis with verification tools to provide quantitative robustness certificates. The goal is to move from empirical evaluation toward provable claims about model safety and reliability, especially in settings where errors carry high societal or economic cost.

References

1. Lakkaraju, K., Valluru, S. L., Srivastava, B., & Valtorta, M. ARC: A Tool to Rate AI Models for Robustness Through a Causal Lens. In IJCAI 2025 Workshop on User-Aligned Assessment of Adaptive AI Systems.
2. Lakkaraju, Kausik, Biplav Srivastava, and Marco Valtorta. "Rating sentiment analysis systems for bias through a causal lens." IEEE Transactions on Technology and Society (2024).
3. Lakkaraju, K., Gupta, A., Srivastava, B., Valtorta, M., & Wu, D. (2023, November). The Effect of Human v/s Synthetic Test Data and Round-Tripping on Assessment of Sentiment Analysis Systems for Bias. In 2023 5th IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA) (pp. 380-389). IEEE.
4. Widener, M., Lakkaraju, K., Aydin, J., & Srivastava, B. (2025). On identifying why and when foundation models perform well on time-series forecasting using automated explanations and rating. In Proceedings of the AAAI Fall Symposium

on AI Trustworthiness and Risk Assessment for Challenged Contexts (ATRACC 2025). AAAI Press.

5. Lakkaraju, K., Valluru, S. L., & Srivastava, B. (2025). Holistic Explainable AI (H-XAI): Extending Transparency Beyond Developers in AI-Driven Decision Making. arXiv preprint arXiv:2508.05792.
6. Lakkaraju, K., Kaur, R., Zeng, Z., Zehtabi, P., Patra, S., Srivastava, B., & Valtorta, M. (2024). Rating Multi-Modal Time-Series Forecasting Models (MM-TSFM) for Robustness Through a Causal Lens. arXiv preprint arXiv:2406.12908.
7. Lakkaraju, K., Kaur, R., Zehtabi, P., Patra, S., Valluru, S. L., Zeng, Z., ... & Valtorta, M. (2025). On Creating a Causally Grounded Usable Rating Method for Assessing the Robustness of Foundation Models Supporting Time Series. arXiv preprint arXiv:2502.12226.
8. Lakkaraju, K., Jones, S. E., Vuruma, S. K. R., Pallagani, V., Muppasani, B. C., & Srivastava, B. (2023, November). Llms for financial advisement: A fairness and efficacy study in personal decision making. In Proceedings of the Fourth ACM International Conference on AI in Finance (pp. 100-107).
9. Srivastava, B., Lakkaraju, K., Gupta, N., Nagpal, V., Muppasani, B. C., & Jones, S. E. (2025). SafeChat: A Framework for Building Trustworthy Collaborative Assistants and a Case Study of its Usefulness. arXiv preprint arXiv:2504.07995.
10. Lakkaraju, K., Muppasani, B., Jones, S. E., & Srivastava, B. (2025). A Dataset and Visualization of Generalizable Election-Related Questions Compiled from Leading Global Democracies for Building AI-Enabled Tools. In PROMISE–PROMoting AI's Safe usage for Elections (pp. 105-113). Cham: Springer Nature Switzerland.
11. Ayisi, A., Deepak, P., Smith, M., Srivastava, B., Nikolich, A., Hickerson, A., ... & Lakkaraju, K. (2025). Towards Better Elections: A Discussion About the United Kingdom and Africa. In PROMISE–PROMoting AI's Safe usage for Elections (pp. 197-207). Cham: Springer Nature Switzerland.
12. Muppasani, B., Pallagani, V., Lakkaraju, K., Lei, S., Srivastava, B., Robertson, B., ... & Narayanan, V. (2023). On safe and usable chatbots for promoting voter participation. AI Magazine, 44(3), 240-247.