Muyang Zheng

sandyyyzheng33@gmail.com | sandyyyzheng.github.io | Google Scholar | LinkedIn

Education

University of California, Davis — Davis, CA, USA
Master of Science in Computer Science — *Sept 2025 - Present*

Hefei University of Technology — Hefei, China
Bachelor of Engineering in Information Security — *Sept 2021 - June 2025*

Research Experience

Research Assistant, PLUM Lab at UC Davis – Advised by Dr. Lifu Huang — Sept 2025 – Present

- **Multimodal Agentic Framework Design:** Built GLIDE, an agentic gameplay video understanding framework that combines **structured reasoning**, **tool use**, and **temporal grounding** to analyze complex glitch events
- **Benchmarking & Performance:** Improved performance of open-source multimodal models after applying GLIDE, increasing average **F1 from 14.47% to 36.05%** and **mIoU from 0.28 to 0.51** across six backbones
- **Annotation Pipeline Design:** Developed a semi-automated annotation pipeline to build our dataset VIDEOGLITCHBENCH, featuring **5,238 gameplay videos from 120 games** with glitch descriptions and timestamps

Research Intern, Binjiang Institute of Zhejiang University – Advised by Dr. Meng Han — Aug 2024 – Apr 2025

- **Performance & Inference Optimization:** Conducted performance testing of in-house LLMs using **BitNet** and **LLMPerf**; analyzed throughput and latency across 10,000+ API call records to optimize system efficiency
- **Algorithm Implementation:** Developed core algorithms for a Large Language Model Safety Evaluation Platform, integrating 15 jailbreak attack algorithms (e.g., GCG, AutoDAN) and 2 evaluation algorithms
- **Automated Red-teaming:** Conducted **1000+ automated jailbreak evaluations** on commercial LLM APIs, generating **20+ statistical reports** analyzing attack success rates and model robustness

Publications

Open-ended Video Game Glitch Detection with Agentic Reasoning and Temporal Grounding (Under review) [Preprint link] — Apr 2026

Muyang Zheng, Tong Zhou, Geyang Wu, Zihao Lin, Haibo Wang, Lifu Huang*

MIST: Jailbreaking Black-box Large Language Models via Iterative Semantic Tuning (Under review) [Preprint link] — June 2025

Muyang Zheng, Yuanzhi Yao*, Changting Lin, Caihong Kai, Yanxiang Chen, Zhiquan Liu

Open-source Projects

Monocular Depth Estimation Model Fine-tuning [Project link] — Nov 2025

- **Efficient Fine-tuning:** Fine-tuned the MiDaS (dpt_large_384) model using LoRA on the NYU Depth V2 dataset, developing a specialized PyTorch training pipeline
- **Precision Improvement:** Outperformed zero-shot baselines by **reducing prediction error by 15.5%** and improving accuracy from 0.796 to 0.831

Jailbreak Evaluation Framework [Project link] — May 2025

- **Enterprise Safety Solution:** Built a comprehensive platform using **Flask**, **SQLite**, and **Vue.js** to test, evaluate, visualize, and analyze jailbreak attacks on closed-source LLMs
- **Algorithm Integration:** Implemented our proposed MIST jailbreak method within the platform, demonstrating a high attack success rate in empirical evaluations
- **Recognition:** Awarded Outstanding Graduation Project in Anhui Province

Skills

AI & Optimization: LangGraph, BitNet, LLMPerf, LoRA Fine-tuning, Prompt Engineering, PyTorch

Coding: Python, Java, SQL, JavaScript, C++, C

Frameworks & Tools: Flask, Vue.js, Spring Boot, Git, Claude Code, LaTeX, MATLAB

Honors

Third-class Scholarship & Merit Student in 2021-2022 Academic Year
Outstanding Member of the Student Union in 2021-2022 Academic Year