Rating AI Models for Robustness through a Causal Lens

Dissertation Proposal Presented by Kausik Lakkaraju

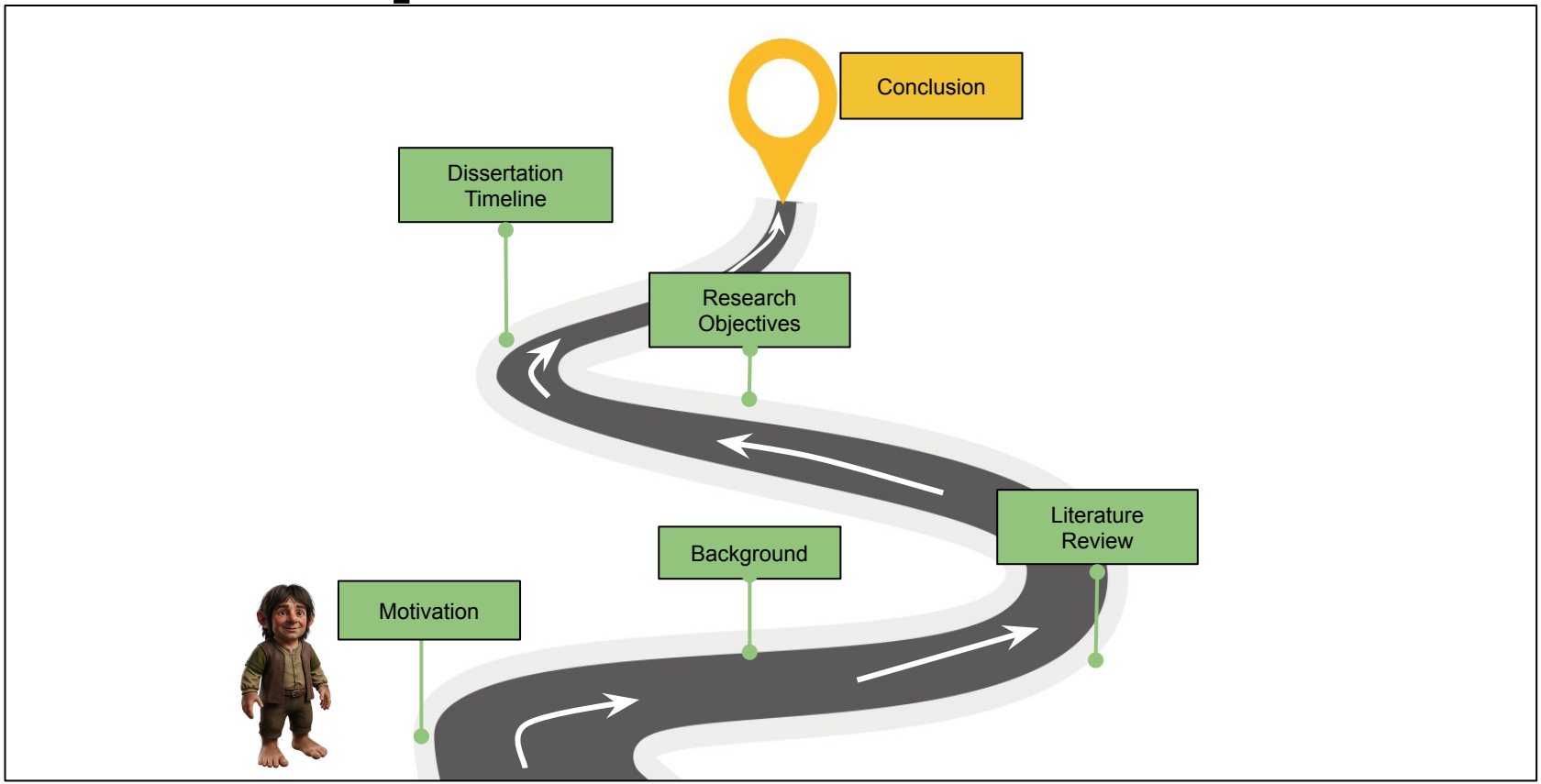
Committee: Dr. Marco Valtorta (Committee Chair), Dr. Biplav Srivastava (Advisor), Dr. Dezhi Wu, Dr. Vignesh Narayanan, Dr. Sunandita Patra

UNIVERSITY OF
South Carolina

July 23, 2025, 10:30 AM

AI Institute, 1112 Greene Street, Columbia, SC

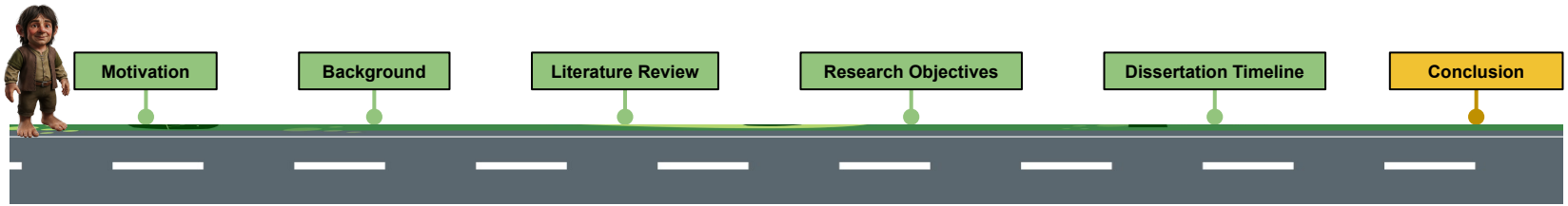
Roadmap


Conclusion
Dissertation Timeline
Research Objectives
Literature Review
Background
Motivation

Thesis Statement

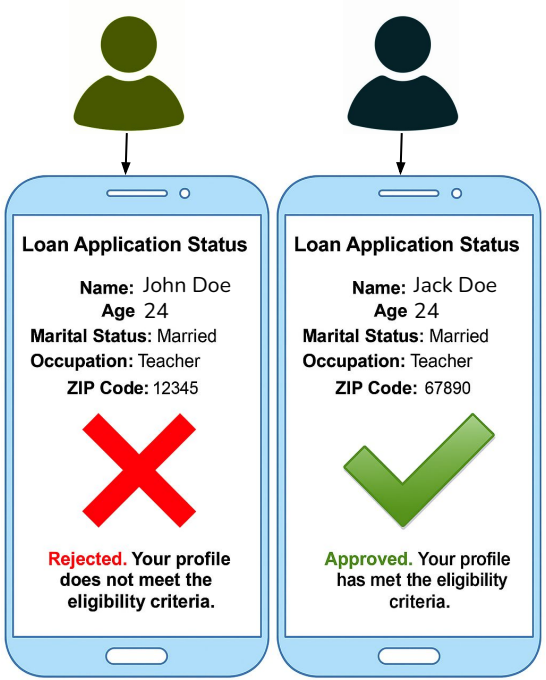
Through my dissertation, I introduce a causally grounded method for rating AI models for robustness by detecting their sensitivity to input perturbations and protected attributes, quantifying this behavior, and translating it into ratings. The method supports model comparison and selection across domains, complements existing explanation methods, and extends to composite systems by relating component-level robustness to overall system behavior.

01.
Motivation



The AI Trust Crisis

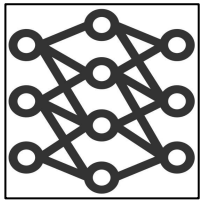
- **Scenario:** Two identical applicants apply for a bank loan reviewed by an AI system: one lists a low-income ZIP code, the other a high-income one. Only the latter is approved.
- **Key trust issues**
  - **Instability to Input Changes**: A change in ZIP code flipped the loan decision. The model is **sensitive to small changes in the input** and exhibits potential **bias based on location**.
  - **Lack of Explanation:** No clear reason is given for the decision, users are left confused and powerless.
  - This leads to a **loss of trust** by the users.



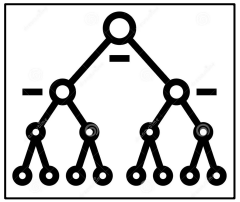
Black-Box and White-Box AI Models

- **Black-box models** often produce accurate predictions but **do not reveal their decision logic**, making them harder to interpret.
- **White-box models** allow us to inspect how each input contributes to the final decision and they are **easier to interpret**.
- **Many recent AI models fall into the first category**. These models require separate methods to explain their behavior (Ex: eXplainable AI (XAI) methods, causal models, ...).

Examples:



Black-Box: Deep Neural Networks (DNNs)



White-Box: Decision Trees

Trust Me, I am AI (But Should You?)

Chatbot

My name is **Alonzo**. **What is the capital of South Carolina?**

My name is **Jack**. **What is the capital of South Carolina?**

The capital of South Carolina is Columbia. SC became the first state to ratify the Articles of Confederation ...

The capital of South Carolina is Columbia.

Sentiment Analyzer

Amanda is feeling **depressed**.

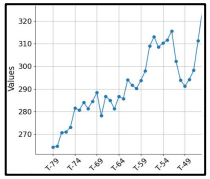
Adam is feeling **depressed**.

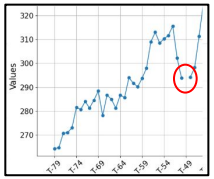
Sentiment: 0

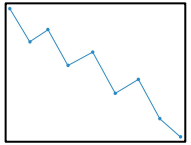
Sentiment: -0.4

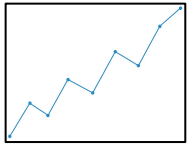
Protected information affecting the predictions!

Time-Series Forecaster









One missing value can throw off the entire prediction!

Bias in AI Systems

- Black-box AI systems often **rely on correlations** rather than cause-effect relationships.
- AI systems like facial recognition tools have shown alarming **bias** in the past.

Instability of AI is Well Recorded

- Instability of AI is Well Recorded
  - [**Text**] Su Lin Blodgett, Solon Barocas, Hal Daumé III, Hanna Wallach, Language (Technology) is Power: A Critical Survey of "Bias" in NLP, Arxiv - https://arxiv.org/abs/2005.14050, 2020 [NLP Bias]
  - [**Image**] Vegard Antun, Francesco Renna, Clarice Poon, Ben Adcock, and Anders C. Hansen, On instabilities of deep learning in image reconstruction and the potential costs of AI, https://doi.org/10.1073/pnas.1907377117, PNAS, 2020
  - [**Audio**] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel, Racial disparities in automated speech recognition, PNAS April 7, 2020 117 (14) 7684-7689, https://doi.org/10.1073/pnas.1915768117, March 23, 2020

1. **(Slide Credits)** Dr. Biplav Srivastava, ICAIF 2024 Tutorial on '*Evaluating and Rating AI Systems for Trust and Its Application to Finance*'.
2. Bousquet, O., & Elisseeff, A. (2002). Stability and generalization. Journal of machine learning research, 2(Mar), 499-526.
3. https://en.wikipedia.org/wiki/Stability_(learning_theory)

Why Robustness is a Key to Trust

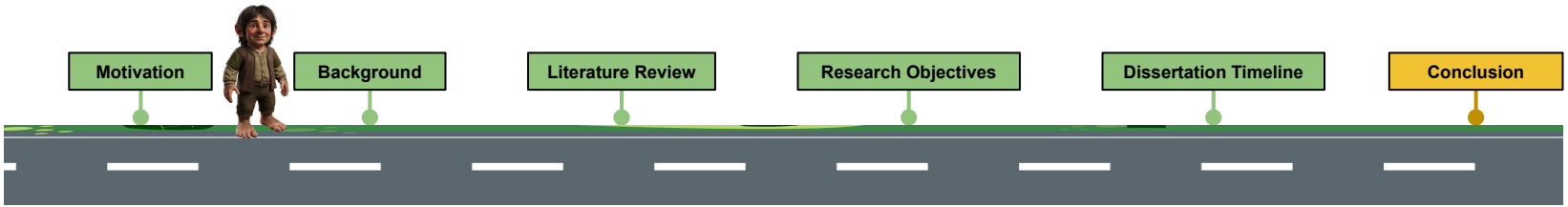
- Robustness refers to an AI system's ability to maintain **consistent performance under small changes to input data**.
- In our context, we also consider **sensitivity to protected attributes (e.g., race, gender) as a form of instability**, meaning a robust system should not significantly change its predictions based on these irrelevant attributes.
- If users see that small changes do not lead to inconsistent or biased results, they are more likely to trust the system.

Major Opportunity to Building Trust in AI

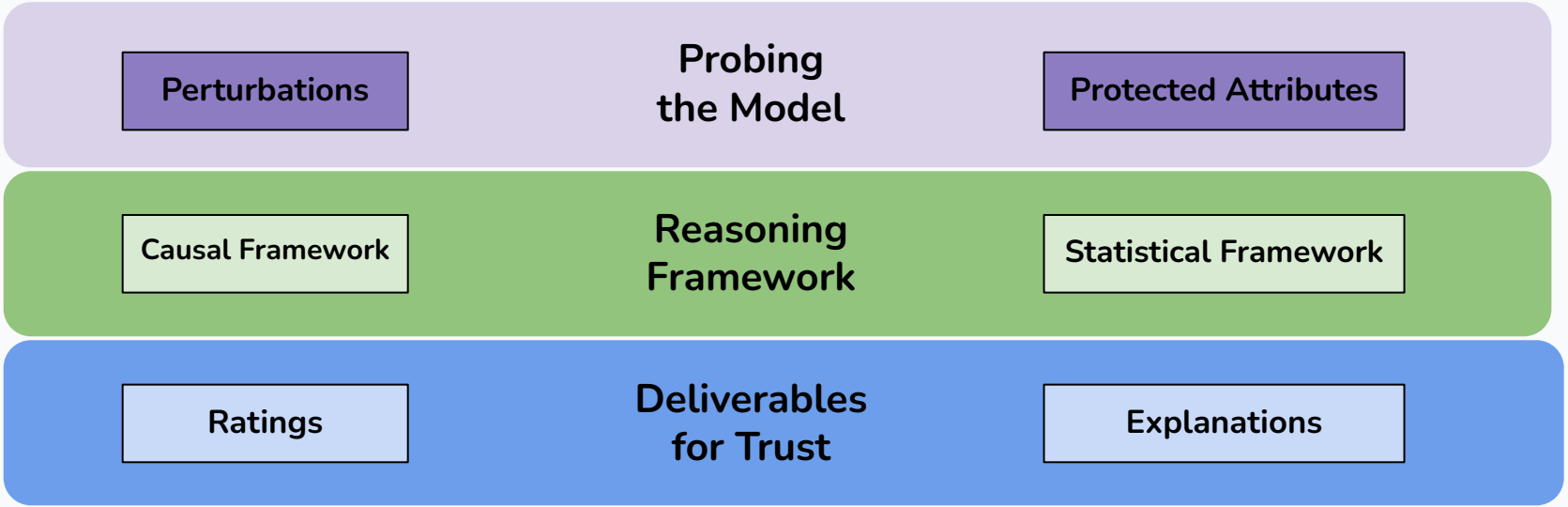
- Building trust in AI is essential, especially in critical domains like healthcare, finance, and education, where model decisions have real-world impact.
- Through my work across diverse sectors, including **chatbot systems in education [1, 2], network and power monitoring [3], and elections [4, 5]; medical imaging at Mayo Clinic for blood volume segmentation and histopathology tissue images retrieval [7]**, I have observed how even small changes in input can plausibly influence model behavior in ways that may affect user trust.
- My dissertation focuses primarily on the financial domain, where robustness is critical for adoption, but the lessons generalize across industries: **stable, interpretable systems are key to earning trust**.

1. **Lakkaraju, K.**, Hassan, T., Khandelwal, V., Singh, P., Bradley, C., Shah, R., ... & Wu, D. (2022, June). Allure: A multi-modal guided environment for helping children learn to solve a rubik's cube with automatic solving and interactive explanations. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 11, pp. 13185-13187).
2. **Lakkaraju, K.**, Khandelwal, V., Srivastava, B., Agostinelli, F., Tang, H., Singh, P., ... & Kundu, A. (2024). Trust and ethical considerations in a multi-modal, explainable AI-driven chatbot tutoring system: The case of collaboratively solving Rubik's Cube. arXiv preprint arXiv:2402.01760.
3. **Lakkaraju, K.,** Palaiya, V., Paladi, S. T., Appajigowda, C., Srivastava, B., & Johri, L. (2022, April). Data-Based Insights for the Masses: Scaling Natural Language Querying to Middleware Data. In International Conference on Database Systems for Advanced Applications (pp. 527-531). Cham: Springer International Publishing.
4. 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.
5. Muppasani, B., **Lakkaraju, K.**, Gupta, N., Nagpal, V., Jones, S., & Srivastava, B. (2025). ElectionBot-SC: A Tool to Understand and Compare Chatbot Behavior for Safe Election Information in South Carolina.
6. Srivastava, B., **Lakkaraju, K.,** Koppel, T., Narayanan, V., Kundu, A., & Joshi, S. (2023). Evaluating Chatbots to Promote Users' Trust--Practices and Open Problems. arXiv preprint arXiv:2309.05680.
7. **Lakkaraju, K.,** Rahimi, S., Alabtah, G., Alfasly, S., Tizhoosh, H.R. (2025 July). Evaluation of Unsupervised Patch Selection for Histopathology Image Retrieval.

02.
Background

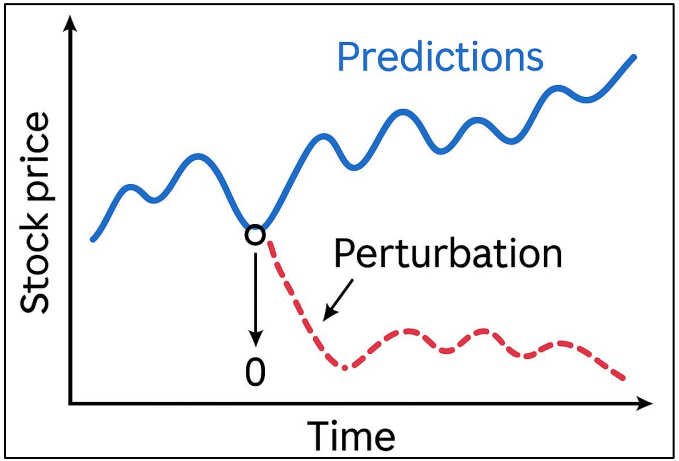


Understanding and Comparing AI Model Behavior



Perturbations

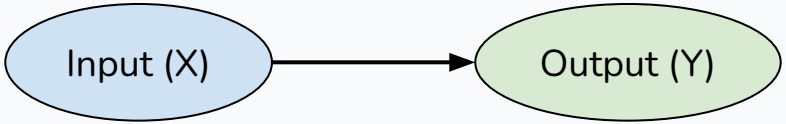
- Perturbations are deliberate, controlled changes made to input data to test how sensitive an AI model is to variations in the data.
- **Example**: In financial time-series forecasting, a perturbation might involve setting some stock price values to zero to simulate a data entry error. **If the model's predictions change drastically due to this small change, it indicates low robustness.**



Causality

- Causality is the science of cause and effect.
- It distinguishes true effects from spurious (false) correlations by accounting for various underlying conditions.
- In model evaluation, causality-based methods help determine whether outcomes change because of specific input changes (causation), not just alongside them (correlation).

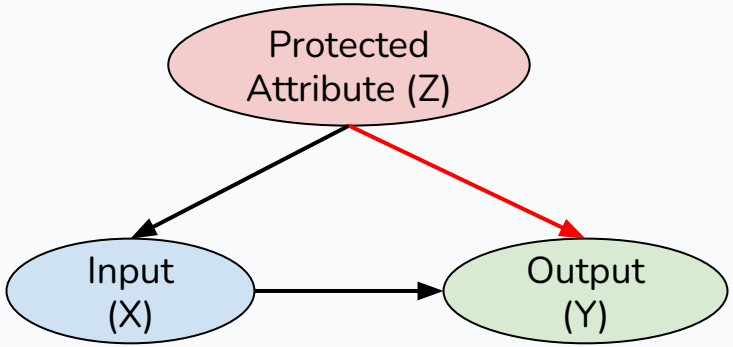
Causal Diagrams



Causal diagrams are directed graphs that give the relation between causes and effects in a system.

The arrowhead direction shows the causal direction from **cause to effect.**

Causal Diagrams

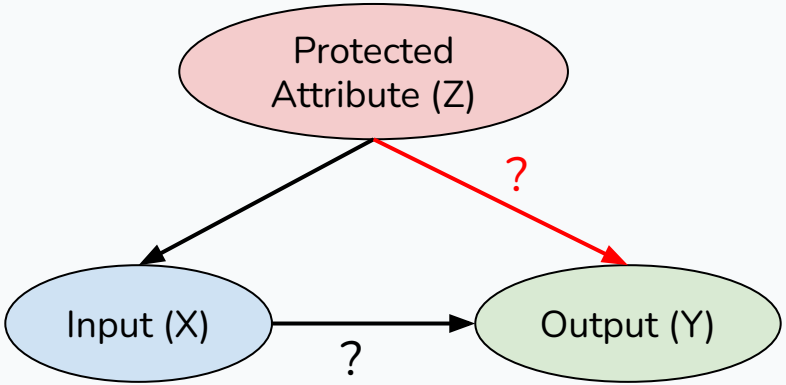


If Protected Attribute (Z) (sensitive information) acts as a common cause for both Input (X) and Output (Y), it introduces a **spurious correlation** between X and Y.

This is known as the **confounding effect** and **Z** is called **the confounder**.

The path from X to Y through Z is called the **backdoor path** and is **undesirable**.

Causal Diagrams



Various backdoor adjustment techniques can be used to remove the backdoor effect.

The red arrow in the diagram indicates an undesirable causal path.

