FEIRAN JIA

feiran.jia@outlook.com ◇ (814)826-8755 ◇ feiran.io ◇ Publications ◇ LinkedIn

EDUCATION

Pennsylvania State University *Sep 2020 - Dec 2025*
Ph.D. in Informatics

Washington University in St. Louis *Aug 2018 - May 2020*
M.S. in Computer Science

ShanghaiTech University *Sep 2014 - Jun 2018*
B.S. in Computer Science; Minor in Finance

INTERNSHIP

Meta, Model Building Foundation and Optimization (Ads CoreML) *May 2024 - Aug 2024*

- Designed and implemented uncertainty quantification for Ads Ranking Models (MTML models) using techniques such as MC Dropout, MC Dropout Connect, Sparse Variational Gaussian Process, Mean Variance networks, Deep Ensemble, and Batch Ensemble, achieving a $\sim 0.26\%$ NE gain.
- Proposed uncertainty-aware learning strategies addressing both aleatoric and epistemic uncertainty to improve the feedback loop problem, potentially enhancing long-term model performance.

SELECTED PROJECTS

Topic Large Language Model Agents
Keywords: Openai API, Huggingface; LLM inference, Prompt Engineer, COT, Reflection, Tool-using, behavioral economics, injection attack

- **Task Alignment for LLM Agents** *Aug 2024-Present*
  - Proposed a task alignment concept that formalizes the relationship between instructions in LLM agent systems, establishing a foundation for ensuring agent behaviors align with user-defined objectives.
  - Developed Task Shield, a practical test-time defense mechanism that dynamically evaluates instructions during inference, demonstrating model-agnostic protection against indirect prompt injection attacks.
  - Achieved state-of-the-art results on AgentDojo benchmark: reduced attack success rate from 47.69% to 2.07% on GPT-4o while improving utility from 50.08% to 69.79%.
- **Trust Behaviors of LLM Agents** *Oct 2023-Jan 2024*
  - Investigated human-like trust behaviors and alignment in language model agents through behavioral economics trust games, analyzing the impact of demographics and reasoning strategies (like CoT, Reflection) on trust.
- **Math Problem Solving with GPT-4 and Tool-using** *Feb 2023-Apr 2023*
  - Developed MathChat, a framework that combines CoT prompting, multi-stage problem decomposition, tool usage, and self-consistency, enabling agents to collaboratively tackle complex math challenges.
  - Enhanced GPT-4's math problem-solving capabilities with MathChat, achieving up to 15% higher accuracy in Algebra and a 6%.

Topic: Robust, Explainable, and Automated Machine Learning
Keywords: Python, Pytorch, Scikit-learn; Deep Learning, Machine Learning, Data Shift, Hyperparameter Tuning

- **Risk in Test-time Adaptation (TTA)** *Nov 2022 - Feb 2023*
  - Uncovered a vulnerability in TTA, showing the impact of malicious test batch samples on benign predictions.
  - Proposed the Distribution Invading Attack to alter model predictions through strategic malicious data injection, demonstrating effectiveness with over a 90% success rate on datasets like CIFAR-100-C and ImageNet-C.
- **Targeted Hyperparameter Optimization (HPO) with Lexicographic Preferences** *Jan 2022 - Oct 2022*
  - Introduced LexiFlow, a randomized directed search method using lexicographic preferences for MO-HPO.
  - Tuned neural networks, XGBoost, and random forests using LexiFlow, demonstrating improvements in accuracy, fairness, feature selection, and overfitting mitigation.
- **Robust Counterfactual Explanations under the Distribution Shift** *Oct 2021 - Sep 2022*
  - Proposed RoCourseNet, a system designed to improve the prediction and explanation capabilities of machine learning models, particularly effective in adapting to data shifts.
  - Employed adversarial training for robust explanation generation, optimizing against worst-case data shifts and achieving over 91% robust validity, a 10% improvement over existing baselines.

Topic Computational Game Theory and Application
Keywords: Pytorch, Networkx, Pandas, Numpy, CPLEX; Gradient-based Optimization, Combinatorial Optimization

- **Gradient Better Response Methods for Solving Large Games** *Jun 2023-Present*
  - Developed the Stochastic Gradient Better Response Method for large-scale game equilibrium search, leveraging stochastic optimization and best response dynamics, enhanced by tensor computations and GPU-acceleration.
  - Managed Blotto games at scales of 4-player-5456-action and 60-player-286-action, a significant advancement beyond the state-of-the-art's 4-player-286-action capability.
- **Content Sharing Design for Social Welfare in Networked Disclosure Game** *Jan 2022-Feb 2023*
  - Modeled user decisions in online personal information sharing as a networked disclosure game, focusing on the balance between privacy and benefit, and developed a P-time graph algorithm to find the optimal equilibrium.
  - Treated content promotion as a network design problem, establishing its NP-hard nature, and created both a MILP exact solution and a scalable heuristic algorithm for real-world network sizes.

