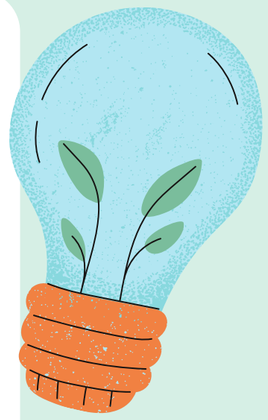
# Exploring the Efficacy of Large Language Models for Post Hoc Explanations in Bail Classification

Zhe Niu, Class of 2024,
Duke Kunshan University,
Instructed by Prof. Luyang Zhang



## Acknowledgments

I extend sincere thanks to "ECON 211: Intelligent Economics: An Explainable AI Approach" and Professor Luyao Zhang for their invaluable guidance and insights. Their support significantly shaped my research in explainable AI. Additionally, I am grateful to my classmates for their constructive peer reviews and suggestions, which greatly enhanced my learning and research development.

## Background and Motivation

The judicial system is increasingly adopting Machine Learning (ML) models for bail decisions, raising concerns about transparency and interpretability. This shift necessitates Explainable AI (XAI), where Large Language Models (LLMs) like GPT-3.5 present novel opportunities for elucidating the decision-making process in bail classification.

## Research Questions

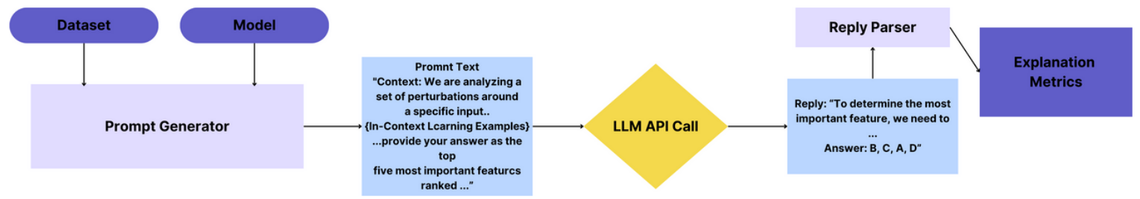
1. Can LLMs offer accurate and reliable post hoc explanations for ML model predictions in bail classification?
2. How do LLMs perform in comparison to established post hoc explanation methods like LIME and SHAP in identifying key features influencing bail decisions?

## Contribution to Literature

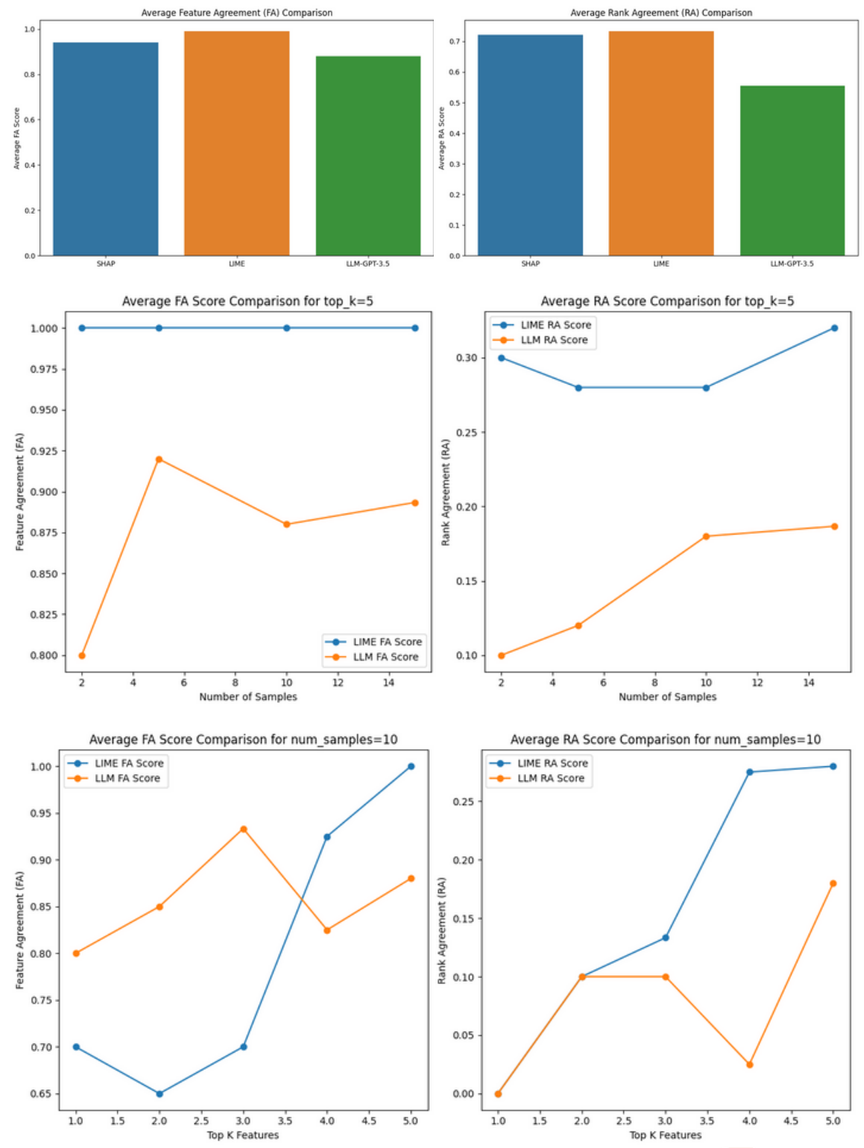
This research expands the XAI field by integrating advanced LLMs as explanatory tools in bail classification, building upon existing studies that predominantly focus on traditional methods like LIME and SHAP. It introduces LLMs into comparative studies of explanation methods in ML, offering new insights into their utility and effectiveness.

## Methodology

The research involves preprocessing the COMPAS dataset, ensuring its suitability for ML model training and LLM-based explanation generation. Logistic Regression, Support Vector Machine (SVM), and Random Forest are utilized to create a robust baseline. The primary focus is on Perturbation-Based In-Context Learning (ICL) using LLMs to determine the most influential features in the models' predictions. The explanations are evaluated for fidelity and interpretability, using Feature Agreement (FA) and Rank Agreement (RA). These metrics assess the alignment of LLM explanations with traditional methods and actual influential features identified by the ML models.

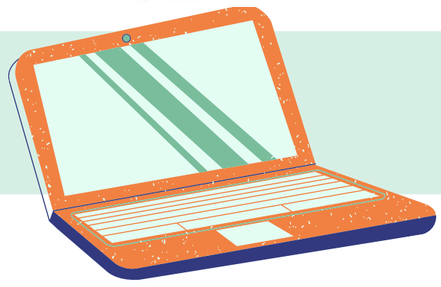



## Results




## Future Research

This study advances XAI methodologies in the context of judicial decision-making, aiming to enhance trust and transparency in ML systems. Future research may explore LLM applications in various domains and refine explanation generation techniques.



## References

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ...& Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI. Information fusion, 58, 82-115.
Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM computing surveys (CSUR), 51(5), 1-42.
ProPublica. (2016). COMPAS recidivism risk score data and analysis [Dataset]. ProPublica. https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis
Kroeger, N., Ley, D., Krishna, S., Agarwal, C., & Lakkaraju, H. (2023). Are Large Language Models Post Hoc Explainers?. arXiv preprint arXiv:2310.05797.
Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys, 55(9), 1-35.