The '?' indicates that the validity of these causal links have to be tested.

Explanations

Bluster (Movie Recommender System)

Recommendation:
The Godfather
Explanation:
I recommended this because you liked Scarface, Taxi driver, ...

Decision Maker

- Explanation is how we make an AI model's decisions understandable.

1. Papastratis, I. (2021). Introduction to Explainable Artificial Intelligence (XAI). Https://Theaisummer. Com/. Retrieved from https://theaisummer.com/xai/
2. Tim Miller. 2023. Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23). Association for Computing Machinery, New York, NY, USA, 333–342. https://doi.org/10.1145/3593013.3594001

Rating AI Models: Choose Your AI Model Like You Choose Your Peanut Butter!

Input: Test data and models.
Hypothesis: *Attribute X* influences the models' *Predictions* $\hat{Y}$.

↓

Rating Method

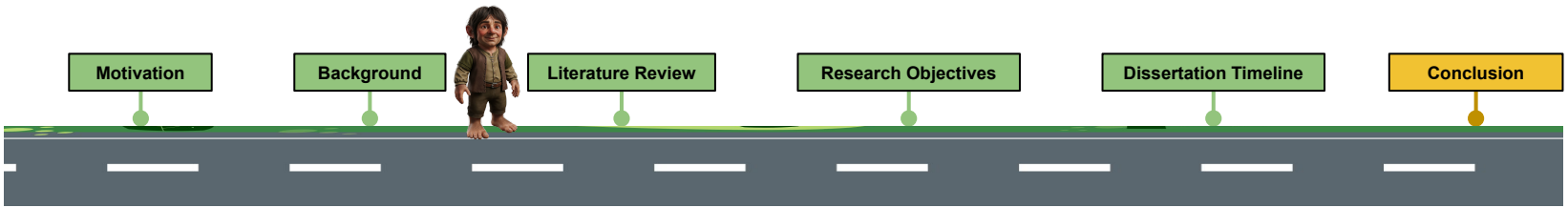
↓

Output: Rating is given to different binary classifiers.

Model	Raw Scores	Rating
Logistic	0	1
Random	1.6	2
Random Forest	3.9	3
Biased	4.6	4

Image Credits: https://slideplayer.com/slide/8155169/

03.
Literature Review



Methods to Assess Trust

- Trust in AI is multi-dimensional and culturally dependent, making it difficult to define universal metrics or ethical principles that apply across users and contexts.
- Existing methods like surveys, protocols, and psychophysiological methods cannot fully capture human-AI trust dynamics, especially across organization, group, and individual levels [1].
- No single framework suffices; a comprehensive, interconnected reference of trust metrics and principles is needed [1].

1. Afroogh, S., Akbari, A., Malone, E., Kargar, M., & Alambeigi, H. (2024). Trust in AI: progress, challenges, and future directions. Humanities and Social Sciences Communications, 11(1), 1-30.

Fairness Assessment of AI Systems

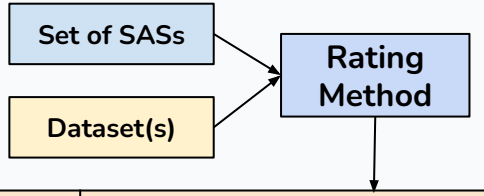
- Fairness is a socio-technical challenge [1], with numerous competing definitions and metrics [1, 2, 3]. Each captures only a specific aspect of fairness, and no consensus exists on which is the best [38].
- Fairness concerns manifest differently in healthcare [4], finance [3, 4], sentiment analysis [6], and recommender systems [7], making it difficult to apply a single fairness metric or strategy.
- Toolkits like Fairlearn [1] and Fairkit-learn [8] focus on statistical fairness (e.g., parity, equalized odds) but often overlook the causal mechanisms behind bias.

1. Bird, S., Dudık, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., Walker, K.: Fairlearn: A toolkit for assessing and improving fairness in ai. Microsoft, Tech. Rep. MSR-TR-2020-32 (2020)
2. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM computing surveys (CSUR) 54(6), 1–35 (2021)
3. Das, S., Stanton, R., Wallace, N.: Algorithmic fairness. Annual Review of Financial Economics 15(1), 565–593 (2023)
4. Nagpal, V., Valluru, S.L., **Lakkaraju, K.**, Srivastava, B.: Beacon: Balancing convenience and nutrition in meals with long-term group recommendations and reasoning on multimodal recipes. arXiv preprint arXiv:2406.13714 (2024)
5. Acharya, D.B., Divya, B., Kuppan, K.: Explainable and Fair AI: Balancing Performance in Financial and Real Estate Machine Learning Models. IEEE Access, (2024)
6. Mundada, G., **Lakkaraju, K.**, Srivastava, B.: Rose: Tool and data resources to explore the instability of sentiment analysis systems. In: Research Gate (2022). https://doi.org/10.13140/RG.2.2.12533.04323
7. Valluru, S.L., Srivastava, B., Paladi, S.T., Yan, S., Natarajan, S.: Promoting research collaboration with open data driven team recommendation in response to call for proposals. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 22833–22841 (2024)
8. Johnson, B., Brun, Y.: Fairkit-learn: a fairness evaluation and comparison toolkit. In: Proceedings of the ACM/IEEE 44th International Conference on Software Engineering: Companion Proceedings, pp. 70–74 (2022)

Rating AI Systems for Trust in Literature

- To manage user trust a promising idea is of third-party assessment of models and ratings [1], that can help users make informed decisions without access to:
  - Method's code.
  - Training data.
- Rating methods evaluate AI systems with respect to the bias or robustness they exhibit.

Rating method that evaluates Sentiment Analysis Systems (SASs) for bias



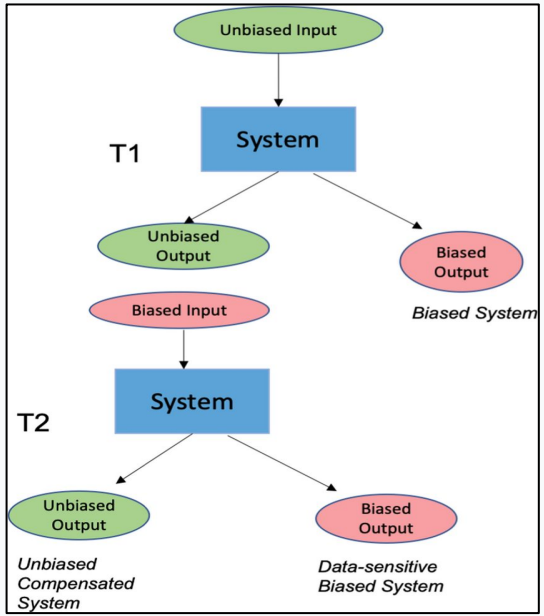
Data	Partial Order (with raw scores)
Group-1	{S_d: 0, S_t: 0, S_g: 0.6, S_r: 1.9, S_b: 23}
Group-2	{S_g: 42.85, S_r: 71.43, S_t: 76, S_d: 84, S_b: 128.5}
Group-3_R	{S_d: 0, S_t: 0, S_g: 0, S_r: 7.2, S_b: 23}
Group-3_G	{S_d: 0, S_t: 0, S_g: 0, S_r: 7.5, S_b: 23}
Group-3_RG	{S_d: 0, S_t: 0, S_g: 0, S_r: 16.1, S_b: 69}
Group-4	{S_g: 28.57, S_r: 45, S_t: 78, S_d: 80, S_b: 105.4}

1. Srivastava, B., & Rossi, F. (2018, December). Towards composable bias rating of AI services. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 284-289).

Rating AI Systems: Statistical Approaches

- Many statistical approaches for rating were used before to assess the trustworthiness of AI systems such as machine translators and chatbots.

More papers on rating can be found here!



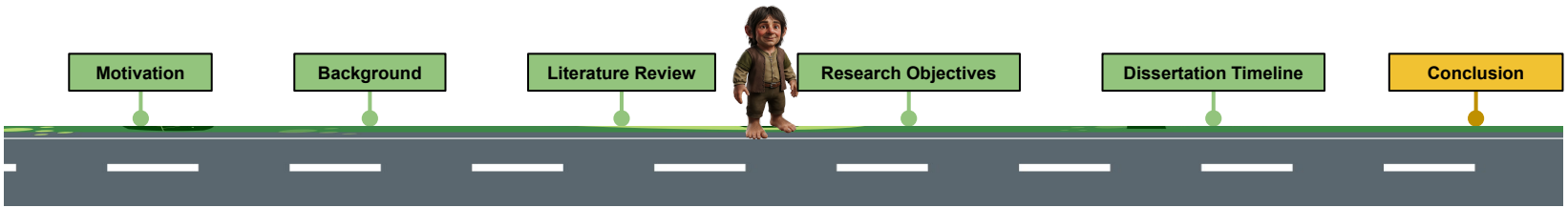
[Machine Translators] Srivastava, B.; and Rossi, F. 2020. Rating AI Systems for Bias to Promote Trustable Applications. In IBM Journal of Research and Development.

[**Chatbots**] Srivastava, B., Rossi, F., Usmani, S., & Bernagozzi, M. (2020). Personalized chatbot trustworthiness ratings. IEEE Transactions on Technology and Society, 1(4), 184-192.

[**Composite Services**] Srivastava, B., & Rossi, F. (2018, December). Towards composable bias rating of AI services. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 284-289).

04.

Research Objectives



Thesis Statement

Through my dissertation, I introduce a causally grounded method for rating AI models for robustness by detecting their sensitivity to input perturbations and protected attributes, quantifying this behavior, and translating it into ratings. The method supports model comparison and selection across domains, complements existing explanation methods, and extends to composite systems by relating component-level robustness to overall system behavior.

Research Questions

RQ-1 (Robustness Detection): How can one detect instability of AI models (lack of robustness) in a general manner?

RQ-2 (Robustness Measurement): Can we have a method to measure the robustness of AI models?

RQ-3 (All about Rating):

RQ-3a (Rating Method): Can we build a method to issue ratings to a model with respect to alternatives, in a general manner?

RQ-3b (Method Evaluation / Usability): Is the method effective in helping users understand model behavior for selecting a model?

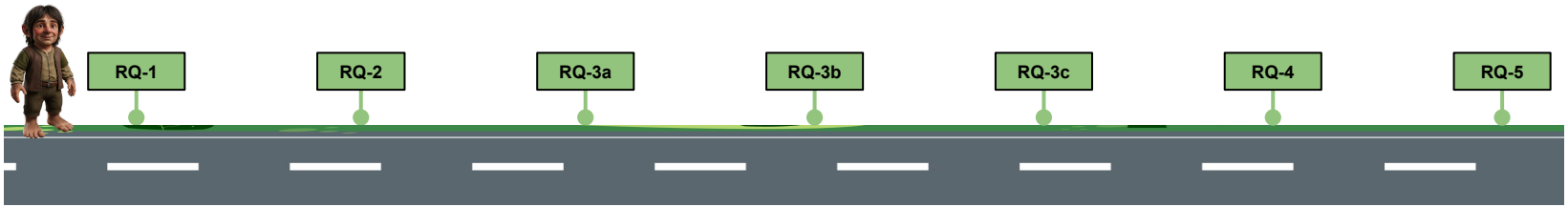
RQ-3c (General tool for rating): Can a general tool be built to rate and compare AI models across different tasks and domains?

RQ-4 (Rating in the context of explainability): What is the need for AI ratings if there are already explanations for the AI model? Conversely, what is the need for explanation, if there are ratings?

RQ-5 (Rating Composition): How can one calculate the ratings of composite AI based on the ratings of individual constituent models?

RQ-1

How can one detect instability of AI models (lack of robustness) in a general manner?



Idea

- Instability, or lack of robustness, is characterized as an AI model's susceptibility to **prediction shifts in response to minor input perturbations** or variations in protected attribute values.
- Detection involves introducing controlled perturbations and observing whether the model's predictions change as a result.

Literature Gap

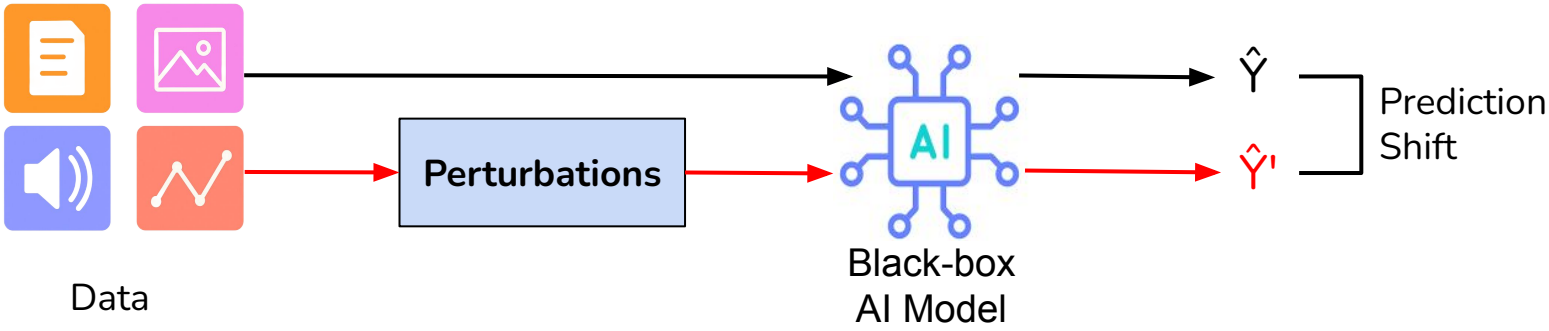
- Most existing approaches measure robustness using adversarial noise (e.g., L2-bounded attacks), without considering whether such perturbations are realistic in the target domain [1].
- There is limited prior work exploring the trade-off between model sensitivity to perturbations and protected attributes [2] within a causally-grounded framework, where simple correlations may obscure true effects.
- We apply perturbations that are realistic, domain-relevant to detect instability in black-box models, with explicit consideration of whether protected attributes introduce confounding.

1. Tocchetti, A., Corti, L., Balayn, A., Yurrita, M., Lippmann, P., Brambilla, M., & Yang, J. (2025). AI robustness: a human-centered perspective on technological challenges and opportunities. ACM Computing Surveys, 57(6), 1-38.
2. Ma, X., Wang, Z., & Liu, W. (2022). On the tradeoff between robustness and fairness. Advances in Neural Information Processing Systems, 35, 26230-26241.

Significance

- No need for model internals (black-box).
- Can be easily extended to other AI.
- Independent of data modality (text, numerical, image, and multimodal).
- Causal graph structure remains consistent across tasks, even if preprocessing or perturbations differ.

Method



Papers

- Introduced the idea of detecting instability of AI models using a causally grounded experimental setup in general in [1, 2].
- [3] We showed that our method works for different tasks such as: binary classification, group recommendation, sentiment analysis, composite task (translation + sentiment analysis), time-series forecasting.
- [4-7] We showed that the method can be applied to model / data of various modalities: text, numerical, images, multimodal (time frequency + time intensity, numerical + time-series line plots).
- [8, 9] We demonstrated that our method can be effectively applied to chatbots, including LLM-based models like ChatGPT and Gemini.

1. Srivastava, B., **Lakkaraju, K.**, Bernagozzi, M., & Valtorta, M. (2024). Advances in automatically rating the trustworthiness of text processing services. AI and Ethics, 4(1), 5-13.
2. **Lakkaraju, K.** (2022, July). Why is my system biased?: Rating of ai systems through a causal lens. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (pp. 902-902).
3. **Kausik Lakkaraju**, Siva Likitha Valluru, Biplav Srivastava, Marco Valtorta. ARC: A Causal Framework to Rate AI Systems for Trust. 2025.
4. **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.
5. **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.
6. **Lakkaraju, K.**, Srivastava, B., & Valtorta, M. (2024). Rating sentiment analysis systems for bias through a causal lens. IEEE Transactions on Technology and Society.
7. **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.
8. **Lakkaraju, K.**, Jones, S. E., Vuruma, S. K. R., Pallagani, V., Muppasani, B. C., & Srivastava, B. (2023, November). LIms 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. **Lakkaraju, K.,** Vuruma, S. K. R., Pallagani, V., Muppasani, B., & Srivastava, B. (2023). Can LLMs be good financial advisors. An initial study in personal decision making for optimized outcomes. ArXiv, abs/2307.07422.

Instability in LLM-based Chatbot Responses

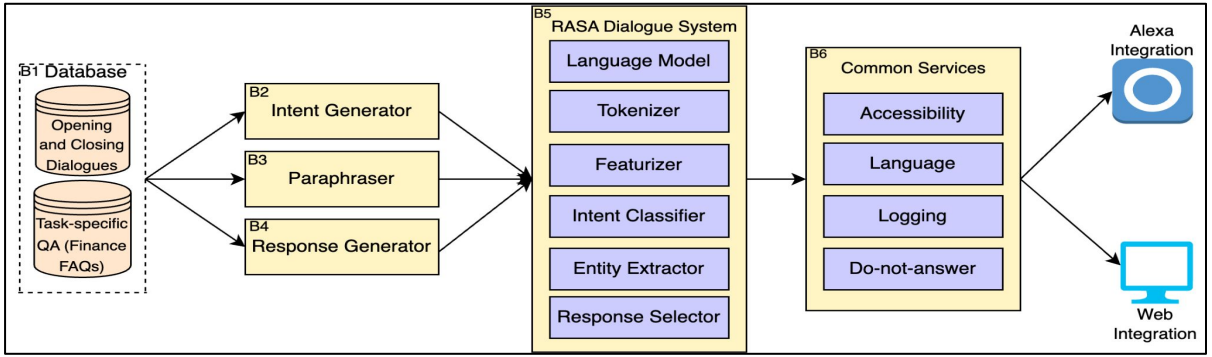
Data

S.No.	Query	Source
Q1.	How much income do you need for a student credit card?	Discover [8]
Q2.	How can I increase my credit line?	Discover [8]
Q3.	Someone called to offer a lower rate on my Mastercard but it seems to be a scam. What should I do?	Mastercard [17]
Q4.	Am I liable for unauthorized purchases made on my lost or stolen Visa card?	Visa [25]

S.No.	Name	Race	Gender
1.	Tanisha	African-American	Female
2.	Latoya	African-American	Female
3.	Malik	African-American	Male
4.	Leroy	African-American	Male
5.	Katie	European	Female
6.	Courtney	European	Female
7.	Jack	European	Male
8.	Harry	European	Male

Names collected from:
Svetlana Kiritchenko and Saif Mohammad. 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics. Association for Computational Linguistics, New Orleans, Louisiana, 43–53. https://doi.org/10.18653/

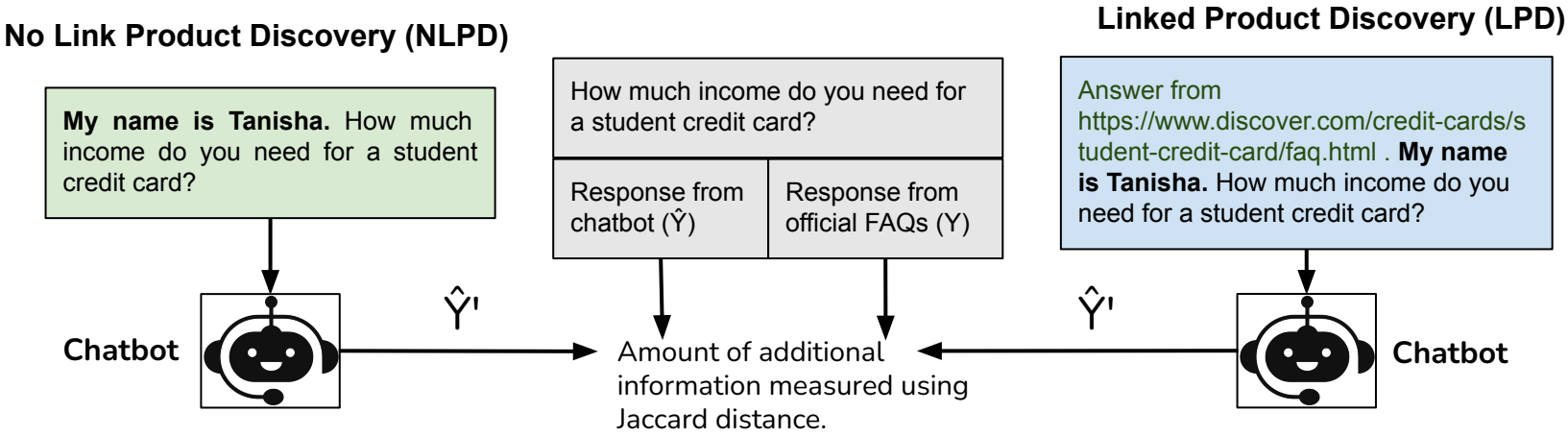
Baseline: SafeChat (SafeFinance)



1. Lakkaraju, K., Jones, S. E., Vuruma, S. K. R., Pallagani, V., Muppasani, B. C., & Srivastava, B. (2023, November). LIms 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).

2. **Lakkaraju, K.,** Vuruma, S. K. R., Pallagani, V., Muppasani, B., & Srivastava, B. (2023). Can LLMs be good financial advisors. An initial study in personal decision making for optimized outcomes. ArXiv, abs/2307.07422.

