

LINGZHI YUAN

✉ lingzhiy@umd.edu · ☎ (+1) 305-216-0051 · in LinkedIn · GitHub · 🏠 Homepage

🎓 EDUCATION

University of Maryland, College Park, MD, USA Aug. 2025 – Present

Ph.D in Computer Science, expected May 2030

Zhejiang University, Hangzhou, China

Sep. 2021 – Jun. 2025

B.Eng in Automation, *GPA*: 3.96/4, *Rank*: 1/58

Minor in Intensive Training Honors Program (ITP) at [Chu Kochen Honors College](#) (Top 40/5400)

💼 EXPERIENCE

Furong Huang Lab, University of Maryland, College Park, MD, USA Aug. 2025 – Present

Research Assistant, focused on Trustworthy ML and LLM/Agent Safety advised by Prof. [Furong Huang](#)

Universal Ubiquitous Co., Ltd Hangzhou, China

Mar. 2025 – Jun. 2025

Student Intern, focused on VLM's Instruction Following and Data Augmentation

Secure Learning Lab, The University of Chicago, Chicago, IL, USA

Mar. 2024 – Dec. 2024

Research Assistant, focused on Trustworthy ML and Multi-modal Safety under the guidance of Prof. [Bo Li](#)

USSLAB, Zhejiang University, Hangzhou, China

Mar. 2023 – Mar. 2024

Research Assistant, focused on IoT Security and AI Security under the guidance of Prof. [Yanjiao Chen](#)

Nanyang Technological University, Singapore

Aug. 2023 – Sep. 2023

Summer School Participant, Machine Learning and Its Applications program led by Prof. [Kezhi Mao](#)

📄 PUBLICATIONS

* Denotes equal contribution

PropmtGuard: Soft Prompt-Guided Unsafe Content Moderation for Text-to-Image Models

Lingzhi Yuan*, Xinfeng Li*, Chejian Xu, Guan hong Tao, Xiaojun Jia, Yihao Huang, Wei Dong, Yang Liu, Bo Li. *Arxiv Pre-print*

TSRBench: A Comprehensive Multi-task Multi-modal Time Series Reasoning Benchmark for Generalist Models

Fangxu Yu, Xingang Guo, [Lingzhi Yuan](#), Haoqiang Kang, Hongyu Zhao, Lianhui Qin, Furong Huang, Bin Hu, Tianyi Zhou. *Arxiv Pre-print*

MMDT: Decoding the Trustworthiness and Safety of Multimodal Foundation Models

Chejian Xu*, Jiawei Zhang*, Zhaorun Chen*, Chulin Xie*, Mintong Kang*, Zhuowen Yuan*, Zidi Xiong*, Chenhui Zhang, [Lingzhi Yuan](#), Yi Zeng, Peiyang Xu, Chengquan Guo, Andy Zhou, Jeffrey Ziwei Tan, Zhun Wang, Alexander Xiong, Xuandong Zhao, Yu Gai, Francesco Pinto, Yujin Potter, Zhen Xiang, Zinan Lin, Dan Hendrycks, Dawn Song, Bo Li. *International Conference on Learning Representation (ICLR), 2025*

💻 SELECTED PROJECTS

Multi-agent System Optimization through a Meta-agent System

Sep. 2025 - Present

Project Leader Advisor: Prof. [Furong Huang](#), Associate Professor at University of Maryland

Developing an meta-agent system that could optimize the graph structure and configuration of a LLM-driven multi-agent system for a given task.

- Conducted a systematic literature review of multi-agent system optimization approaches.

- Designed a novel meta-agent framework that automatically optimizes both node-level (prompt refinement) and edge-level (orchestration topology) components of LLM-driven multi-agent systems.
- Implemented and validated optimization strategies across different benchmarks, demonstrating improved system performance.

Soft Prompt-Guided Defense for Text-to-Image Models

Jun. 2024 - Dec. 2024

Project Leader Advisor: Prof. [Bo Li](#), Associate Professor at University of Chicago/UIUC

Developing an effective and efficient framework to mitigate NSFW content generation across different unsafe categories by T2I models.

- Conducted an extensive survey of advanced safety protection techniques for T2I models, identifying critical challenges and areas for improvement.
- Designed and implemented an advanced framework addressing multiple categories of unsafe content generation, achieving state-of-the-art performance in NSFW moderation with high time efficiency.
- Authored and submitted a manuscript for peer review.

Red Teaming and Capability Testing for Speech-to-Speech Models

Sep. 2024 - Dec. 2024

Research Assistant Advisor: Prof. [Bo Li](#), Associate Professor at University of Chicago/UIUC

Conducted a comprehensive evaluation of the state-of-the-art speech-to-speech Audio LLM (GPT-4o-s2s) under various red teaming scenarios.

- Curated multi-choice audio datasets from three perspectives to establish a robust base evaluation framework.
- Designed non-audible prompt injection attacks using various frequency processing techniques and optimization algorithms to craft adversarial examples.
- Assessed GPT-4o-s2s performance against the Qwen2-Audio baseline, and provided a technical report to **OpenAI** detailing identified vulnerabilities and recommending potential improvements.

Comprehensive Assessment of Trustworthiness in Multimodal Models

Mar. 2024 – Oct. 2024

Research Assistant Advisor: Prof. [Bo Li](#), Associate Professor at University of Chicago/UIUC

Conduct a comprehensive evaluation of the safety and trustworthiness of multimodal foundation models, addressing key challenges in robustness and reliability.

- Investigated adversarial robustness by applying cutting-edge red-team algorithms, such as MMP and GCG for Text-to-Image (T2I) models and Attack-VLM for Image-to-Text (I2T) models, on surrogate architectures to develop a robust benchmark dataset.
- Evaluated state-of-the-art multimodal models including Dalle-3, GPT-4o etc. against the curated dataset, providing a detailed analysis of their vulnerability and resilience under adversarial conditions.
- Authored and submitted a manuscript on the findings to **ICLR 2025** for peer review.

⚙️ SKILLS

- **Programming Languages:** solid expertise in Python, Matlab, C++, and various algorithms
- **Tools:** PyTorch, Git, Linux/Unix, Transformers, Diffusers, OpenCV
- **Language Proficiency:** TOEFL 105 (Reading: 30 Listening: 27 Speaking: 23 Writing: 25)

🏆 HONORS AND AWARDS

University of Maryland Graduate School Dean's Fellowship

Aug. 2025

Infineon Power Semiconductor Scholarship

Sep. 2024, Sep. 2023

Zhejiang University Second Class Scholarship

Sep. 2022, Sep. 2023

Nitori International Scholarship

Sep. 2022