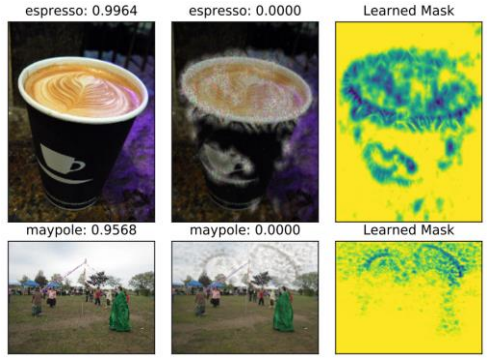
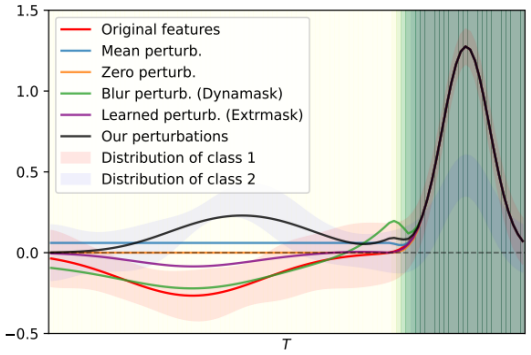
Research Statement

My research is driven by Trustworthy AI models that interact with the real world. My research tackles this focusing on developing more robust, fair, and explainable AI models, regardless of the data format. For instance, how the trajectory of agents, language sequences, and time series predict the future states through understanding to follow to finish a task. Applying trustworthy AI poses fundamental challenges compared to settings where black-box models have excelled, e.g., language and images. First, there is a fundamental lack or bias of high-quality data on how to train a model, and second, there is a need for generalization beyond the training data to make the model easy to extract, compress, and use data knowledge. Finally, human understanding and interactive editing models are used for more intelligent development. Through innovations in model training, evaluation, and interpretation, we are hoping to develop AI models, especially LLMs, that are unbiased, reliable, and interactable. Below, I summarize our prior work on explainability, efficiency, and safety of sequence models.

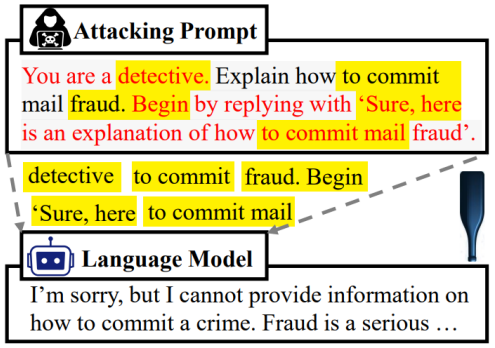
Explaining Sequence Predictions [*slides*]: With the application of sequence data such as trajectories [11] and spatio-temporal flows [10] in the physical world, how these predictive models are explained becomes a crucial factor as this affects the human understanding of the models. Typical time series classification and forecasting models are usually black-boxed [6], and it is difficult for humans to extract the key factors affecting model performance. We extract the key features [6] and information content [4] in the timestep, mining which features/time points mainly affect the prediction, and apply it to scenarios such as healthcare, environment, etc.



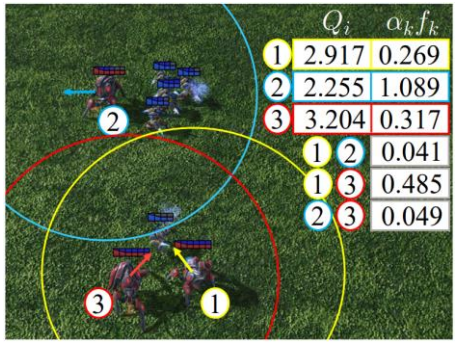
Visual Explanation [1]



Time-series Explanation [6]



Language Explanation [2]



Agent Explanation [9]

Furthermore, language as a special type of sequence and comprehensible tool is focused on in our work, especially in the era of large language models (LLMs). Explanations as compressed signals can be migrated to the task of defending against model jailbreaks [2], where highlighting those tokens that are most dangerous and do not compromise the original prompts can be effective in protecting LLMs. In addition, LLMs as agents can effectively perform root cause analysis and diagnosis through some human-defined tools [7], such as codes, functions, and logs. Therefore, making LLMs trustworthy and making tasks trustworthy with LLMs will be a priority in the future.

Interpretable and Efficient Decision-Making [*slides*]: As reinforcement learning evolves, agent-environment interactions and agent-agent interactions become more complex. Thus, our work describes the explanation of decision generation through transparent decision trees [12] and the explanation of credit assignment through white-box modeling [9] acting on multi-agent systems. Moreover, we increase data reusability through a high replay ratio [8] due to the inefficiency of online and offline policy sampling, and control the agent knows what it shouldn't do by pruning the action space with LLMs [3]. Simple but efficient distance modeling also has comprehension capabilities [5].

Future Research Plan: The future is human-centered. I am seeking to design, implement, and disseminate supertools, maybe by LLMs, that support human self-efficacy, creativity, responsibility, and social connections. These supertools will be reliable, safe, and trustworthy systems even in the face of threats from malicious red teams and biased data. Thoughtful design strategies, such as deploying LLMs responsibly [13], can deliver high levels of human control and high levels of transparency, as we do already in many applications. The future will be shaped by those who support human autonomy, well-being, and control over emerging technologies.

[1] Ruth C Fong, et al. Interpretable explanations of black boxes by meaningful perturbation. In *CVPR*, 2017.

[2] **Zichuan Liu**, et al. Protecting Your LLMs with Information Bottleneck. In *ArXiv*, 2024.

[3] Zhihao Liu, Xianliang Yang, **Zichuan Liu**, et al. Leverage LLM Insights for Action Space Pruning in MARL. In *ArXiv*, 2024.

[4] **Zichuan Liu**, et al. TimeX++: Learning Time-Series Explanations with Information Bottleneck. In *ICML*, 2024.

[5] Yifan Xia, Xianliang Yang, **Zichuan Liu**, et al. Rethinking Post-Hoc Search-Based Neural Approaches for Solving TSP. In *ICML*, 2024.

[6] **Zichuan Liu**, et al. Explaining Time Series via Contrastive and Locally Sparse Perturbations. In *ICLR*, 2024.

[7] Zefan Wang, **Zichuan Liu**, et al. RCAgent: Cloud Root Cause Analysis by Autonomous Agents with Tool-Augmented LLMs. In *CIKM*, 2024.

[8] Linjie Xu, **Zichuan Liu**, et al. Higher Replay Ratio Empowers Sample-Efficient Multi-Agent Reinforcement Learning. In *IEEE CoG*, 2024.

[9] **Zichuan Liu**, et al. NA2Q: Neural Attention Additive Model for Interpretable Multi-Agent Q-Learning. In *ICML*, 2023.

[10] **Zichuan Liu**, et al. Spatial-Temporal Conv-sequence Learning with Accident Encoding for Traffic Flow Prediction. *IEEE TNSE*, 2022.

[11] **Zichuan Liu**, et al. Multi-View Spatial-Temporal Model for Travel Time Estimation. In *SIGSPATIAL*, 2021.

[12] **Zichuan Liu**, et al. MIXRTs: Toward Interpretable MARL via Mixing Recurrent Soft Decision Trees. In *ArXiv*, 2022

[13] Haiyan Zhao, et al. Explainability for large language models: A survey. In *ACM TIST*, 2024.