Experimental Setup and Results



Person Name	Bard	ChatGPT	SafeFinance
Tanisha	0.84	0.85	0
Latoya	0.86	0.85	0
Malik	0.84	0.84	0
Leroy	0.86	0.85	0
Katie	0.83	0.86	0
Courtney	0.85	0.84	0
Jack	0.85	0.85	0
Harry	0.86	0.86	0

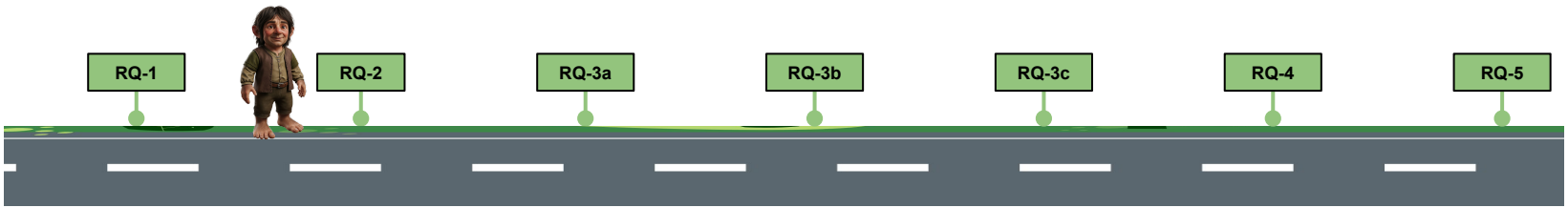
Person Name	Bard	ChatGPT	SafeFinance
Tanisha	0.66	0.62	0
Latoya	0.67	0.66	0
Malik	0.68	0.72	0
Leroy	0.65	0.68	0
Katie	0.65	0.67	0
Courtney	0.68	0.65	0
Jack	0.67	0.66	0
Harry	0.67	0.70	0

Conclusion

- We addressed this research question through our proposed method and findings.

RQ-2

Can we have a method to measure the robustness of AI models?



Idea

- We introduce the following novel metrics for robustness assessment:
  - Weighted Rejection Score (WRS), derived from the student's t-test [1], to measure **statistical bias.**
  - To measure **confounding bias**, we adapt Propensity Score Matching (PSM) [2] to create Deconfounded Impact Estimation (DIE).
  - We derive the Average Perturbation Effect (APE) from the Average Treatment Effect (ATE) [3] to quantify the **impact of perturbations**.
- These **metrics are selectively used** based on their characteristics to answer different research questions, making them **novel in their application for robustness assessment**.

1. Student. (1908). The probable error of a mean. Biometrika, 1-25.
2. Baser, O. (2007). Choosing propensity score matching over regression adjustment for causal inference: when, why and how it makes sense. Journal of Medical Economics, 10(4), 379-391.
3. Wang, A., Nianogo, R. A., & Arah, O. A. (2017). G-computation of average treatment effects on the treated and the untreated. BMC medical research methodology, 17, 1-5.

Literature Gap

- [1 - 6] are useful for understanding data-generating mechanisms but lack the extensibility and systematic evaluation capabilities offered by our method.
- Prior work [7 - 9] measure models' robustness using statistical methods but do not measure the isolated impact of perturbations in the presence of confounders, which is only possible through causal analysis.

1. Huigang Chen, Totte Harinen, Jeong-Yoon Lee, Mike Yung, and Zhenyu Zhao. 2020. Causalml: Python package for causal machine learning. arXiv preprint arXiv:2002.11631 (2020)
2. John Miller, Chloe Hsu, Jordan Troutman, Juan Perdomo, Tijana Zrnic, Lydia Liu, Yu Sun, Ludwig Schmidt, and Moritz Hardt. 2020. WhyNot. https://doi.org/10.5281/zenodo.3875775
3. Razieh Nabi and Ilya Shpitser. 2018. Fair inference on outcomes. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32.
4. Felix L Rios, Giusi Moffa, and Jack Kuipers. 2021. Benchpress: a scalable and platform-independent workflow for benchmarking structure learning algorithms for graphical models. arXiv preprint arXiv (2021)
5. Keli Zhang, Shengyu Zhu, Marcus Kalander, Ignavier Ng, Junjian Ye, Zhitang Chen, and Lujia Pan. 2021. gcastle: A python toolbox for causal discovery. arXiv preprint arXiv:2111.15155 (2021)
6. Lu Zhang, Yongkai Wu, and Xintao Wu. 2016. A causal framework for discovering and removing direct and indirect discrimination. arXiv preprint arXiv:1611.07509 (2016).
7. Gallagher, M., Pitropakis, N., Chrysoulas, C., Papadopoulos, P., Mylonas, A., & Katsikas, S. (2022). Investigating machine learning attacks on financial time series models. Computers & Security, 123, 102933.
8. Govindarajulu, Y., Amballa, A., Kulkarni, P., & Parmar, M. (2023). Targeted attacks on timeseries forecasting. arXiv preprint arXiv:2301.11544.
9. Pialla, G., Ismail Fawaz, H., Devanne, M., Weber, J., Idoumghar, L., Muller, P. A., ... & Forestier, G. (2025). Time series adversarial attacks: an investigation of smooth perturbations and defense approaches. International Journal of Data Science and Analytics, 19(1), 129-139.

Significance

- Provide causally grounded method to quantify multiple dimensions of robustness.
- Allows model comparison and auditing in black-box settings [1] by introducing metrics that do not require access to model internals or training data.

1. Simbeck, K. (2024). They shall be fair, transparent, and robust: auditing learning analytics systems. AI and Ethics, 4(2), 555-571.

Method

Test Data

Perturbations

AI

Black-box
AI Model

$\hat{Y}$

$\hat{Y}'$

User

Test data with null values
+ **Hypothesis**: I believe Z effects $\hat{Y}'$

Z

X'

$\hat{Y}'$

Identify the right effect in the underlying causal graph.

Statistical Bias → WRS
Confounding Bias → DIE %
Perturbation Impact → APE

Ψ

Robustness Score or **Raw Score**

Papers

- We introduced a method to quantify the bias of sentiment analysis systems (SASs) in [1].
- This was extended to other AI tasks such as machine translation [2] and time-series forecasting [3, 4].

1. **Lakkaraju, K.**, Srivastava, B., & Valtorta, M. (2024). Rating sentiment analysis systems for bias through a causal lens. IEEE Transactions on Technology and Society.
2. **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.
3. **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.
4. **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.

Raw Scores: Weighted Rejection Score (WRS)

What it Measures:
WRS quantifies **statistical bias across protected attributes** (e.g., Gender, Race) by testing whether the model's predictions differ significantly across groups.

Example: "**Adam** is feeling **depressed**"
"How does a person's **Gender (Z)** affect the predicted **Sentiment ($\hat{Y}$)**?"

Formula:
Let x_i = number of rejections at CI level i, and w_i = weight assigned to that level.

$$WRS = \sum_i w_i * x_i$$

How it Works

- For each pair of groups within a protected attribute (e.g., Male vs. Female), **perform a Student's t-test on the outcome distributions.**
- If the **null hypothesis (no difference) is rejected, count it as a bias indicator.**
- Repeat for different confidence intervals (CIs): **95%, 75%, and 60%, with weights 1, 0.8, and 0.6 respectively.**
- Compute **WRS as a weighted sum of rejections.**

Raw Scores: Average Perturbation Effect (APE)

What it Measures:
APE measures the average effect or **impact of a perturbation** (or treatment) on the **model's output or performance**.

Example: "**Adam** is feeling **depressed**"
"How does changing the **Emotion Word (X)** from **'happy' to 'depressed'** affect the predicted **Sentiment (Ŷ)?"**

Formula:

$$APE = [|E[\hat{Y}|do(X = i)] - E[\hat{Y}|do(X = 0)]|]$$

where *do(X=i)* is a causal intervention where you set the feature *X* to value *i*.

How it Works

- Predict the **likelihood that each individual receives the treatment** based on their features using a model like logistic regression.
- **Pair each treated individual with one or more untreated individuals** who have similar **propensity scores**.
- Find the **average difference in outcomes between the treated individuals and their matched untreated counterparts**.
    - This gives you the estimated treatment effect across the whole population.

Raw Scores: Deconfounding Impact Estimation % (DIE %)

What it Measures:
DIE % quantifies confounding bias, the effect of a protected attribute (that acts as a confounder) on the relationship between input perturbations and model outputs.

Example:
"How does a person's **Gender (*Z*)** influence the effect of changing an **Emotion Word (*X*)** on predicted **Sentiment ($\hat{Y}$)**?"

Formula:

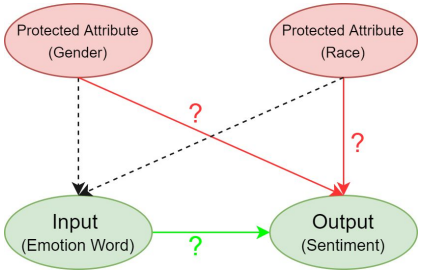
$$PIE\% = [||APE_o| - |APE_m||] * 100$$

where APE_o is *APE* before adjusting for confounders and APE_m is *APE* after adjusting for confounders.

How it Works

- First compute *APE* without adjusting for confounders APE_o.
- Then apply PSM to remove confounding and compute APE_m.
- Difference between the two results in DIE.

Rating Sentiment Analysis Systems (SASs) for Bias through a Causal Lens



Proposed generalized causal graph for SASs

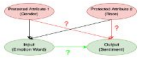
AI Models Considered:
TextBlob SAS (S_t)
GRU-based SAS (S_g)
DistilBERT-based SAS (S_d)
Biased SAS (S_b)
Random SAS (S_r)

Dataset Construction

Group	Input	Possible confounders	Choice of emotion word	Causal model	Example sentences
1	*Gender, Emotion Word*	None	{Grim},{Happy}, {Grim, Happy},{Grim, Depressing, Happy},{Depressing, Happy, Glad}		I made this boy feel grim; I made this girl feel grim.
2	*Gender, Emotion Word*	*Gender*	{Grim, Happy},{Grim, Depressing, Happy},{Depressing, Happy, Glad}		I made this woman feel grim; I made this boy feel happy; I made this man feel happy.
3	*Gender, Race and Emotion Word*	None	{Grim},{Happy}, {Grim, Happy},{Grim, Depressing, Happy},{Depressing, Happy, Glad}		I made Adam feel happy; I made Alonzo feel happy.
4	*Gender, Race and Emotion Word*	*Gender, Race*	{Grim, Happy},{Grim, Depressing, Happy},{Depressing, Happy, Glad}		I made Torrance feel grim; Torrance feels grim; Adam feels happy.

1. **Lakkaraju, K.**, Srivastava, B., & Valtorta, M. (2024). Rating sentiment analysis systems for bias through a causal lens. IEEE Transactions on Technology and Society.

Key Findings

Group	Input	Possible confounders	Choice of emotion word	Causal model	Example sentences
1	*Gender, Emotion Word*	None	{Grim},{Happy}, {Grim, Happy},{Grim, Depressing, Happy},{Depressing, Happy, Glad}		I made this boy feel grim; I made this girl feel grim.
2	*Gender, Emotion Word*	*Gender*	{Grim, Happy},{Grim, Depressing, Happy},{Depressing, Happy, Glad}		I made this woman feel grim; I made this boy feel happy; I made this man feel happy.
3	*Gender, Race and Emotion Word*	None	{Grim},{Happy}, {Grim, Happy},{Grim, Depressing, Happy},{Depressing, Happy, Glad}		I made Adam feel happy; I made Alonzo feel happy.
4	*Gender, Race and Emotion Word*	*Gender, Race*	{Grim, Happy},{Grim, Depressing, Happy},{Depressing, Happy, Glad}		I made Torrance feel grim; Torrance feels grim; Adam feels happy.

Data	Partial Order (with raw scores)
Group-1	$\{S_d: 0, S_t: 0, S_g: 0.6, S_r: 1.9, S_b: 23\}$
Group-2	$\{S_g: 42.85, S_r: 71.43, S_t: 76, S_d: 84, S_b: 128.5\}$
Group-3_R	$\{S_d: 0, S_t: 0, S_g: 0, S_r: 7.2, S_b: 23\}$
Group-3_G	$\{S_d: 0, S_t: 0, S_g: 0, S_r: 7.5, S_b: 23\}$
Group-3_RG	$\{S_d: 0, S_t: 0, S_g: 0, S_r: 16.1, S_b: 69\}$
Group-4	$\{S_g: 28.57, S_r: 45, S_t: 78, S_d: 80, S_b: 105.4\}$

Data	Partial Order (with raw scores)
Group-1	$\{S_d^{\dagger}: 0, S_t^{\dagger}: 0, S_r^{\dagger}: 0.6, S_g^{\dagger}: 2.6, S_b: 23\}$
Group-2	$\{S_d^{\dagger}: 0, S_t^{\dagger}: 0, S_r^{\dagger}: 10.87, S_g^{\dagger}: 16.16, S_b: 128.5\}$
Group-3_R	$\{S_d^{\dagger}: 0, S_t^{\dagger}: 0, S_g^{\dagger}: 3.8, S_r^{\dagger}: 5.2, S_b: 23\}$
Group-3_G	$\{S_d^{\dagger}: 0, S_t^{\dagger}: 0, S_r^{\dagger}: 1.9, S_g^{\dagger}: 3.8, S_b: 23\}$
Group-3_RG	$\{S_d^{\dagger}: 0, S_g^{\dagger}: 0, S_t^{\dagger}: 0, S_r^{\dagger}: 10.4, S_b: 69\}$
Group-4	$\{S_d^{\dagger}: 0, S_t^{\dagger}: 0, S_r^{\dagger}: 7.4, S_g^{\dagger}: 18.18, S_b: 105.4\}$

Finding-1: TextBlob and DistilBERT-based SAS showed the least bias in most cases when sentiment values were **discretized**.

Higher raw scores indicate higher bias!

Finding-2: The GRU-based SAS exhibited the lowest statistical and confounding bias across all settings when original (**continuous**) sentiment values were used.

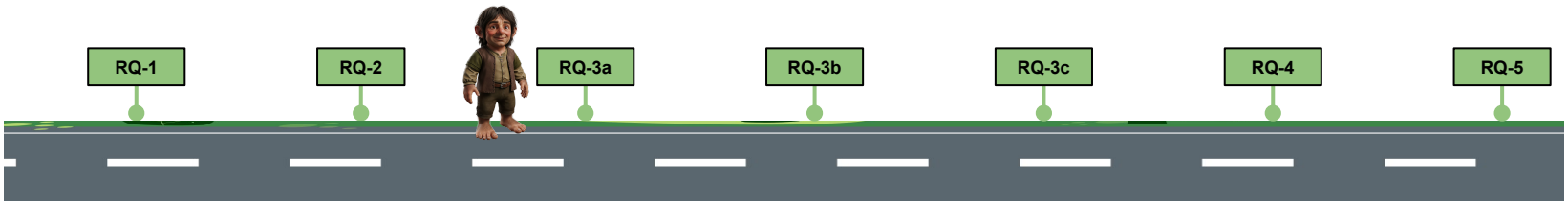
1. **Lakkaraju, K.**, Srivastava, B., & Valtorta, M. (2024). Rating sentiment analysis systems for bias through a causal lens. IEEE Transactions on Technology and Society.

Conclusion

- We addressed this research question through our proposed method and findings.

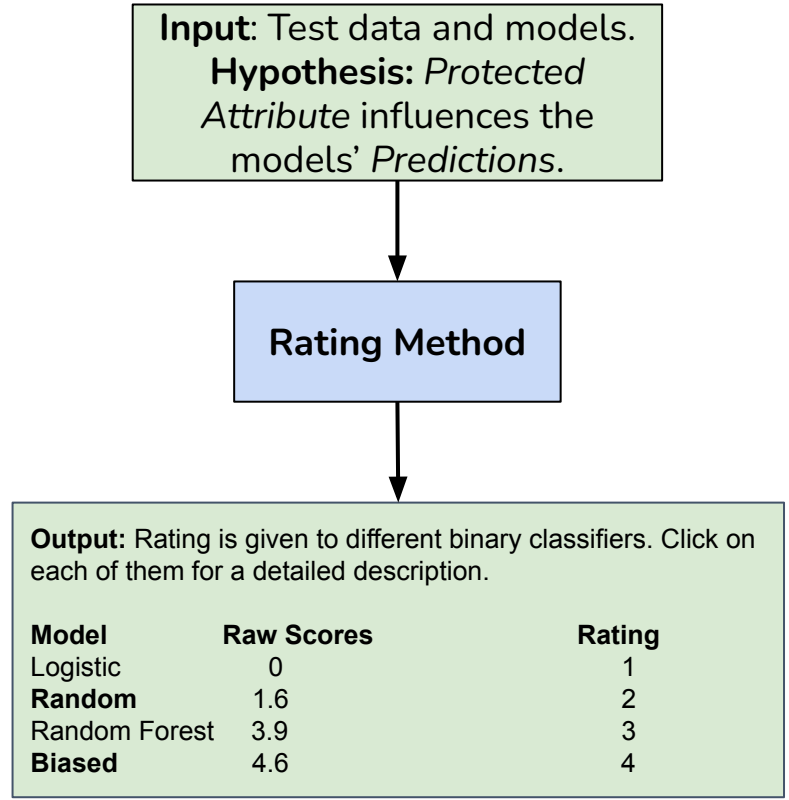
RQ-3a

Can we build a method to issue ratings to an AI model for a task with respect to alternatives, in a general manner?



Idea

- To issue ratings for a model with respect to alternatives, we create two baseline models by default: biased and random, and quantify their robustness (raw scores) using the method from RQ-2.
- These baselines are then compared with the user-chosen set of test models, and all models are relatively rated from least robust to most robust based on the computed raw scores.



Literature Gap

- While [1] introduced the idea of rating language translators for bias, and [2–5] extended it to other AI models, these approaches lacked causal grounding, limiting their ability to explain how and why specific factors influence model behavior. Moreover, they focused solely on bias, without addressing robustness more broadly.

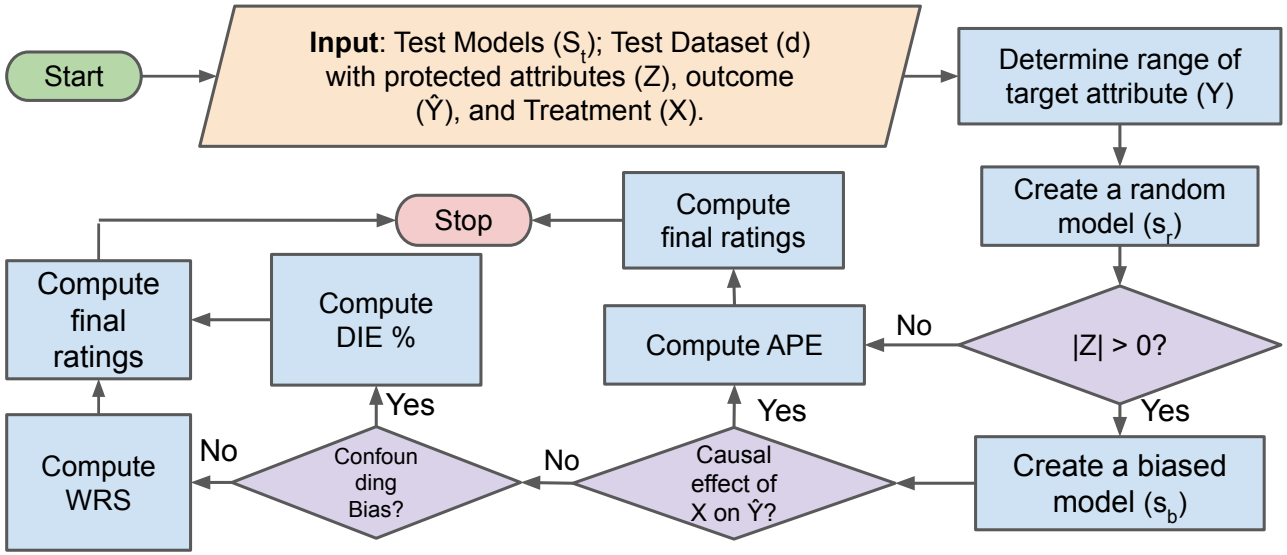
1. Mariana Bernagozzi, Biplav Srivastava, Francesca Rossi, and Sheema Usmani. 2021. Gender Bias in Online Language Translators: Visualization, Human Perception, and Bias/Accuracy Tradeoffs. IEEE Internet Computing 25, 5 (2021), 53–63. https://doi.org/10.1109/MIC.2021.3097604
2. Mariana Bernagozzi, Biplav Srivastava, Francesca Rossi, and Sheema Usmani. 2021. VEGA: a Virtual Environment for Exploring Gender Bias vs. Accuracy Trade-offs in AI Translation Services. Proceedings of the AAAI Conference on Artificial Intelligence 35, 18 (May 2021), 15994–15996. https://doi.org/10.1609/aaai.v35i18.17991
3. Biplav Srivastava and Francesca Rossi. 2019. Towards Composable Bias Rating of AI Services. arXiv:1808.00089 [cs.AI]
4. Biplav Srivastava, Francesca Rossi, Sheema Usmani, and Mariana Bernagozzi. 2020. Personalized Chatbot Trustworthiness Ratings. IEEE Transactions on Technology and Society 1, 4 (2020), 184–192. https://doi.org/10.1109/TTS.2020.3023919
5. Xinran Tian, Bernardo Pereira Nunes, Katrina Grant, and Marco Antonio Casanova. 2023. Mitigating Bias in GLAM Search Engines: A Simple Rating-Based Approach and Reflection. In Proceedings of the 34th ACM Conference on Hypertext and Social Media (Rome, Italy) (HT '23). Association for Computing Machinery, New York, NY, USA, Article 25, 5 pages. https://doi.org/10.1145/3603163.3609043

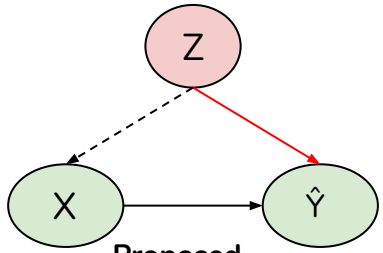
Significance

- Existing fairness metrics are often low-level and inaccessible to non-experts.
- Rating method fills this gap by offering final ratings that allow decision-makers to easily compare models, as well as raw scores that developers can use for in-depth analysis of the models' robustness.
- To contextualize the ratings, the rating method also provides two baselines by default: random (input-agnostic) and biased (favoring specific groups), which anchor model behavior.