TEACHING ASSISTANT

Penn State University (2022-2023): Data Integration;Programming Models for Big Data; Data Science Capstone
Washington University in St. Louis (2019): Introduction to Machine Learning
ShanghaiTech University (2017-2018): Introduction to Algorithmic Game Theory; Computer Architecture

RESEARCH PAPERS

1. **Feiran Jia**, Tong Wu, Xin Qin, Anna Squicciarini. The Task Shield: Enforcing Task Alignment to Defend Against Indirect Prompt Injection in LLM Agents. Submitted to ARR, 2024.
2. **Feiran Jia**, Sixie Yu, Anna Squicciarini, and Yevgeniy Vorobeychik. Gradient Better Response Methods for Solving Large Games. Paper in Progress, 2024.
3. Chengxing Xie*, Canyu Chen*, **Feiran Jia**, Ziyú Ye, Kai Shu, Adel Bibi, Ziniu Hu, Philip Torr, Bernard Ghanem, Guohao Li. Can Large Language Model Agents Simulate Human Trust Behaviors? *NeurIPS*, 2024.
4. Yiran Wu, **Feiran Jia**, Shaokun Zhang, Hangyu Li, Erkang Zhu, Yue Wang, Yin Tat Lee, Richard Peng, Qingyun Wu, Chi Wang. MathChat: Converse to Tackle Challenging Math Problems with LLM Agents, *ICLR Workshop on LLM Agents*, 2024
5. Hangzhi Guo, **Feiran Jia**, Jinghui Chen, Anna Squicciarini, Amulya Yadav. RoCourseNet: Distributionally Robust Training of a Prediction Aware Recourse Model. In *CIKM*, 2023 (Oral).
6. Yiran Wu, **Feiran Jia**, Shaokun Zhang, Qingyun Wu, Hangyu Li, Erkang Zhu, Yue Wang, Yin Tat Lee, Richard Peng, Chi Wang. An Empirical Study on Challenging Math Problem Solving with GPT-4. In *arXiv*, 2023.
7. **Feiran Jia**, Chenxi Qiu, Sarah Rajtmajer, Anna Squicciarini. Content Sharing Design for Social Welfare in Networked Disclosure Game. In *UAI*, 2023.
8. Tong Wu, **Feiran Jia**, Xiangyu Qi, Jiachen T. Wang, Vikash Sehwag, Saeed Mahloujifar, Prateek Mittal. Uncovering Adversarial Risks of Test-Time Adaptation. In *ICML*, 2023.
9. Shaokun Zhang, **Feiran Jia**, Chi Wang, and Qingyun Wu. LexicoFlow: Multi-objective Hyperparameter Optimization with Lexicographic Preference. In *ICLR* 2023 (Oral).
10. Zun Li, **Feiran Jia**, Aditya Mate, Shahin Jabbari, Mithun Chakraborty, Milind Tambe, and Yevgeniy Vorobeychik. Solving Structured Hierarchical Games Using Differential Backward Induction. In *UAI*, 2022 (Oral).
11. Christopher Griffin, Anna Squicciarini, and **Feiran Jia**. Consensus in complex networks with noisy agents and peer pressure. In *Physica A: Statistical Mechanics and its Applications*, 2022.
12. **Feiran Jia**, Aditya Mate, Zun Li, Shahin Jabbari, Mithun Chakraborty, Milind Tambe, Michael Wellman, and Yevgeniy Vorobeychik. A Game-Theoretic Approach for Hierarchical Policy-Making. In *AASG*, 2021
13. **Feiran Jia**, Kai Zhou, Charles Kamhoua, and Yevgeniy Vorobeychik. Blocking Adversarial Influence in Social Networks. In *GameSec*, 2020.

HONORS & AWARDS

Meritorious Winner in Mathematical Contest in Modeling 2017 | UAI Scholarship 2023

PROFESSIONAL SERVICE

- **Conference Reviewer** UAI 2023, SecureComm 2023, AAMAS 2023, NeurIPS 2024, AAAI 2025, ICLR 2025, ICML 2025 (incoming)
- **Workshop Reviewer** NLP4Science workshop at EMNLP 2024, ICML 2024 Workshop FM-Wild
- **Journal Reviewer** DMLR
- **Volunteer** UAI 2023