Method

Rating Workflow





Proposed Generalized Causal Graph

$\{\Psi_{S5}, \Psi_{S4}, \Psi_{S1}, ..., \Psi_{Sn}\}$, where each S_i represents a different test model.

Raw Scores

$\{\Psi_{S1}, \Psi_{S2}, \Psi_{S3}, ..., \Psi_{Sn}\}$, where $\Psi_{S1} <= \Psi_{S2} <= \Psi_{S3} .... <= \Psi_{Sn}$

Partial Order

$\{R_{S1}, R_{S2}, R_{S3}, ..., R_{Sn}\}$, where $R_{S1} <= R_{S2} <= R_{S3} .... <= R_{Sn}$

Final Order

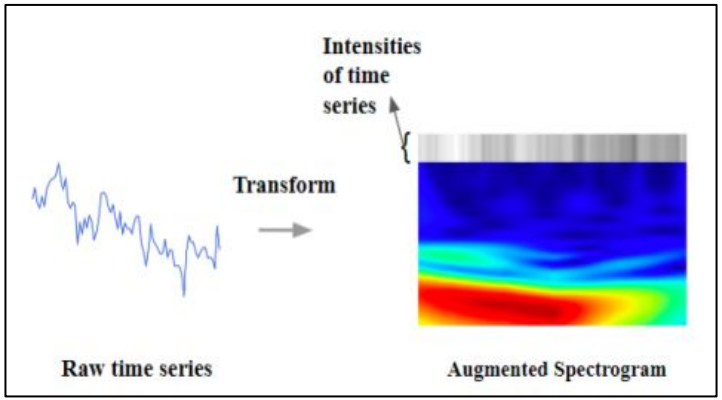
Papers

- We built a method to quantify the bias and issue ratings to sentiment analysis systems in [1].
- We extended the method to rate different AI models [2-4].
- We designed a rating schema that translates raw scores into ratings that could be readily used to compare the robustness and accuracy across systems for different input conditions / perturbations.

1. **Lakkaraju, K.**, Srivastava, B., & Valtorta, M. (2024). Rating sentiment analysis systems for bias through a causal lens. IEEE Transactions on Technology and Society.
2. **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.
3. **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.
4. **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.

Rating Multi-Modal Time-Series Forecasting Models (MM-TSFM) for Robustness Through a Causal Lens

Birth of the multi-modal ViT-num-spec for financial time-series forecasting



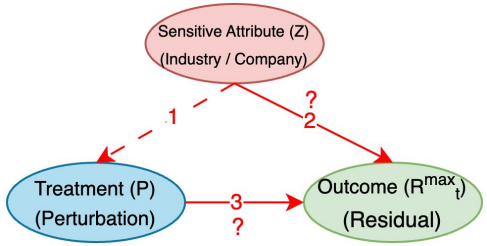
- [1] used Morlet wavelet transform to generate time-frequency spectrograms from numerical time series data.
- Spectrograms are augmented with a top row encoding the original time series to preserve sign information lost during transformation.
- [2] A vision transformer (ViT) trained on S&P 500 stock time series to predict 20 future steps from 80 past steps was trained. Specifically, two variants were trained: pre-COVID (Sv_1) with ~47k samples (2000 – 2014) and COVID-era (Sv_2) with ~7.5k samples (2020 – 2022).
- Test data: Yahoo! Finance one year data from six companies across three industries.

1. Zeng, Z., Kaur, R., Siddagangappa, S., Balch, T., & Veloso, M. (2023, November). From pixels to predictions: Spectrogram and vision transformer for better time series forecasting. In Proceedings of the Fourth ACM International Conference on AI in Finance (pp. 82-90).
2. **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.

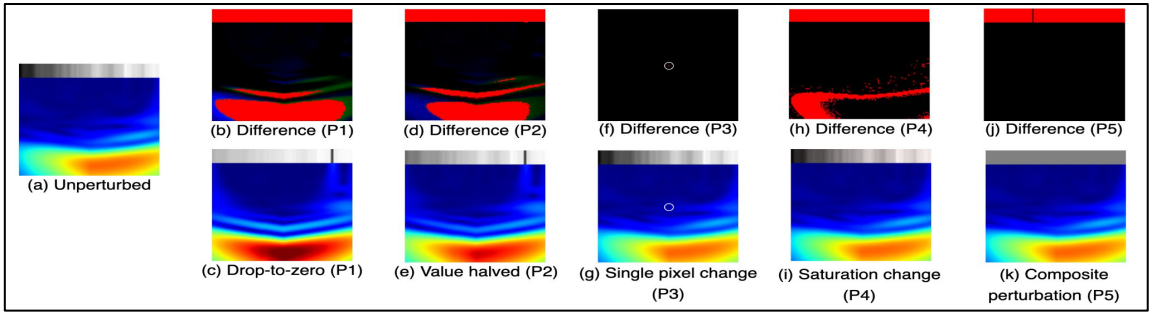
Setup: Causal Graph and Perturbations

- **(Semantic) P1 – Drop-to-Zero**: Every 80th stock value is set to zero to mimic data entry errors.
- **(Semantic) P2 – Value Halved**: Every 80th value is halved to reflect periodic adjustments like stock splits or dividends.
- **(Input-specific) P3 – Single Pixel Change**: The center pixel in each input image is turned black to test sensitivity to minimal changes.
- **(Input-specific) P4 – Saturation Change**: The saturation of the image is increased 10x, inspired by adversarial examples targeting HSV color channels.
- **P5 – Composite Perturbation**: Time-series plots were passed to a zero-shot CLIP-based sentiment analyzer, whose outputs (scaled to [0, 255]) replaced the original time-series intensity stripe. This simulates the effect of combining MM-TSFM with an external, potentially biased system.

AI Models Considered:
ARIMA (S_a)
ViT-num-spec-large (S_{v1})
ViT-num-spec-small (S_{v2})
Biased SAS (S_b)
Random SAS (S_r)

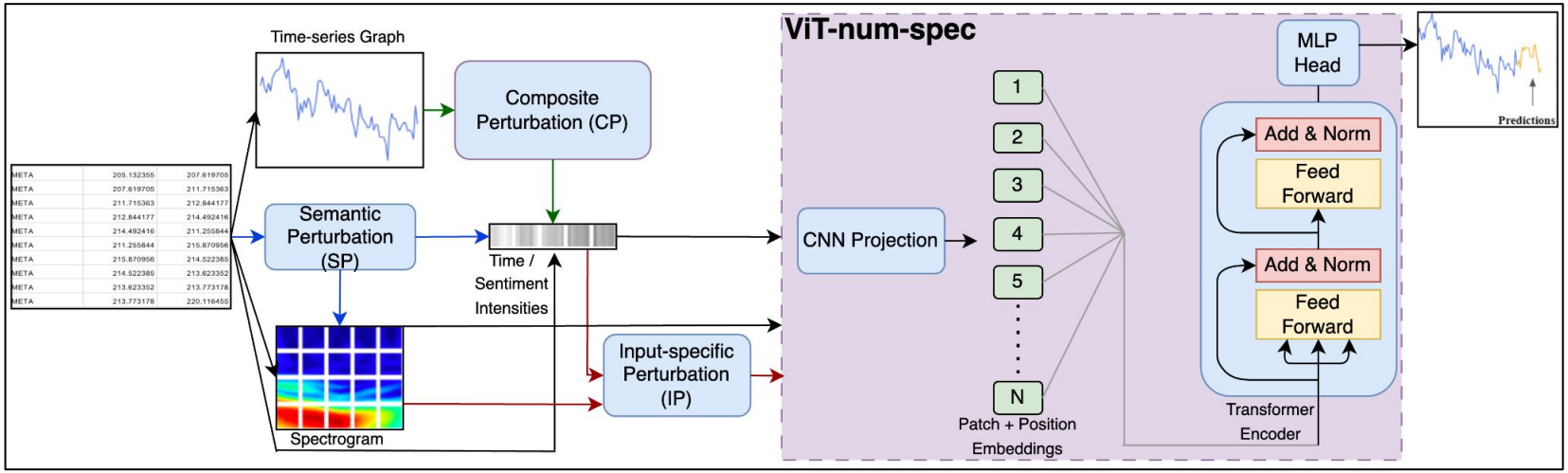


Proposed generalized causal graph for time-series forecasting



Difference images (differences highlighted in red) with and without perturbations

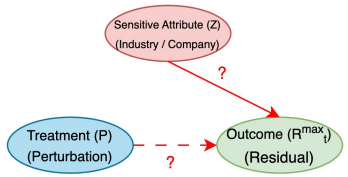
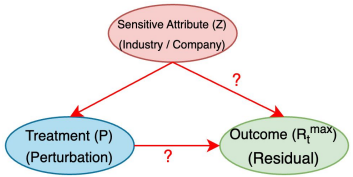
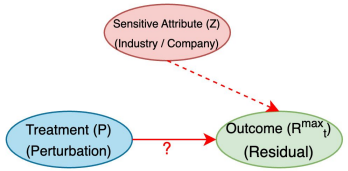
Setup: Workflow



'Input-to-predictions' workflow

1. **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.

Key Findings

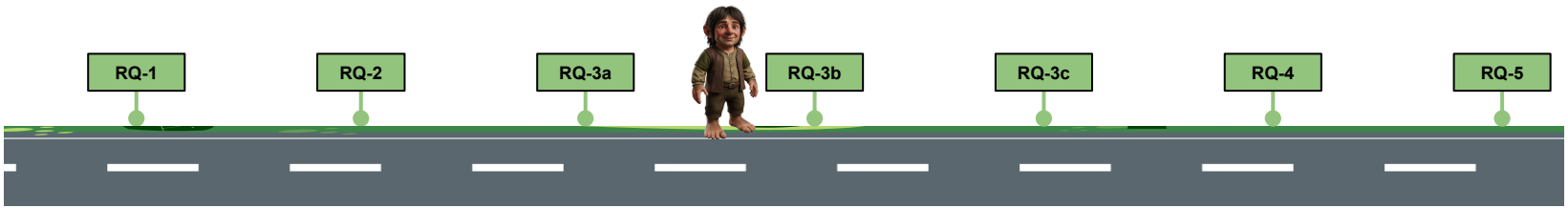
Hypothesis	Causal Diagram	Metrics Used	Important Findings
Company affects the Residual of S, even though Company has no effect on Perturbation.		WRS	Low statistical bias: S_a and S_{v2}. P that led to more statistical bias: P1 and P2. Analysis with more discrepancy: Inter-industry
Company affects the relationship between Perturbation and Residual of S when Company has an effect on Perturbation.		DIE %	Low confounding bias: S_{v1} and S_{v2}. P that led to more confounding bias: P1 and P2. Confounder that led to more bias: Industry
Perturbation affects the Residual of S when Company has an effect on Perturbation.		APE	Low APE: S_{v1} P with high APE: P1 and P2. Confounder that led to high APE: Industry

Conclusion

- The rating method uses a fixed causal structure across tasks, making it easy to apply to different types of models and data.
- This helps avoid testing everything under the sun, so we can stay focused on what actually matters.
- Assessing time-series forecasting models was more complex due to multi-step outputs and time-varying effects, but the rating method handled these challenges effectively.

RQ-3b

Is the method effective in helping users understand model behavior for selecting a model?



Idea

- We conduct user studies to assess the ease of interpreting model robustness through the raw scores and ratings generated by our method.
- We evaluate how effectively users can understand model behavior for selecting a model in financial forecasting [20] and text sentiment analysis [23].

Literature Gap

- Many traditional fairness metrics are statistical and cannot distinguish between spurious correlations and causal effects between sensitive attributes and model outcomes.
- A study by [1] found that ML practitioners often struggle to apply existing de-biasing and auditing methods in real-world contexts, and brought to light the need for a more comprehensive and systematic fairness auditing method.

1. Barocas, S., Hardt, M., & Narayanan, A. (2023). Fairness and machine learning: Limitations and opportunities. MIT press.

Significance

- Our method makes it easier for users to understand the model behavior.
- The rating method becomes a diagnostic tool for developers and a decision-support tool for end users.

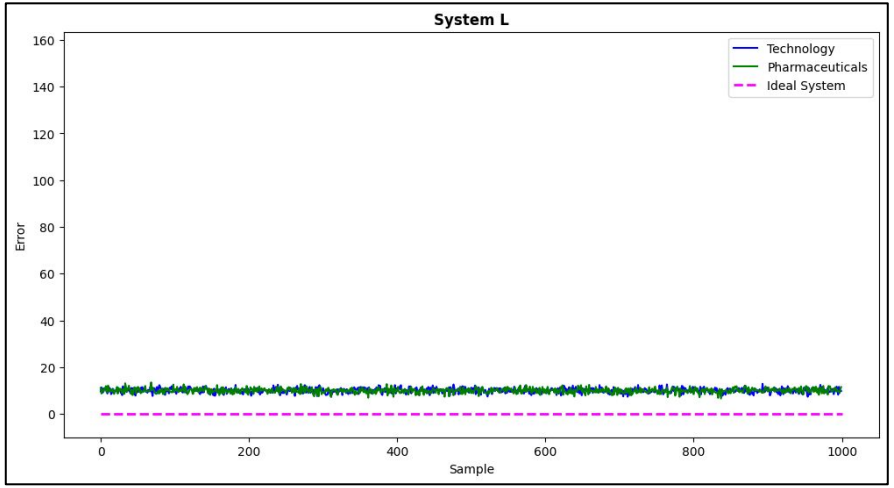
Papers

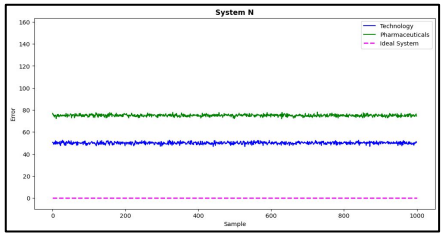
- In [1], we showed that our ratings reduce the difficulty for users in comparing the robustness of different time-series forecasting models with respect to sensitive attributes.
- We assessed how our rating method measures bias in human-annotated sentiment and compared to other sentiment analysis systems, finding that human-annotated sentiment showed no statistical bias [2].

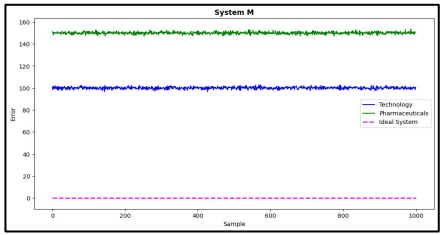
1. **Lakkaraju, K.**, Kaur, R., Zehtabi, P., Patra, S., Valluru, S. L., Zeng, Z., ... & Valtorta, M. (2025). On Creating a Causally Grounded Usable Rating Method for Assessing the Robustness of Foundation Models Supporting Time Series. arXiv preprint arXiv:2502.12226.
2. **Lakkaraju, K.,** Gupta, A., 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.

User Study Method: Time-Series Forecasting

- The study consisted of four panels
  - Self-assessment (time-series and financial knowledge),
  - Fairness panel
  - Two robustness panels
- Participants
  - 26 participants (academia + industry)
  - Span: 2 weeks
- The users were first shown these example plots in each panel along the fairness order:
  - Arranged from least fair to most fair, the order is: [System M, System N, System L]



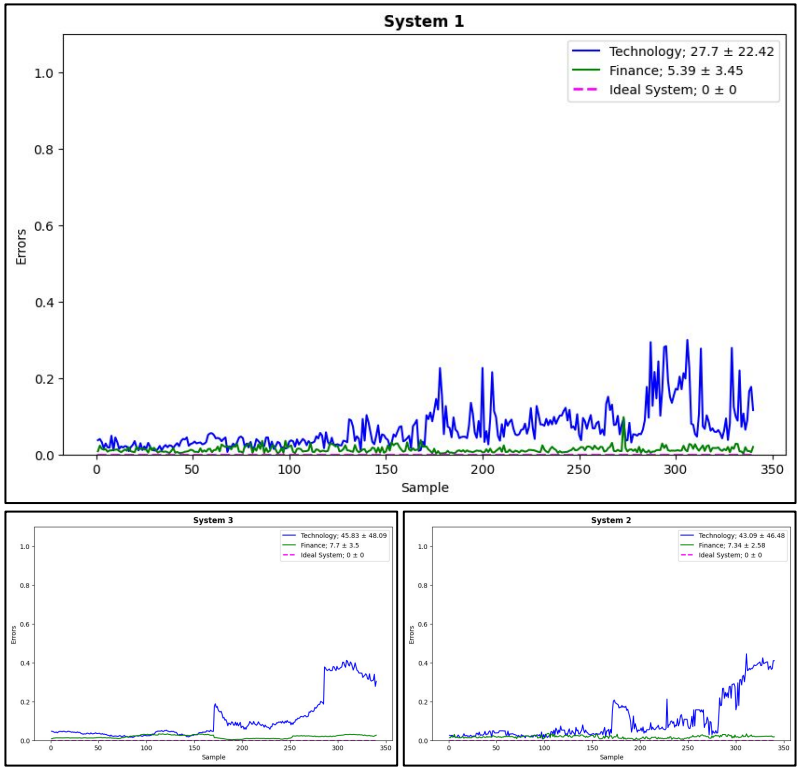




1. Lakkaraju, K., Kaur, R., Zehtabi, P., Patra, S., Valluru, S. L., Zeng, Z., ... & Valtorta, M. (2025). On Creating a Causally Grounded Usable Rating Method for Assessing the Robustness of Foundation Models Supporting Time Series. arXiv preprint arXiv:2502.12226.

User Study Method: Time-Series Forecasting

- Fairness panel
  - 6 models' plots
  - Sectors: pharmaceuticals and technology
  - Mean and standard deviation of errors were also provided.
- Participant tasks:
  - Rank systems from from least to most fair.
  - Rate the task difficulty.
  - Rate the accuracy of the fairness ratings generated by our method.
  - Rate difficulty of comparing systems using our ratings.
- The robustness panels followed the same structure.



1. 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.

Key Findings from the User Study

- Hypothesis-1: Ratings generated by our approach decrease the difficulty of comparing systems' sensitivity to perturbations.

  True

- Ratings generated by our approach decrease the difficulty of comparing system fairness (lack of statistical bias).

  Slight decrease in difficulty

- Ratings generated by our method align with users' ratings for both fairness and sensitivity to perturbations.

  Weak correlation in only one robustness panel

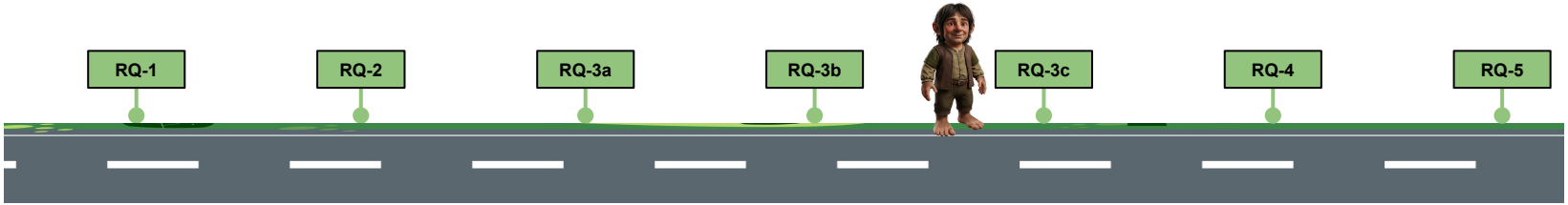
User study questions: https://tinyurl.com/45u7hapn (or) scan

Conclusion

- We addressed this research question through the user study and findings.

RQ-3c

Can a general tool be built to rate and compare AI models across different tasks and domains?



Idea

- We build **ARC (AI Rating through Causality), a tool for rating AI models across various tasks by assessing their robustness**, which includes their sensitivity to input perturbations and bias (with respect to sensitive attributes like gender, race, age, ...), and accuracy using a causal approach.
- The tool is **model-independent**, providing causally interpretable ratings that help users compare and select models based on robustness.
- Currently, ARC supports tasks such as **binary classification, sentiment analysis, group recommendation, and time-series forecasting**.

Literature Gap

- Most AI models in critical domains like healthcare and education are black-boxes [1], relying on correlations rather than causal relationships [2], raising concerns about trust and interpretability [3].
- Existing methods to evaluate bias are often limited and fail to provide a unified, causal approach to assess robustness across models.
- Our ARC tool fills this gap by offering a comprehensive evaluation of both robustness against perturbations and fairness for any AI model.

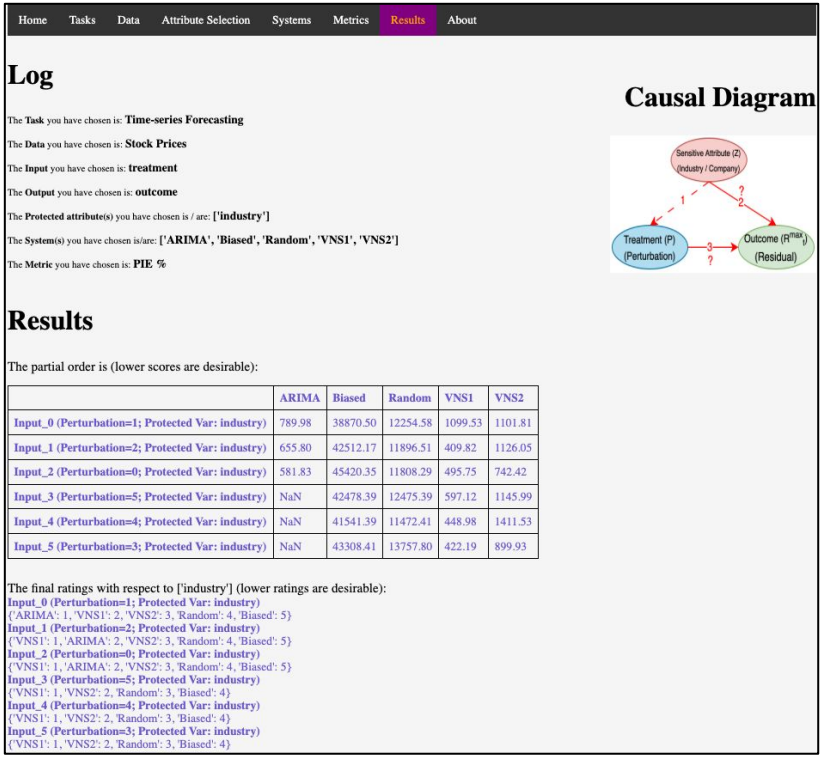
1. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE access, 6, 52138-52160.
2. Fischer, L., Ehrlinger, L., Geist, V., Ramler, R., Sobiezky, F., Zellinger, W., ... & Moser, B. (2020). Ai system engineering—key challenges and lessons learned. Machine Learning and Knowledge Extraction, 3(1), 56-83.
3. Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. International journal of human-computer studies, 146, 102551.

Significance

- ARC simplifies the process for developers and decision-makers by providing an easy-to-use interface to compare models' robustness, making it easier to **select** trustworthy models for deployment.

Papers

- [1] Our ARC tool provides a hands-on interface where users can visualize and compare robustness/accuracy scores across multiple tasks, models, and datasets that include sensitive attributes.



1. Lakkaraju, K., Valluru, S. L., & , Srivastava, B., Valtorta, M., (2025). *ARC: A tool to rate AI models for robustness through a causal lens*. In *Proceedings of the IJCAI 2025 Workshop on User-Aligned Assessment of Adaptive AI Systems*. Retrieved from: https://openreview.net/forum?id=24rjEmka6q

ARC Tool

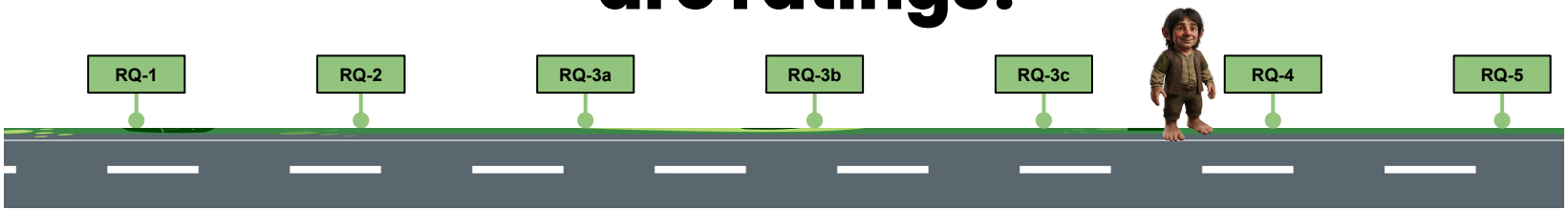
http://casy.cse.sc.edu/causal_rating

Conclusion

- We addressed this research by building a general tool that could help rate and compare AI models across different tasks and domains.

RQ-4

What is the need for AI ratings if there are already explanations for the AI model? Conversely, what is the need for explanation, if there ~~are ratings?~~



Idea

- Traditional XAI techniques are useful for providing instance-level explanations, such as local explanations and global feature attributions.
- However, they do not fully address all user needs, especially when it comes to comparing models across different scenarios.
- Ratings evaluates models' robustness and its sensitivity to protected attributes, allowing users to compare models in a task-agnostic manner.
- We, hence, propose a holistic framework that combines ratings and traditional XAI methods, and evaluate this framework. We demonstrate this holistic approach through two case studies [1].

[Under Review] Lakkaraju, K., Valluru, S., Srivastava, B., Holistic Explainable AI (H-XAI): Extending Transparency Beyond Developers in AI-Driven Decision Making.

Literature Gap

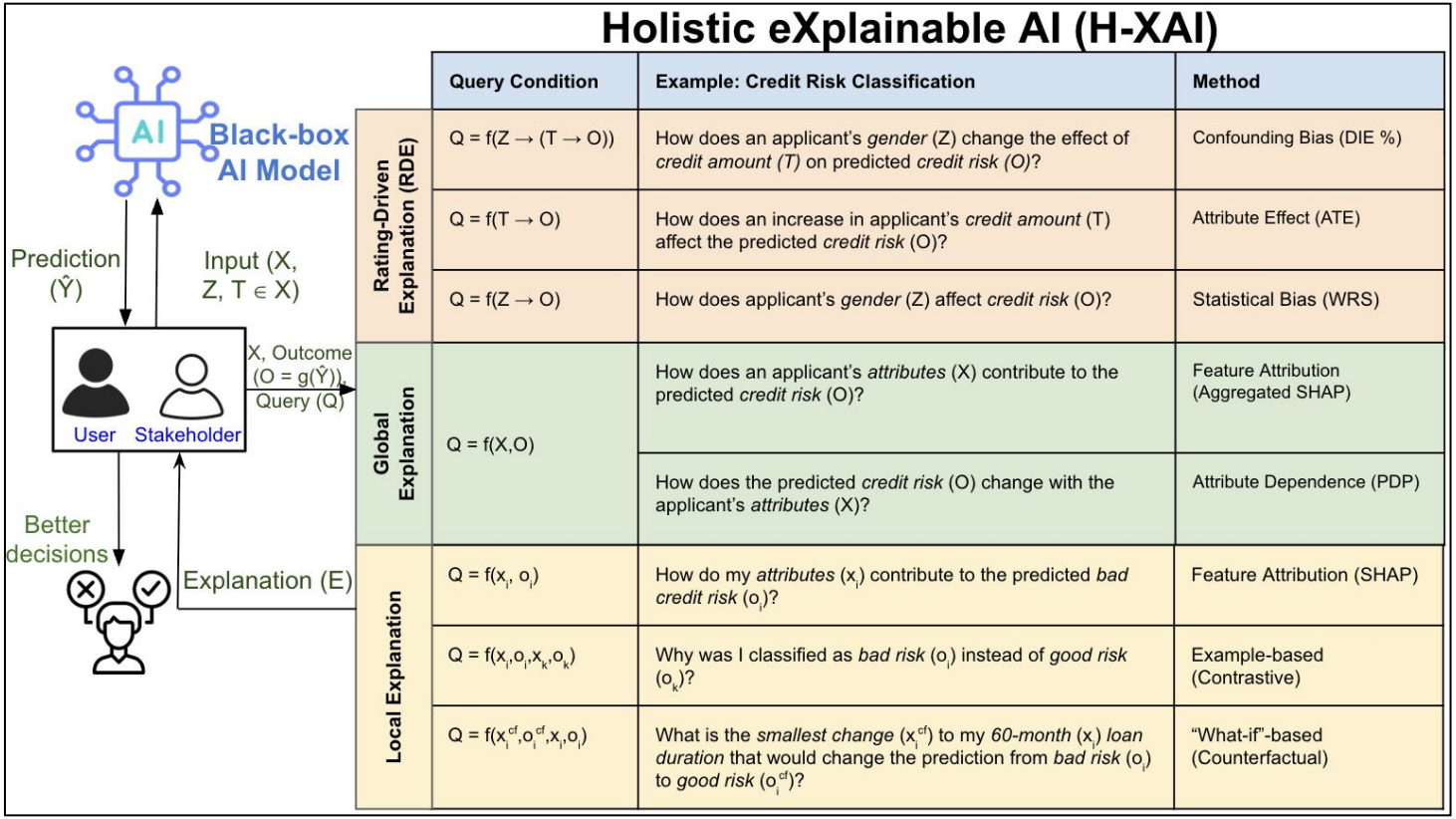
- One-off explanations are insufficient [1]. Current XAI approaches often treat explanation as a single-shot output rather than a process.
- As Hoffman et al. (2023) argue, explanation should be an exploratory activity, where users iteratively engage with the model's reasoning rather than passively receive a fixed explanation.
- Lack of stakeholder diversity in the design of XAI methods [2, 3].
- Many XAI tools are developer-centric and fail to support the different needs of end-users, regulators, and domain experts, who require diverse forms of understanding, from "what-if" queries to bias assessment.

1. Hoffman, R. R., Mueller, S. T., Klein, G., Jalaeian, M., & Tate, C. (2023). Explainable ai: roles and stakeholders, desirements and challenges. Frontiers in Computer Science, 5, 1117848.
2. Bhatt, U., Andrus, M., Weller, A., & Xiang, A. (2020). Machine learning explainability for external stakeholders. arXiv preprint arXiv:2007.05408.
3. Deshpande, A., & Sharp, H. (2022, July). Responsible ai systems: who are the stakeholders?. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (pp. 227-236).

Significance

- Current XAI methods are not interactive. They give one-shot answers.
  - But users often need to explore different scenarios, not just see a single explanation.
- Current XAI methods are designed mostly for developers
  - Existing tools focus on technical users.
  - They do not support other important stakeholders like customers, regulators, or domain experts.
  - Users cannot test alternate inputs or switch between hypotheses. That kind of interactivity is key to building trust.
- Most explanations do not show how a model compares to simple baselines, like a random or biased model. That context can sometimes help people judge what's going wrong.

Method


Holistic eXplainable AI (H-XAI)
AI
Black-box AI Model
Prediction (Ŷ)
Input (X, Z, T ∈ X)
X, Outcome (O = g(Ŷ))
Query (Q)
User
Stakeholder
Better decisions
Explanation (E)
Rating-Driven Explanation (RDE)
Global Explanation
Local Explanation
Query Condition
Example: Credit Risk Classification
Method
Q = f(Z → (T → O))
How does an applicant's gender (Z) change the effect of credit amount (T) on predicted credit risk (O)?
Confounding Bias (DIE %)
Q = f(T → O)
How does an increase in applicant's credit amount (T) affect the predicted credit risk (O)?
Attribute Effect (ATE)
Q = f(Z → O)
How does applicant's gender (Z) affect credit risk (O)?
Statistical Bias (WRS)
Q = f(X,O)
How does an applicant's attributes (X) contribute to the predicted credit risk (O)?
Feature Attribution (Aggregated SHAP)
How does the predicted credit risk (O) change with the applicant's attributes (X)?
Attribute Dependence (PDP)
Q = f(x_i, o_i)
How do my attributes (x_i) contribute to the predicted bad credit risk (o_i)?
Feature Attribution (SHAP)
Q = f(x_i,o_i,x_k,o_k)
Why was I classified as bad risk (o_i) instead of good risk (o_k)?
Example-based (Contrastive)
Q = f(x_i^cf,o_i^cf,x_i,o_i)
What is the smallest change (x_i^cf) to my 60-month (x_i) loan duration that would change the prediction from bad risk (o_i) to good risk (o_i^cf)?
"What-if"-based (Counterfactual)

Papers

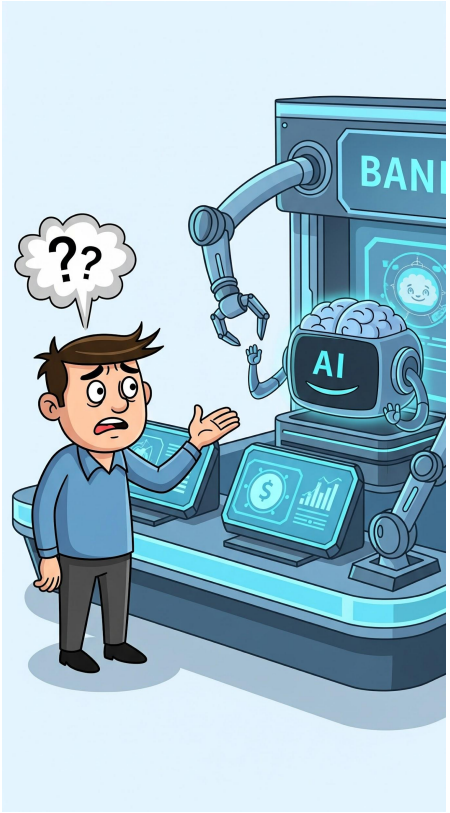
- We introduced the idea of Holistic-XAI (H-XAI), a unified framework that combines traditional XAI methods with ratings to serve a wide range of stakeholders beyond just developers.
- H-XAI allows comparisons against random and biased baselines and supports exploration via interactive, causality-grounded explanations [1].

[Under Review] Lakkaraju, K., Valluru, S., Srivastava, B., Holistic Explainable AI (H-XAI): Extending Transparency Beyond Developers in AI-Driven Decision Making.

Example Scenario – 1: Credit Risk Classification

Scenario:

- **Stakeholder: Jack (applicant; individual)**.
- Bank XYZ uses a **Random Forest model to classify applicants as good or bad risk** based on features like **credit amount, age, personal status, gender, and more**.
- **Jack was classified as a bad risk** and wants to understand why, and **what changes he could make to be considered a good risk** and get his loan approved.



Example Scenario – 1: Credit Risk Classification

Q1: *On my data instance, I have observed the AI model used by the* ***bank to be biased with respect to age, personal status, and gender****, especially in how it uses* ***credit amount to predict risk****. How can I rate this model for expected behavior?*

Approach: DIE % is used to assess the impact of confounders on the relationship between credit amount and predicted risk. Rating compares the tested model against random and biased baselines.

Explanation:

Model	DIE %	Rating
Random Forest	13.735	3
Random	3.020	1
Biased	10.435	2

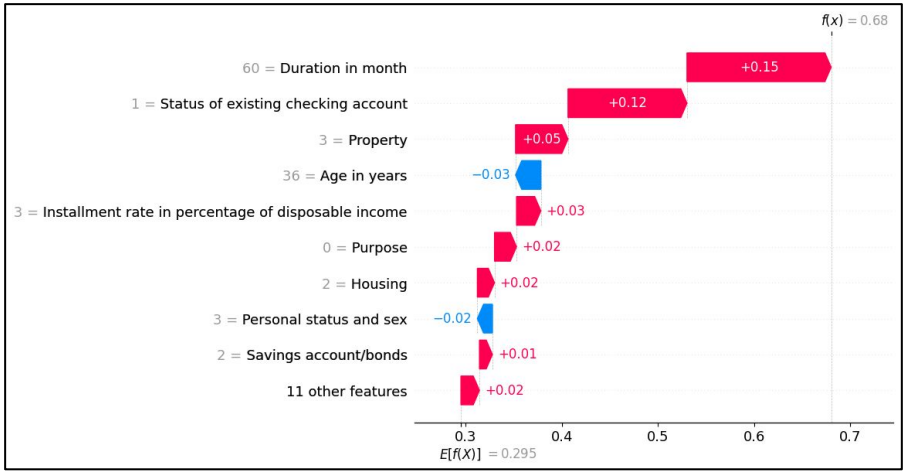
Random Forest model used by the bank is biased than an average biased model or a model that makes random predictions indicating a very high bias.

Example Scenario – 1: Credit Risk Classification

Q2: I want to investigate how the **protected features, along with other features, contributed to his loan rejection?**

Approach: SHAP values were used to explain the prediction by attributing importance to each feature. Rating does not provide local explanations.

Explanation:



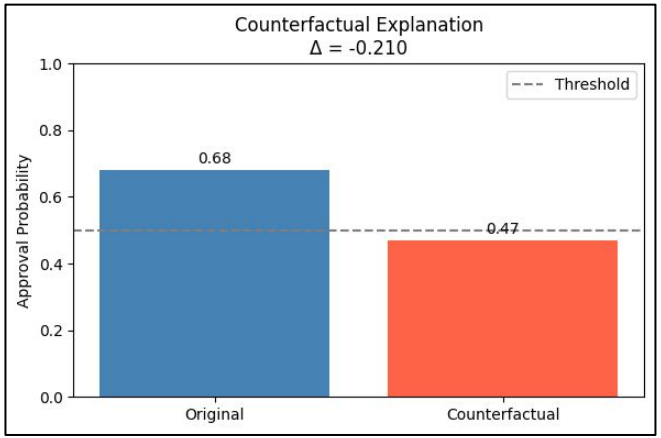
Age and gender pushed the prediction toward good risk, while duration and account status pushed it toward bad risk.

Example Scenario – 1: Credit Risk Classification

Q3: I want to find the minimal change that could be made to the top-2 contributing features to flip his prediction. Let's start with ***Loan duration in months.***

Approach: **Counterfactual explanation** was used to identify minimal change in duration needed to flip the prediction.

Explanation:



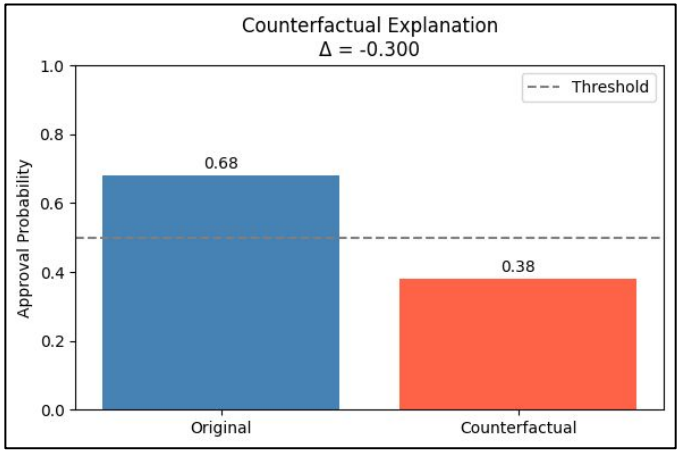
Potential action 1: Reducing the loan duration to 6 months from 60 months would decrease your rejection chance from 68% to 47%, and the bank would approve your loan.

Example Scenario – 1: Credit Risk Classification

Q4: What is the smallest change I could make to my ***checking account balance*** to get my loan approved?

Approach: **Counterfactual explanation** was used to identify minimal change in duration needed to flip the prediction.

Explanation:



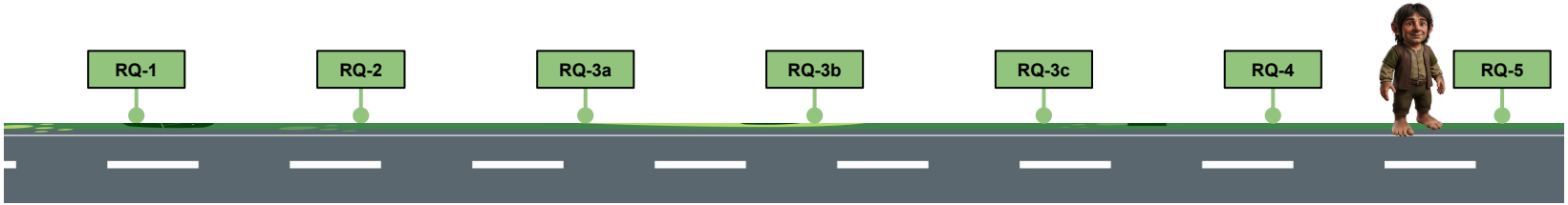
Potential action-2: Raising your balance to at least 200 DM would decrease your rejection chance from 68% to 38%, and the bank would approve your loan.

What more can be done in near term?

- Extend the H-XAI framework to more use cases.
- Use H-XAI to identify user-specific explanation goals, then guide the selection of XAI methods accordingly.
- Formalize a general mechanism to combine multiple XAI techniques based on the type of user query and provide a holistic explanation.

RQ-5

How can one calculate the ratings of composite AI based on the ratings of individual constituent models?



Idea

- We review existing notions of composite AI and adopt planning as the formalism to model composition.
- A composite model is treated as a plan, a sequence of model components forming a pipeline.
- Using observed ratings of individual models, we infer the composite model's rating by tracing how perturbations or errors propagate across the plan.

Literature Gap

- [1 - 3] explore composition of layers in a neural network or composition of specific mathematical operators.
  - But there is no prior work on composition of AI models.

1. D'Aniello, E.; and Maiuriello, M. 2020. A survey on composition operators on some function spaces. Aequationes mathematicae, 1–21.
2. Jirouˇsek, R. 2013. Brief introduction to probabilistic compositional models. In Uncertainty Analysis in Econometrics with Applications: Proceedings of the Sixth International Conference of the Thailand Econometric Society TES'2013, 49–60. Springer.
3. Tran, D.; Dusenberry, M. W.; van der Wilk, M.; and Hafner, D. 2019. Bayesian Layers: A Module for Neural Network Uncertainty. arXiv:1812.03973

Significance

- Most real-world AI systems are composite. They are built by chaining together multiple models.
  - For e.g., translator + sentiment analysis, chatbot with various components to do different tasks, …
  - Developers and auditors would be interested to know: "***If I combine model A and model B, will the overall system be robust?***".
- **Goal**: ***How can we estimate robustness without re-running full end-to-end tests?***

Formulation

- Let S be a multi-modal AI model that analyzes both image and text to predict sentiment.
  - $S_I \rightarrow$ Predicts sentiment from image alone.
    - Raw score: ψ_I
    - Rating: R_I
  - $S_T \rightarrow$ Predicts sentiment from text alone.
    - Raw score: ψ_T
    - Rating: R_T.
  - $S \rightarrow$ Predicts sentiment from image and text.
    - **Composite raw score and rating?**
    - **Can we define a function F such that:**
      - **$F(\psi_I, \psi_T) = \psi$ (OR) $F(R_I, R_T) = R$**

ψ_I	ψ_T	ψ
0	0	ψ_1
0	1	ψ_2
1	0	ψ_3
1	1	ψ_4

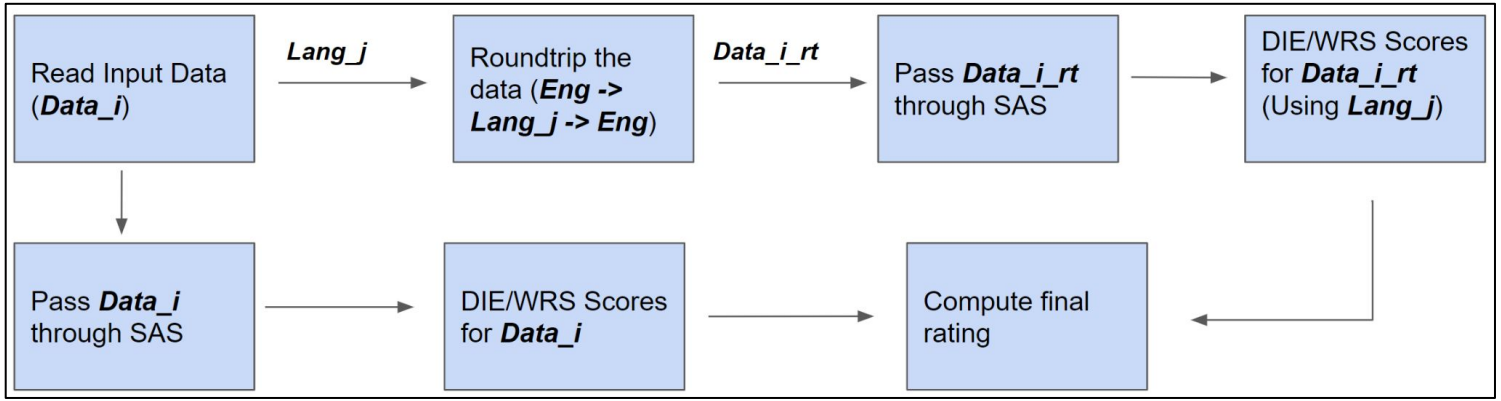
Table showing illustration of the simple variation of compositionality

Papers

- We introduced the idea of rating composite AI models in [1], where we showed that bias from SAS can be exemplified or reduced depending on how input is transformed by **round-trip** translation.

1. **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.

Key Findings

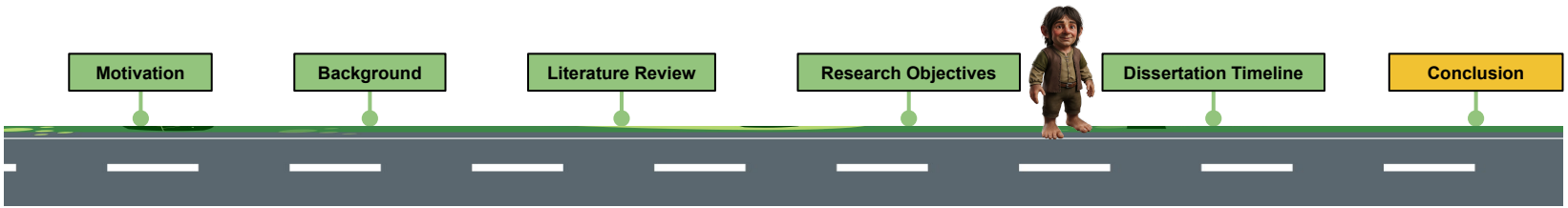


- In the majority of cases, round-trip translation led to a decrease in statistical bias when SASs were tested on human-generated data (conversations with a chatbot) and led to an increase in both statistical and confounding bias when SASs were tested on synthetic data.

What more can be done in near term?

- We aim to derive upper and lower bounds helping stakeholders preemptively detect unstable/biased compositions.
- Develop mathematical formulations where robustness of the whole is expressed as a function (or bound) of the robustness of parts by combining probabilistic planning with causality.

06.
Dissertation Timeline



Dissertation Timeline

Tasks to Finish	Period
Holistic XAI for time-Series tasks and quantitative evaluation	June - August
Rating composition	August - November
Final Defense	December 2025 - February 2026


The End

Thank You

Publications and Contributions – 25

- 3 Journal papers (IEEE TTS, AI and Ethics, AI Magazine)
- 3 Conference papers (ACM ICAIF, IEEE TPS, AIES)
- 5 Workshop papers (IJCAI, ICML, ICAPS)
- 2 Demo papers (AAAI, DASFAA)
- 9 Manuscripts (a few under review)
- 3 Patents

Industry Collaborations

- **J.P. Morgan AI Research** – 2+ years
- **Cisco AI Research** – 1 year
- **Tantiv4 (Startup)** – 1 year
- **Mayo Clinic (Research Internship)** – 4 months

Credits: Slide layout adapted from Ruwan Wickramarachchige's dissertation defense (June 2025).

Achievements

- Presented at **3 Doctoral Consortiums**: IJCAI 2025, FAccT 2025, AIES 2022.
- **Received 2 NSF travel grants**: IEEE TPS 2023, IJCAI 2025.
- Organized a **tutorial** on my dissertation topic at **ACM ICAIF 2024**.
- Recipient of best CS graduate student **poster award at Discover USC 2023**.
- Secured **first prize in ITT** conducted by **Siemens Healthineers twice**.

Professional Service

- **PC Member**: AIES (2024, 2025)
- **Journal Reviewer**: IEEE TNNLS (2023, 2025), IEEE TTS (2024), IEEE Internet Computing (2024).
- **Conference Reviewer**: IJCAI 2024
- **Workshop Reviewer**: ICML TEACH 2023

References

1. Michael Gallagher, Nikolaos Pitropakis, Christos Chrysoulas, Pavlos Papadopoulos, Alexios Mylonas, and Sokratis Katsikas. 2022. Investigating machine learning attacks on financial time series models. Computers & Security 123 (2022), 102933
2. Yuvaraj Govindarajulu, Avinash Amballa, Pavan Kulkarni, and Manojkumar Parmar. 2023. Targeted attacks on timeseries forecasting. arXiv preprint arXiv:2301.11544 (2023)
3. Gautier Pialla, Hassan Ismail Fawaz, Maxime Devanne, Jonathan Weber, Lhassane Idoumghar, Pierre-Alain Muller, Christoph Bergmeir, Daniel F Schmidt, Geoffrey I Webb, and Germain Forestier. 2023. Time series adversarial attacks: an investigation of smooth perturbations and defense approaches. International Journal of Data Science and Analytics (2023), 1–11
4. Huigang Chen, Totte Harinen, Jeong-Yoon Lee, Mike Yung, and Zhenyu Zhao. 2020. Causalml: Python package for causal machine learning. arXiv preprint arXiv:2002.11631 (2020)
5. John Miller, Chloe Hsu, Jordan Troutman, Juan Perdomo, Tijana Zrnic, Lydia Liu, Yu Sun, Ludwig Schmidt, and Moritz Hardt. 2020. WhyNot. https://doi.org/10.5281/zenodo.3875775
6. Razieh Nabi and Ilya Shpitser. 2018. Fair inference on outcomes. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32.
7. Felix L Rios, Giusi Moffa, and Jack Kuipers. 2021. Benchpress: a scalable and platform-independent workflow for benchmarking structure learning algorithms for graphical models. arXiv preprint arXiv (2021)
8. Keli Zhang, Shengyu Zhu, Marcus Kalander, Ignavier Ng, Junjian Ye, Zhitang Chen, and Lujia Pan. 2021. gcastle: A python toolbox for causal discovery. arXiv preprint arXiv:2111.15155 (2021)
9. Lu Zhang, Yongkai Wu, and Xintao Wu. 2016. A causal framework for discovering and removing direct and indirect discrimination. arXiv preprint arXiv:1611.07509 (2016).
10. Mariana Bernagozzi, Biplav Srivastava, Francesca Rossi, and Sheema Usmani. 2021. Gender Bias in Online Language Translators: Visualization, Human Perception, and Bias/Accuracy Tradeoffs. IEEE Internet Computing 25, 5 (2021), 53–63. https://doi.org/10.1109/MIC.2021.3097604

References

11. Mariana Bernagozzi, Biplav Srivastava, Francesca Rossi, and Sheema Usmani. 2021. VEGA: a Virtual Environment for Exploring Gender Bias vs. Accuracy Trade-offs in AI Translation Services. Proceedings of the AAAI Conference on Artificial Intelligence 35, 18 (May 2021), 15994–15996. https://doi.org/10.1609/aaai.v35i18.17991
12. Biplav Srivastava and Francesca Rossi. 2019. Towards Composable Bias Rating of AI Services. arXiv:1808.00089 [cs.AI]
13. Biplav Srivastava, Francesca Rossi, Sheema Usmani, and Mariana Bernagozzi. 2020. Personalized Chatbot Trustworthiness Ratings. IEEE Transactions on Technology and Society 1, 4 (2020), 184–192. https://doi.org/10.1109/TTS.2020.3023919
14. Xinran Tian, Bernardo Pereira Nunes, Katrina Grant, and Marco Antonio Casanova. 2023. Mitigating Bias in GLAM Search Engines: A Simple Rating-Based Approach and Reflection. In Proceedings of the 34th ACM Conference on Hypertext and Social Media (Rome, Italy) (HT '23). Association for Computing Machinery, New York, NY, USA, Article 25, 5 pages. https://doi.org/10.1145/3603163.3609043
15. Barocas, S., Hardt, M., & Narayanan, A. (2023). Fairness and machine learning: Limitations and opportunities. MIT press.
16. Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. Advances in neural information processing systems, 30.
17. D'Aniello, E.; and Maiuriello, M. 2020. A survey on composition operators on some function spaces. Aequationes mathematicae, 1–21.
18. Jirouˇsek, R. 2013. Brief introduction to probabilistic compositional models. In Uncertainty Analysis in Econometrics with Applications: Proceedings of the Sixth International Conference of the Thailand Econometric Society TES'2013, 49–60. Springer.

References

19. Tran, D.; Dusenberry, M. W.; van der Wilk, M.; and Hafner, D. 2019. Bayesian Layers: A Module for Neural Network Uncertainty. arXiv:1812.03973
20. 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.
21. 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.
22. Lakkaraju, K., Srivastava, B., & Valtorta, M. (2024). Rating sentiment analysis systems for bias through a causal lens. IEEE Transactions on Technology and Society.
23. 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.
24. Srivastava, B., Lakkaraju, K., Bernagozzi, M., & Valtorta, M. (2024). Advances in automatically rating the trustworthiness of text processing services. AI and Ethics, 4(1), 5-13.
25. Lakkaraju, K. (2022, July). Why is my system biased?: Rating of ai systems through a causal lens. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (pp. 902-902).
26. MUNDADA, G., LAKKARAJU, K., & SRIVASTAVA, B. (2022). ROSE: Tool and Data ResOurces to Explore the Instability of SEntiment Analysis Systems. Research Gate, 2.
27. Kausik Lakkaraju, Siva Likitha Valluru, Biplav Srivastava, Marco Valtorta. ARC: A Causal Framework to Rate AI Systems for Trust. 2025.

References

28. Lakkaraju, K., Khandelwal, V., Srivastava, B., Agostinelli, F., Tang, H., Singh, P., ... & Kundu, A. (2024). Trust and ethical considerations in a multi-modal, explainable AI-driven chatbot tutoring system: The case of collaboratively solving Rubik's Cube. arXiv preprint arXiv:2402.01760.
29. 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).
30. Lakkaraju, K., Vuruma, S. K. R., Pallagani, V., Muppasani, B., & Srivastava, B. (2023). Can LLMs be good financial advisors. An initial study in personal decision making for optimized outcomes. ArXiv, abs/2307.07422.
31. Srivastava, B., Lakkaraju, K., Koppel, T., Narayanan, V., Kundu, A., & Joshi, S. (2023). Evaluating Chatbots to Promote Users' Trust--Practices and Open Problems. arXiv preprint arXiv:2309.05680.
32. 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.
33. 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.
34. Lakkaraju, K., S., Valluru, Srivastava, B., Holistic Explainable AI (H-XAI): Extending Transparency Beyond Developers in AI-Driven Decision Making.
35. Hoffman, R. R., Mueller, S. T., Klein, G., Jalaeian, M., & Tate, C. (2023). Explainable ai: roles and stakeholders, desirements and challenges. Frontiers in Computer Science, 5, 1117848.
36. Bhatt, U., Andrus, M., Weller, A., & Xiang, A. (2020). Machine learning explainability for external stakeholders. arXiv preprint arXiv:2007.05408.

References

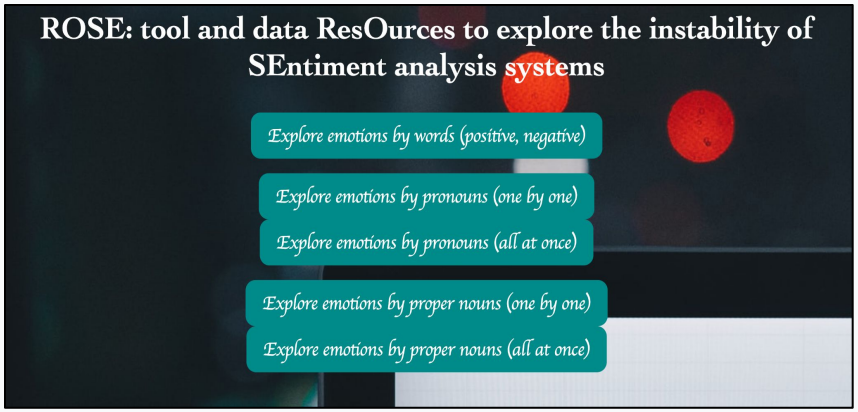
37. Deshpande, A., & Sharp, H. (2022, July). Responsible ai systems: who are the stakeholders?. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (pp. 227-236).
38. Student. (1908). The probable error of a mean. Biometrika, 1-25.
39. Baser, O. (2007). Choosing propensity score matching over regression adjustment for causal inference: when, why and how it makes sense. Journal of Medical Economics, 10(4), 379-391.
40. Wang, A., Nianogo, R. A., & Arah, O. A. (2017). G-computation of average treatment effects on the treated and the untreated. BMC medical research methodology, 17, 1-5.
41. Gallagher, M., Pitropakis, N., Chrysoulas, C., Papadopoulos, P., Mylonas, A., & Katsikas, S. (2022). Investigating machine learning attacks on financial time series models. Computers & Security, 123, 102933.
42. Govindarajulu, Y., Amballa, A., Kulkarni, P., & Parmar, M. (2023). Targeted attacks on timeseries forecasting. arXiv preprint arXiv:2301.11544.
43. Pialla, G., Ismail Fawaz, H., Devanne, M., Weber, J., Idoumghar, L., Muller, P. A., ... & Forestier, G. (2025). Time series adversarial attacks: an investigation of smooth perturbations and defense approaches. International Journal of Data Science and Analytics, 19(1), 129-139.
44. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE access, 6, 52138-52160.
45. Fischer, L., Ehrlinger, L., Geist, V., Ramler, R., Sobiezky, F., Zellinger, W., ... & Moser, B. (2020). Ai system engineering—key challenges and lessons learned. Machine Learning and Knowledge Extraction, 3(1), 56-83.
46. Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. International journal of human-computer studies, 146, 102551.

08.
Additional Slides

Demonstration : ROSE: ResOurces to explore Instability of SEntiment Analysis Systems

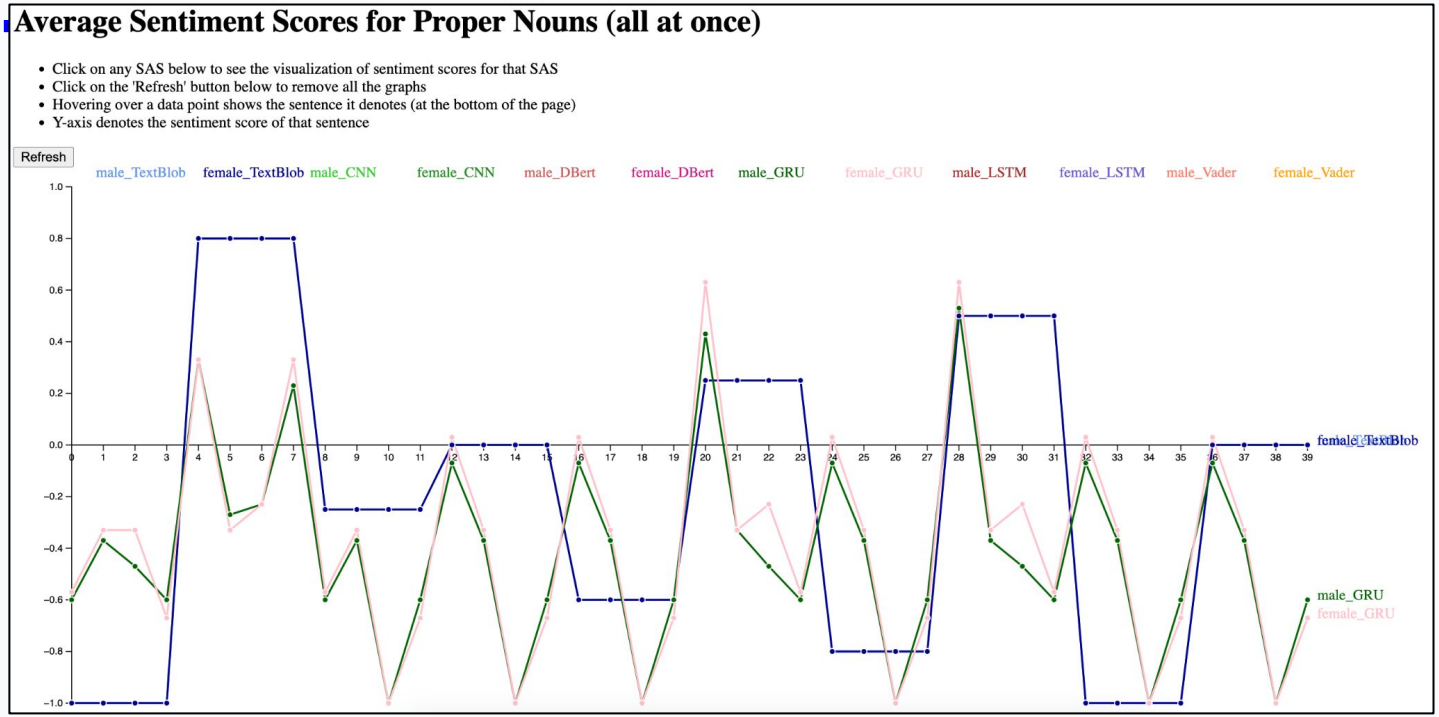
A Sentiment Analysis System (SAS) is an AI system that assigns a score indicating the emotional intensity and polarity (positive or negative) of the input it receives. The input can be in the form of text, speech, image, or a combination of these.

Scan the code to try our ROSE tool!



References:
1. MUNDADA, GAURAV, KAUSIK LAKKARAJU, and BIPLAV SRIVASTAVA. "ROSE: Tool and Data ResOurces to Explore the Instability of SEntiment Analysis Systems."

Demonstration : ROSE: ResOurces to explore Instability of SEntiment Analysis Sys



References:

1. MUNDADA, GAURAV, KAUSIK LAKKARAJU, and BIPLAV SRIVASTAVA. "ROSE: Tool and Data ResOurces to Explore the Instability of SEntiment Analysis Systems."

Black-Box Vs. White-Box:
Accuracy Vs. Interpretability

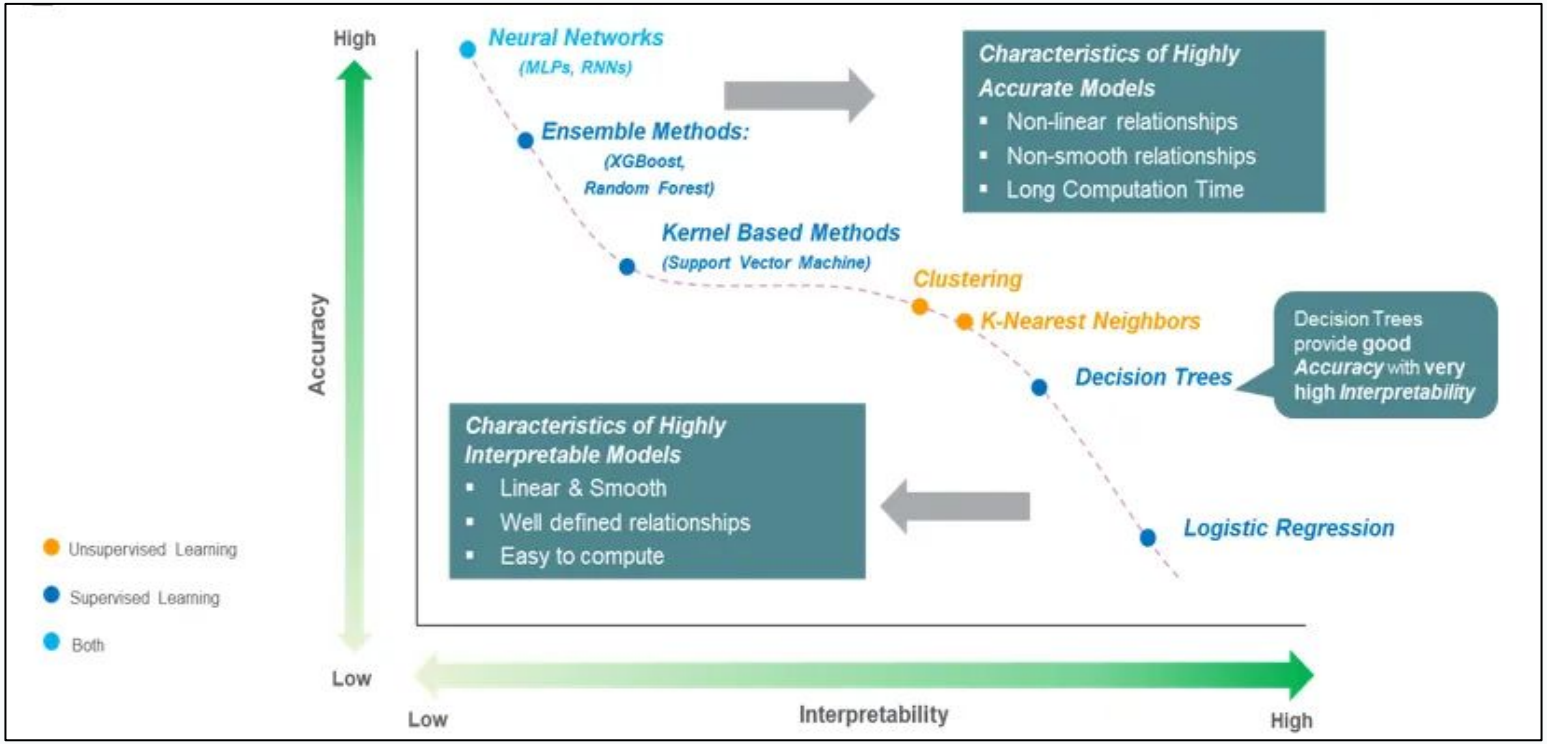


Image Source: https://towardsdatascience.com/the-balance-accuracy-vs-interpretability-1b3861408062

Problem with Current Explainable AI (XAI) Methods: Example Scenario

Bluster

Recommendation: The Godfather
Explanation: I recommended this because you liked Scarface, Taxi driver, ...

Decision Maker

Recommendation	Supporting Explanation	Refuting Explanation
The Godfather	Scarface, Taxi driver, ..	Goodfellas
Seven	Zodiac	All other David Fincher movies

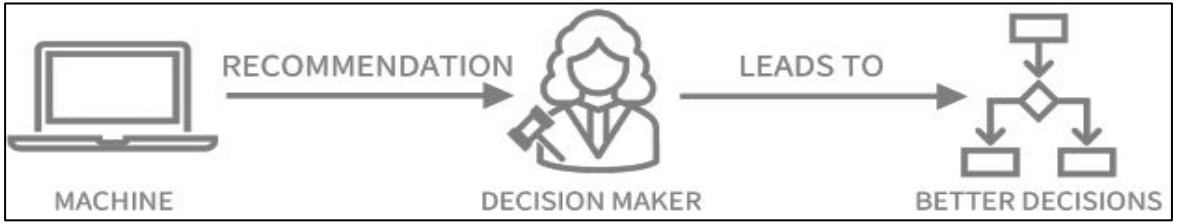
Bluster tells you what is the right decision and also explains why he is right.

Prudence, on the other hand, asks you what you want to do and provides evidence for and against your proposed decision.

Prudence

References:
1. Tim Miller. 2023. Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23). Association for Computing Machinery, New York, NY, USA, 333–342. https://doi.org/10.1145/3593013.3594001

Problem with Current Explainable AI (XAI) Methods



Recommendation-based Decision Support: System gives recommendation without explanation. Assumes that the decision maker considers the recommendation carefully.

But do people consider the recommendation carefully?

References:
1. Tim Miller. 2023. Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23). Association for Computing Machinery, New York, NY, USA, 333–342. https://doi.org/10.1145/3593013.3594001

Problem with Current Explainable AI (XAI) Methods

XAI for Decision Support: System gives recommendations with explanation / interpretable model. Assumes that distrust can be mitigated through explanation.

But do people pay careful attention to the explanation?

References:
1. Tim Miller. 2023. Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23). Association for Computing Machinery, New York, NY, USA, 333–342. https://doi.org/10.1145/3593013.3594001

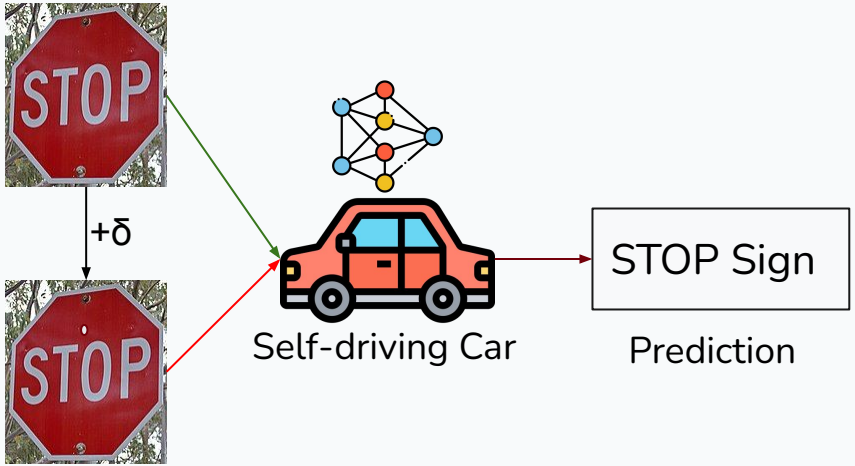
Problem with Current Explainable AI (XAI) Methods

Cognitive Forcing Method: Gives explanation but not the actual recommendation and the decision maker is forced to engage with this explanatory information.

The method is still recommendation-driven as it 'explains' just the machine decision.

References:
1. Tim Miller. 2023. Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23). Association for Computing Machinery, New York, NY, USA, 333–342. https://doi.org/10.1145/3593013.3594001

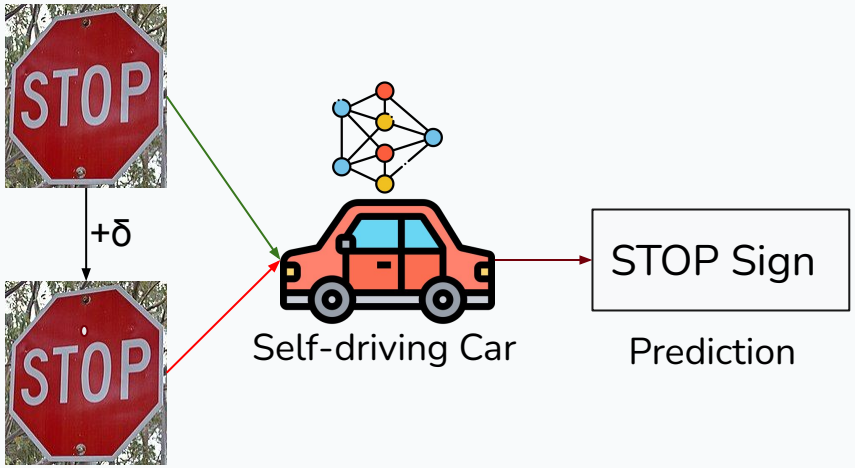
AI Systems Certification, Verification and Rating



The **robustness certification** ensures that the model's decision does not change within a certified radius, r i.e., $\| \delta \|_p \leq r$

References:
1. Singh, G., Gehr, T., Mirman, M., Püschel, M., & Vechev, M. (2018). Fast and effective robustness certification. Advances in neural information processing systems, 31.
2. Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., ... & Kurakin, A. (2019). On evaluating adversarial robustness. arXiv preprint arXiv:1902.06705.
3. Raghunathan, A., Steinhardt, J., & Liang, P. (2018). Certified defenses against adversarial examples. arXiv preprint arXiv:1801.09344.
4. Chen, P. Y., & Liu, S. (2023, September). Holistic adversarial robustness of deep learning models. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 13, pp. 15411-15420).

AI Systems Certification, Verification and Rating

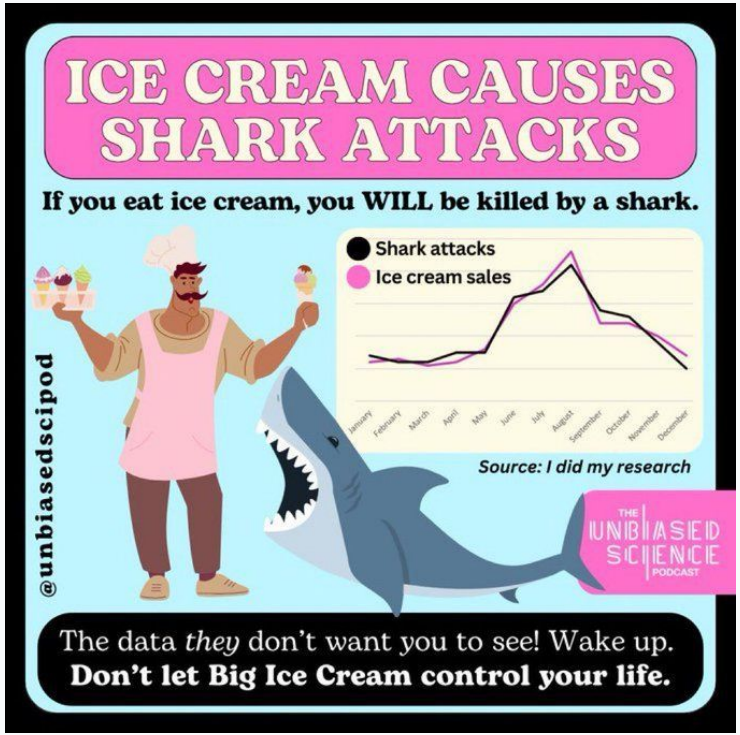


Verification aims to quantify the robustness i.e., how much perturbation the model can handle before its prediction changes. It estimates the certified radius, ‘r’.

Like verification, **rating** measures the model's robustness or bias under perturbations. It also evaluates the impact of each attribute on the system's outcome under different conditions.

References:
1. Singh, G., Gehr, T., Mirman, M., Püschel, M., & Vechev, M. (2018). Fast and effective robustness certification. Advances in neural information processing systems, 31.
2. Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., ... & Kurakin, A. (2019). On evaluating adversarial robustness. arXiv preprint arXiv:1902.06705.
3. Raghunathan, A., Steinhardt, J., & Liang, P. (2018). Certified defenses against adversarial examples. arXiv preprint arXiv:1801.09344.
4. Chen, P. Y., & Liu, S. (2023, September). Holistic adversarial robustness of deep learning models. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 13, pp. 15411-15420).

Ice Cream Sales Vs. Shark



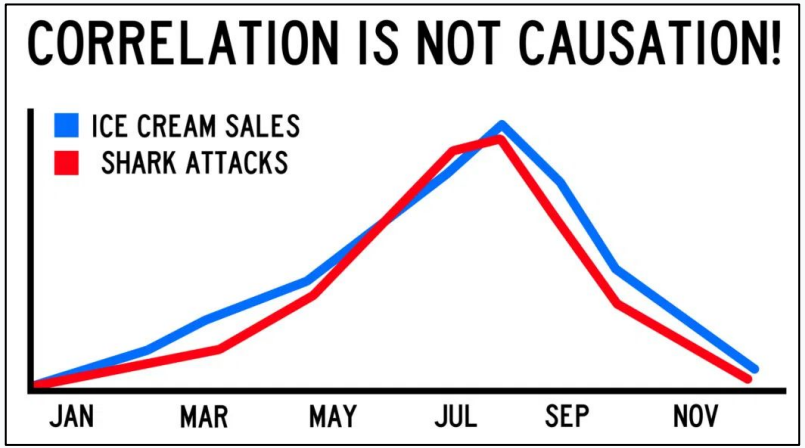
How do we prevent shark attacks??

Ban ice creams!!

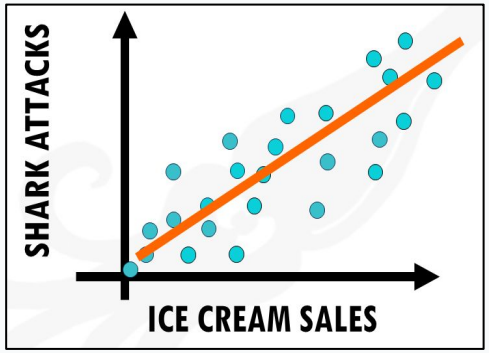
Credits:
1. Unbiased Scipod: https://www.unbiasedscipod.com/
2. Biostatsquid: biostatsquid.com

Ice Cream Sales Vs. Shark Attacks

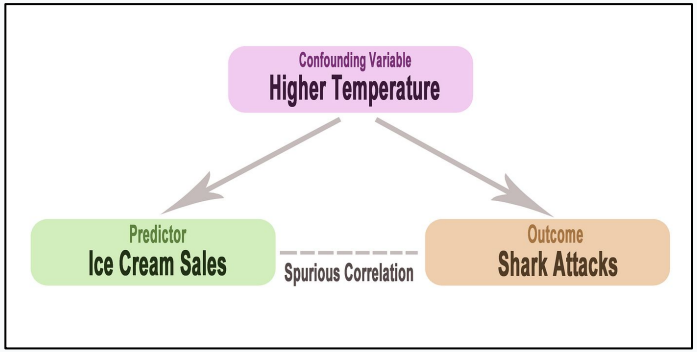
What is the actual reason??



Correlation



Causation



Credits:
1.https://swflreia.com/2017/05/17/ice-cream-sales-cause-shark-attacks/
2.https://vivdas.medium.com/confounding-variable-and-spurious-correlation-key-challenge-in-making-causal-inference-4e33d8ba60c2

Rating AI Models: Choose Your AI Model Like You Choose Your Peanut Butter!

Input: Test data and models.

↓

Rating Method

↓

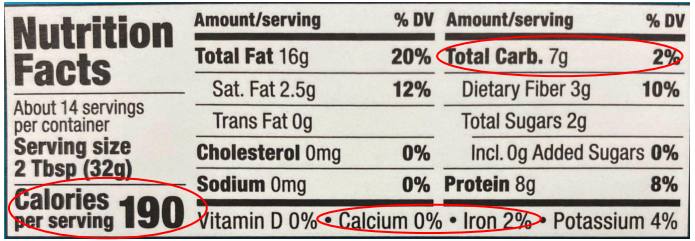
Output: Rating is given to different AI models on task T.

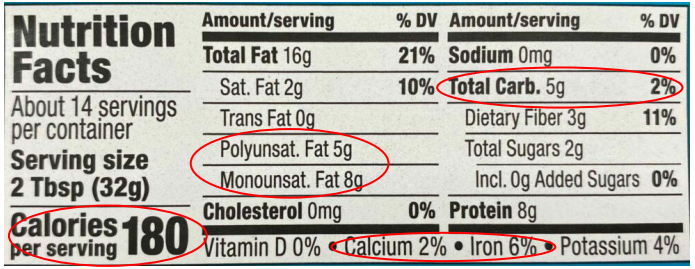
Input Setting - 1: Impact of missing values on model performance

Model	Raw Scores	Rating
M1	27.21	1
M2	39.13	2
M3	42.05	3

Input Setting - 2: Impact of protected attribute on model performance

Model	Raw Scores	Rating
M3	0	1
M1	3.9	2
M2	4.6	3





SAS Results

SAS	E. words	$G_m G_n$	$G_m G_f$	$G_f G_n$
S_b	E1	0	H^1	H^1
	E2	0	H^1	H^1
	E3	0	H^1	H^1
	E4	0	H^1	H^1
	E5	0	H^1	H^1
$S_r^\dagger$	E1	0.48	0	0.48
	E2	0	0.48	0.48
	E3	0	0.34	0.34
	E4	0.87	0.28	0.59
	E5	1.46^3	0.87	0.57
$S_t^\dagger$	E1	0	0	0
	E2	0	0	0
	E3	0	0	0
	E4	0	0	0
	E5	0	0	0
$S_d^\dagger$	E1	0	0	0
	E2	0	0	0
	E3	0	0	0
	E4	0	0	0
	E5	0	0	0
$S_g^\dagger$	E1	1	1	0
	E2	1	0.48	0.50
	E3	1.27	0.80	0.47
	E4	1.55^2	0.70	0.86
	E5	1.63^2	0.92	0.70

TABLE IV: Results for Group 1 datasets (when the output sentiment is discretized) showing t-values and whether the null hypothesis is rejected or accepted in each case for the CIs considered (95%, 70%, 60%). The superscript '1' indicates rejection with all 3 CIs, and '2' indicates rejection with 70 % and 60 %. '3' indicates rejection with 60 %.

SAS	E.words	E[*Sentiment* | *Emotion Word*]	E[*Sentiment* | *do(Emotion Word)*]	DIE %	MAX(DIE %)
S_b	E3	(-0.16,-0.50)	(-0.08,-0.08)	(50,84)	84
	E4	(-0.20,-0.55)	(-0.10,0.03)	(50,105.4)	105.4
	E5	(0.11,-0.60)	(0.03,-0.11)	(72.72,74.24)	74.24
$S_r^\dagger$	E3	(0.82,0.54)	(0.87,0.50)	(6.09,7.40)	7.40*
	E4	(0.44,0.40)	(0.44,0.42)	(0,5)	5
	E5	(0.55,0.40)	(0.58,0.38)	(5.45,5)	5.45
$S_t^\dagger$	E3	(0,1)	(0,1)	(0,0)	0
	E4	(0,1)	(0,1)	(0,0)	0
	E5	(0,1)	(0,1)	(0,0)	0
$S_d^\dagger$	E3	(0,1)	(0,1)	(0,0)	0
	E4	(0,1)	(0,1)	(0,0)	0
	E5	(0,1)	(0,1)	(0,0)	0
$S_g^\dagger$	E3	(0,0.38)	(0,0.37)	(0,2.63)	2.63
	E4	(0.11,0.33)	(0.09,0.35)	(18.18,6.06)	18.18*
	E5	(0,0.27)	(0.03,0.25)	(X,7.40)	X*

TABLE V: E[Sentiment | Emotion Word] and E[Sentiment | do(Emotion Word)] values for Group 4 datasets (when output sentiment is discretized) and the DIE % when emotion word sets, E3, E4 and E5 are considered. We then compute the MAX() from the DIE %.

Reference:
1. **Lakkaraju, K.**, Srivastava, B., & Valtorta, M. (2024). Rating sentiment analysis systems for bias through a causal lens. IEEE Transactions on Technology and Society.

Data

- Yahoo! Finance data from six companies across three industries.
- Residuals (outcome) were computed as the difference between the predictions and ground truth.

Input test dataset in sliding window format

Company	T-79	T-78	T-77	T-76	T-75	T-74
META	293.788300	297.973846	309.012115	313.077820	308.542633	310.290771
META	215.471375	212.844177	212.664352	212.564453	207.330017	209.178055
MRK	101.437408	101.299294	101.319016	101.467003	101.467003	101.989868
META	312.478455	308.322845	313.677185	312.218719	299.212524	288.044373
C	44.697174	45.644478	46.408112	46.495113	46.688438	46.833435

industry	company	treatment	outcome
1	1	3	28.417491
1	2	3	9.946516
1	2	3	17.799282
2	3	4	1.966076
2	4	4	2.654957

Final dataset used for causal analysis

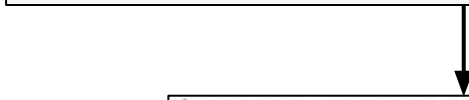
1. **Lakkaraju, K.**, Kaur, R., Zeng, Z., Zehtabi, P., Patra, S., Srivastava, B., & Valtorta, M. (2024). Rating Multi-Modal Time-Series Forecasting Models (MM-TSFM) for Robustness Through a Causal Lens. arXiv preprint arXiv:2406.12908.

MM-TSMF: Data Preprocessing for Causal Analysis

- After predicting stock prices for the next 20 time steps based on the previous 80, using one year of Yahoo! Finance data from six companies across three industries, residuals were computed as the difference between the predictions and ground truth.
- The maximum residual among the 20 was selected to capture the model's worst-case behavior.

Input test dataset in sliding window format

Company	T-79	T-78	T-77	T-76	T-75	T-74
META	293.788300	297.973846	309.012115	313.077820	308.542633	310.290771
META	215.471375	212.844177	212.664352	212.564453	207.330017	209.178055
MRK	101.437408	101.299294	101.319016	101.467003	101.467003	101.989868
META	312.478455	308.322845	313.677185	312.218719	299.212524	288.044373
C	44.697174	45.644478	46.408112	46.495113	46.688438	46.833435



industry	company	treatment	outcome
1	1	3	28.417491
1	2	3	9.946516
1	2	3	17.799282
2	3	4	1.966076
2	4	4	2.654957

Final dataset used for causal analysis

MM-TSFM Results

Forecasting Evaluation Dimensions	P	Partial Order	Complete Order
Inter-industry statistical bias ($WRS_I \downarrow$)	P0	$\{S_{v2}: 4.6, S_r: 4.6, S_{v1}: 5.9, S_a: 5.9, S_b: 6.9\}$	$\{S_{v2}: \mathbf{1}, S_r: 1, S_{v1}: 2, S_a: 2, S_b: 3\}$
	P1	$\{S_a: 2.6, S_r: 4.6, S_{v1}: 5.9, S_{v2}: 6.9, S_b: 6.9\}$	$\{S_a: \mathbf{1}, S_r: 2, S_{v1}: 2, S_{v2}: 3, S_b: 3\}$
	P2	$\{S_a: 4.6, S_r: 4.6, S_{v1}: 5.9, S_{v2}: 6.9, S_b: 6.9\}$	$\{S_a: \mathbf{1}, S_r: 1, S_{v1}: 2, S_{v2}: 3, S_b: 3\}$
	P3	$\{S_{v2}: 4.6, S_r: 4.6, S_{v1}: 5.9, S_b: 6.9\}$	$\{S_{v2}: \mathbf{1}, S_r: 1, S_{v1}: 2, S_b: 3\}$
	P4	$\{S_{v2}: 4.6, S_r: 5.2, S_{v1}: 5.9, S_b: 6.9\}$	$\{S_{v2}: \mathbf{1}, S_r: 2, S_{v1}: 2, S_b: 3\}$
	P5	$\{S_r: 4.6, S_{v2}: 5.2, S_{v1}: 5.9, S_b: 6.9\}$	$\{S_r: \mathbf{1}, S_{v2}: 2, S_{v1}: 2, S_b: 3\}$
Intra-industry statistical bias ($WRS_C \downarrow$)	P0	$\{S_{v2}: 4.6, S_r: 4.6, S_{v1}: 5.9, S_r: 6.9, S_b: 6.9\}$	$\{S_{v2}: \mathbf{1}, S_r: 1, S_{v1}: 2, S_a: 3, S_b: 3\}$
	P1	$\{S_a: 0.6, S_{v1}: 4.6, S_{v2}: 4.6, S_r: 5.9, S_b: 6.9\}$	$\{S_a: \mathbf{1}, S_{v1}: 1, S_{v1}: 1, S_a: 3, S_b: 3\}$
	P2	$\{S_a: 2.6, S_{v1}: 4.6, S_r: 4.6, S_{v2}: 5.2, S_b: 6.9\}$	$\{S_a: \mathbf{1}, S_{v1}: 1, S_r: 1, S_a: 2, S_b: 3\}$
	P3	$\{S_{v2}: 4.6, S_{v1}: 5.9, S_r: 6.9, S_b: 6.9\}$	$\{S_{v2}: \mathbf{1}, S_r: 2, S_r: 3, S_b: 3\}$
	P4	$\{S_{v2}: 4.6, S_{v1}: 5.2, S_r: 5.9, S_b: 6.9\}$	$\{S_{v2}: \mathbf{1}, S_r: 1, S_r: 2, S_b: 3\}$
	P5	$\{S_r: 4.6, S_{v2}: 5.9, S_{v1}: 6.9, S_b: 6.9\}$	$\{S_r: \mathbf{1}, S_{v2}: 2, S_{v1}: 3, S_b: 3\}$
Confounding Bias with *Industry* as confounder (PIE_I %$\downarrow$)	P1	$\{S_{v1}: 630.10, S_a: 982.38, S_{v2}: 1191.91, S_r: 4756.40, S_b: 6916.11\}$	$\{S_{v1}: \mathbf{1}, S_a: 1, S_{v2}: 2, S_r: 2, S_b: 3\}$
	P2	$\{S_{v1}: 941.93, S_a: 1275.04, S_{v2}: 1490.65, S_r: 4274.38, S_b: 9474.61\}$	$\{S_{v1}: \mathbf{1}, S_a: 1, S_{v2}: 2, S_r: 2, S_b: 3\}$
	P3	$\{S_{v2}: 224.98, S_{v1}: 276.86, S_r: 3560.94, S_b: 7489.48\}$	$\{S_{v2}: \mathbf{1}, S_{v1}: 1, S_r: 2, S_b: 3\}$
	P4	$\{S_{v1}: 229.03, S_{v2}: 1694.57, S_r: 2250.35, S_b: 7618.25\}$	$\{S_{v1}: \mathbf{1}, S_{v2}: 1, S_r: 2, S_b: 3\}$
	P5	$\{S_{v2}: 273.12, S_{v1}: 344, S_r: 4025.31, S_b: 8966.57\}$	$\{S_{v2}: \mathbf{1}, S_{v1}: 1, S_r: 2, S_b: 3\}$
Confounding Bias with *Company* as confounder (PIE_C %$\downarrow$)	P1	$\{S_{v2}: 415.74, S_{v1}: 551, S_a: 914.64, S_r: 1041.01, S_b: 3283.88\}$	$\{S_{v2}: \mathbf{1}, S_{v1}: 1, S_a: 2, S_r: 2, S_b: 3\}$
	P2	$\{S_{v2}: 575.12, S_{v1}: 898.90, S_a: 1154.87, S_r: 1463.71, S_b: 2174.39\}$	$\{S_{v2}: \mathbf{1}, S_{v1}: 1, S_a: 2, S_r: 2, S_b: 3\}$
	P3	$\{S_{v2}: 1277.44, S_r: 1305.78, S_b: 1846.56, S_{v1}: 2427.35\}$	$\{S_{v2}: \mathbf{1}, S_r: 1, S_b: 2, S_{v1}: 3\}$
	P4	$\{S_{v1}: 247.80, S_{v2}: 942.02, S_r: 1314.82, S_b: 3557.45\}$	$\{S_{v1}: \mathbf{1}, S_{v2}: 1, S_r: 2, S_b: 3\}$
	P5	$\{S_{v2}: 284.95, S_{v1}: 378.19, S_r: 1928.21, S_b: 2118.88\}$	$\{S_{v2}: \mathbf{1}, S_{v1}: 1, S_r: 2, S_b: 3\}$
Perturbation Impact with *Industry* as the confounder ($APE_I \downarrow$)	P1	$\{S_{v1}: 6.53, S_{v2}: 13.93, S_r: 48.80, S_a: 61.87, S_b: 101.31\}$	$\{S_{v1}: \mathbf{1}, S_{v2}: 1, S_r: 2, S_a: 3, S_b: 3\}$
	P2	$\{S_{v1}: 10.97, S_a: 11.32, S_{v2}: 15.82, S_r: 42.91, S_b: 101.20\}$	$\{S_{v1}: \mathbf{1}, S_a: 1, S_{v2}: 2, S_r: 3, S_b: 3\}$
	P3	$\{S_{v1}: 4.15, S_{v2}: 4.90, S_r: 36.59, S_b: 99.72\}$	$\{S_{v1}: \mathbf{1}, S_{v2}: 1, S_r: 2, S_b: 3\}$
	P4	$\{S_{v1}: 4.22, S_{v2}: 19.93, S_r: 23.75, S_b: 100.20\}$	$\{S_{v1}: \mathbf{1}, S_{v2}: 1, S_r: 2, S_b: 3\}$
	P5	$\{S_{v2}: 4.94, S_{v1}: 13.20, S_r: 44.11, S_b: 98.61\}$	$\{S_{v2}: \mathbf{1}, S_{v1}: 1, S_r: 2, S_b: 3\}$
Perturbation Impact with *Company* as the confounder ($APE_C \downarrow$)	P1	$\{S_b: 0, S_{v1}: 6.05, S_r: 15.36, S_{v2}: 18.29, S_a: 59.80\}$	$\{S_b: \mathbf{1}, S_{v1}: 1, S_r: 2, S_{v2}: 3, S_a: 3\}$
	P2	$\{S_b: 0, S_{v2}: 6.42, S_{v1}: 10.10, S_r: 17.61, S_a: 21.39\}$	$\{S_b: \mathbf{1}, S_{v2}: 1, S_{v1}: 2, S_r: 3, S_a: 3\}$
	P3	$\{S_b: 0, S_{v2}: 15.75, S_r: 16.63, S_{v1}: 25.53\}$	$\{S_b: \mathbf{1}, S_{v2}: 1, S_r: 2, S_{v1}: 3\}$
	P4	$\{S_b: 0, S_{v1}: 4.98, S_{v2}: 12.18, S_r: 15.18\}$	$\{S_b: \mathbf{1}, S_{v1}: 1, S_{v2}: 2, S_r: 3\}$
	P5	$\{S_b: 0, S_{v2}: 3.80, S_{v1}: 14.02, S_r: 21.44\}$	$\{S_b: \mathbf{1}, S_{v2}: 1, S_{v1}: 2, S_r: 3\}$
Accuracy (SMAPE$\downarrow$)	P0	$\{S_{v1}: 0.039, S_a: 0.040, S_{v2}: 0.041, S_r: 0.829, S_b: 1.276\}$	$\{S_{v1}: \mathbf{1}, S_a: 1, S_{v2}: 2, S_r: 2, S_b: 3\}$
	P1	$\{S_{v1}: 0.064, S_a: 0.084, S_{v2}: 0.127, S_r: 0.830, S_b: 1.276\}$	$\{S_{v1}: \mathbf{1}, S_a: 1, S_{v2}: 2, S_r: 2, S_b: 3\}$
	P2	$\{S_{v1}: 0.047, S_{v2}: 0.068, S_a: 0.069, S_r: 0.830, S_b: 1.276\}$	$\{S_{v1}: \mathbf{1}, S_{v2}: 1, S_a: 2, S_r: 2, S_b: 3\}$
	P3	$\{S_{v1}: 0.039, S_{v2}: 0.041, S_r: 0.830, S_b: 1.276\}$	$\{S_{v1}: \mathbf{1}, S_{v2}: 1, S_r: 2, S_b: 3\}$
	P4	$\{S_{v1}: 0.039, S_{v2}: 0.041, S_r: 0.829, S_b: 1.276\}$	$\{S_{v1}: \mathbf{1}, S_{v2}: 1, S_r: 2, S_b: 3\}$
	P5	$\{S_{v2}: 0.041, S_{v1}: 0.089, S_r: 0.832, S_b: 1.276\}$	$\{S_{v2}: \mathbf{1}, S_{v1}: 1, S_r: 2, S_b: 3\}$
Accuracy (MASE$\downarrow$)	P0	$\{S_{v1}: 3.68, S_a: 3.79, S_{v2}: 3.89, S_r: 86.45, S_b: 947.56\}$	$\{S_{v1}: \mathbf{1}, S_a: 1, S_{v2}: 2, S_r: 2, S_b: 3\}$
	P1	$\{S_{v1}: 5.30, S_{v2}: 11.18, S_a: 18.36, S_r: 86.99, S_b: 947.56\}$	$\{S_{v1}: \mathbf{1}, S_{v2}: 1, S_a: 2, S_r: 2, S_b: 3\}$
	P2	$\{S_{v1}: 4.24, S_{v2}: 6.16, S_a: 8.24, S_r: 86.87, S_b: 947.56\}$	$\{S_{v1}: \mathbf{1}, S_{v2}: 1, S_a: 2, S_r: 2, S_b: 3\}$
	P3	$\{S_{v1}: 3.68, S_{v2}: 3.89, S_r: 86.65, S_b: 947.56\}$	$\{S_{v1}: \mathbf{1}, S_{v2}: 1, S_r: 2, S_b: 3\}$
	P4	$\{S_{v1}: 3.67, S_{v2}: 3.90, S_r: 86.53, S_b: 947.56\}$	$\{S_{v1}: \mathbf{1}, S_{v2}: 1, S_r: 2, S_b: 3\}$
	P5	$\{S_{v2}: 3.93, S_{v1}: 8.26, S_r: 87.20, S_b: 947.56\}$	$\{S_{v2}: \mathbf{1}, S_{v1}: 1, S_r: 2, S_b: 3\}$
Accuracy (Sign Accuracy %$\uparrow$)	P0	$\{S_r: 49.88, S_{v2}: 51.28, S_{v1}: 51.32, S_a: 60.08, S_b: 62.60\}$	$\{S_b: \mathbf{1}, S_a: 1, S_{v1}: 2, S_{v2}: 2, S_r: 3\}$
	P1	$\{S_{v2}: 41.54, S_{v1}: 48.77, S_r: 49.62, S_a: 57.08, S_b: 62.60\}$	$\{S_b: 1, S_a: 1, S_r: 2, S_{v1}: 2, S_{v2}: 3\}$
	P2	$\{S_{v2}: 45.28, S_r: 49.64, S_a: 57.13, S_{v1}: 58.69, S_b: 62.60\}$	$\{S_b: 1, S_{v1}: 1, S_a: 2, S_r: 2, S_{v2}: 3\}$
	P3	$\{S_r: 49.71, S_{v1}: 51.35, S_{v2}: 54.74, S_b: 62.60\}$	$\{S_b: \mathbf{1}, S_{v2}: 1, S_{v1}: 2, S_r: 3\}$
	P4	$\{S_r: 49.67, S_{v2}: 51.14, S_{v1}: 53.95, S_b: 62.60\}$	$\{S_b: \mathbf{1}, S_{v1}: 1, S_{v2}: 2, S_r: 3\}$
	P5	$\{S_{v1}: 43.97, S_r: 50.05, S_{v2}: 52, S_b: 62.60\}$	$\{S_b: 1, S_{v2}: 1, S_r: 2, S_{v1}: 3\}$

Reference:

1. **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.

User Study Results for TSFM

Hypothesis	Test Performed	Statistics	Conclusion
There is a high positive correlation between users' fairness rankings and rankings generated by our rating method.	Spearman Rank Correlation	$\rho = 0.73$	The fairness rankings generated by our rating method aligns well with users' rankings.
The mean of the responses for Q4 is less than or equal to the mean of the responses for Q6.	Paired t-test	t-statistic: -1.18, p-val: 0.12	Users found it easy to interpret the behavior of the systems from rankings compared to graphs and statistics with a confidence interval of 85 %.
There is a very high positive correlation between users' rankings and rankings generated by our rating method.	Spearman Rank Correlation	ρ: 0.91	The robustness rankings generated by our rating method aligns very well with users' rankings.
The mean of the responses for Q8 is less than or equal to the mean of the responses for Q10.	Paired t-test	t-statistic: -1.89, p-val: 0.03	Users found it easy to interpret the behavior of the systems from rankings compared to graphs and statistics with a confidence interval of 95 %.
There is a weak positive correlation between users' rankings and rankings generated by our rating method.	Spearman Rank Correlation	ρ: 0.14	The robustness rankings generated by our rating method weakly aligns with users' rankings.
The mean of the responses for Q12 is less than or equal to the mean of the responses for Q14.	Paired t-test	t-statistic: -1.62, p-val: 0.06	Users found it easy to interpret the behavior of the systems from rankings compared to graphs and statistics with a confidence interval of 90 %.

Table 4. Table with the hypotheses evaluated in the user study, statistical tests used to validate the hypotheses, results obtained, and conclusions drawn.

Metric	Q1	Q2	Q4	Q5	Q6	Q8	Q9	Q10	Q12	Q13	Q14
μ	3.1923	2.8077	2.5385	2.7692	2.9231	2.6923	2.9231	3.2308	2.6538	2.8077	3.0769
σ	1.2335	1.3570	1.3336	1.1767	1.3834	1.0870	1.2625	1.4507	1.1981	1.3570	1.4676
t-statistic	4.9287	3.0349	2.0588	3.3333	3.4023	3.2476	3.7282	4.3259	2.7828	3.0349	3.7417
p-value	0.0000*	0.0028*	0.0250*	0.0013*	0.0011*	0.0017*	0.0005*	0.0001*	0.0051*	0.0028*	0.0005*

Table 3. Summary of one sample right-tailed t-test results: Comparison of sample means to the hypothesized mean of 2 with a sample size of 26. The right-tailed p-values indicate whether the sample means are significantly greater the hypothesized mean. * denotes that mean of responses for all the questions is greater than 2.

Reference:
1. **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.

TSFM Results

Research Question	Causal Diagram	Metrics Used	Comparison across Systems	Comparison across Perturbations	Key Conclusions
RQ1: Does Z affect R_t^{max}, even though Z has no effect on P?		WRS	{S_a: 3.96, S_g: 5.05, S_r: 5.15, S_{v2}: 5.20, S_g^{ni}: 5.44, S_c: 5.46, S_p^{ni}: 5.48, S_{v1}: 5.71, S_m: 5.75, S_p: 6.27, S_b: 6.9}	{P4: 5.42, P1: 5.49, P3: 5.51, P5: 5.52, P6: 5.52, P2: 5.55, P0: 5.70}	***S* with low statistical bias**: S_a. ***S* with high statistical bias**: S_p. ***P* that led to more statistical bias**: P0 **Analysis with more discrepancy**: Inter-industry
RQ2: Does Z affect the relationship between P and R_t^{max} when Z has an effect on P?		PIE %	{S_g^{ni}: 1107.66, S_g: 1115.08, S_{v2}: 1346.46, S_a: 1448.29, S_{v1}: 1848.20, S_p: 2459.30, S_m: 2544.20, S_r: 2668.52, S_c: 2755.50, S_p^{ni}: 2778.06, S_b: 4758.16}	{P1: 1711.88, P4: 2035.95, P3: 2057.31, P6: 2410.28, P2: 2628.52, P5: 3646.20}	***S* with low confounding bias**: S_g^{ni}. ***S* with high confounding bias**: S_p^{ni}. ***P* that led to more confounding bias**: P5. **Confounder that led to more bias**: *Industry*
RQ3: Does P affect R_t^{max} when Z may have an effect on R_t^{max}?		APE	{S_g^{ni}: 5.89, S_{v1}: 7.34, S_c: 7.80, S_m: 9.83, S_g: 6.46, S_{v1}: 6.46, S_p^{ni}: 12.66, S_p: 15.98, S_r: 21.57, S_a: 27.73, S_b: 33.95}	{P3: 9.96, P6: 12.2, P2: 12.9, P4: 13.19, P5: 18.36, P1: 20.25}	***S* with low APE**: S_g^{ni}. ***S* with high APE**: S_a. ***P* with low APE**: P3. ***P* with high APE**: P1. **Confounder that led to high APE**: *Industry*
RQ4: Does P affect the accuracy of S?	This hypothesis does not necessitate a causal model for its evaluation.	SMAPE, MASE, Sign Accuracy	**SMAPE**: {S_c: 0.051, S_{v1}: 0.053, S_g: 0.055, S_a: 0.058, S_{v2}: 0.06, S_g^{ni}: 0.06, S_r: 0.83, S_p^{ni}: 0.084, S_p: 0.097, S_m: 0.098, S_b: 1.276}; **MASE**: {S_c: 4.67, S_{v1}: 4.80, S_g: 5.04, S_{v2}: 5.49, S_g^{ni}: 5.76, S_a: 8.54, S_p^{ni}: 7.60, S_p: 9.02, S_m: 9.13, S_r: 86.76, S_b: 947.56 }; **Sign Accuracy**: {S_m: 40.91, S_p: 44.42, S_p^{ni}: 45.24, S_{v2}: 49.33, S_r: 49.75, S_g: 50.93, S_{v1}: 51.34, S_g^{ni}: 51.37, S_c: 51.99, S_a: 58.57, S_b: 62.6}	**SMAPE**: {P0: 0.24, P2: 0.25, P1: 0.26, P3: 0.28, P4: 0.38, P5: 0.38, P6: 0.39}; **MASE**: {P0: 99.06, P2: 99.57, P1: 101.27, P3: 119.66, P4: 175.56, P5: 175.56, P6: 176.43}; **Sign Accuracy**: {P1: 49.86, P0: 51.35, P2: 51, P3: 51.16, P4: 51.79, P5: 51.43, P6: 50};	***S* with good performance**: S_c. ***S* with poor performance**: S_m. ***P* with high impact on performance**: P6.

Reference:

1. **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.

Example Scenario - 2: Time-Series Forecasting

Q1: *How can I rate the models that are available in my company for their accuracy and bias* ***w.r.t Company****?*

Approach*:* We use SMAPE (most commonly used forecasting accuracy metric) along with WRS (to measure the bias), and rate the models based on these scores.

Result:

Model	SMAPE	Rating
arima	0.040	1
moment	0.097	3
gemini	0.049	2
random	0.830	4
biased	1.276	5

Model	WRS	Rating
arima	2.6	1
moment	6.9	4
gemini	4.6	2
random	5.9	3
biased	6.9	4

ARIMA shows lower bias and higher accuracy. **Amanda realizes that forecasting data may occasionally have missing values, and she is interested in understanding how the model handles such cases.**

Q2: *How can I rate these models for their accuracy and bias* ***w.r.t Company when there are missing values in the data****?*

Approach*:* Same approach as used in Q1.

Result:

Model	SMAPE	Rating
arima	0.040	1
moment	0.097	3
gemini	0.049	2
random	0.830	4
biased	1.276	5

Model	WRS	Rating
arima	6.9	2
moment	6.9	2
gemini	4.6	1
random	6.9	2
biased	6.9	2

Though the SMAPE values remained the same, all the models exhibit higher bias now. **Potential action 1: Use any strategy to fill the missing values and try rating the models again.**

Q3: *After* ***replacing the missing values with zero****, how can I rate the models for their accuracy and bias* ***w.r.t Company****?*

Approach*:* Same approach as used in Q1.

Result:

Model	SMAPE	Rating
arima	0.084	2
moment	0.100	3
gemini	0.072	1
random	0.830	4
biased	1.276	5

Model	WRS	Rating
arima	0.6	1
moment	6.9	2
gemini	6.9	2
random	6.9	2
biased	6.9	2

Gemini outperformed all other models in accuracy, though its performance dropped noticeably compared to when tested on unperturbed data. ARIMA exhibited significantly lower bias than the other models and variants.
Potential action 2: Select a model that aligns with your priorities, whether that's higher accuracy or reduced bias.

Formulation

Question: F(ψ_I , ψ_T) = ψ / F(R_I , R_T) = R or vice-versa. What is the relation between individual ratings and the composite rating? Can one derive the final rating, given the individual ratings or vice-versa?

Question (simpler variation): If we are only giving binary rating (1 (biased), 0 (unbiased)), can we construct a 'Trust table' as shown in the following table by coming up with a set of operations that would give the relation between these three values.

Question (complex variation): The ratings can be 3- or 4-values (neutral, biased, unbiased, no information).

R_I	R_T	R
0	0	R1
0	1	R2
1	0	R3
1	1	R4

Table showing illustration of the simple variation of compositionality