Jingyang Zhang

zhjy227@gmail.com | +1 (984) 245-5792 | https://zjysteven.github.io/ | Google Scholar

EDUCATION

Duke University, Durham, NC, USA — Aug 2019 – Dec 2024

- Ph.D. in Electrical and Computer Engineering
  - Advisor: Prof. Yiran Chen, Prof. Hai (Helen) Li
  - Focus: distribution shifts, adversarial robustness, AI safety

Tsinghua University, Beijing, China — Sep 2015 – Jul 2019

- B.S. in Electronic Engineering

PROFESSIONAL EXPERIENCE

Research Scientist, *Virtue AI*, San Francisco, CA — Aug 2025 – Current

- Research and development on LLM **jailbreaking** and **guardrail**
- Research and development on **redteaming** for LLM agent systems

Machine Learning Engineer, *Sciforium*, Mountain View, CA — Jan 2025 – Aug 2025

- In-house JAX/Flax implementation of transformers with efficiency enhanced components such as **non-complex RoPE**, advanced **KV cache** implementation, **MoE**, and **FP8 GEMM**
- Design and implementation of **byte-based LLMs** with native multi-modal support
- Implementation of scalable uni- and multi-modal data transformation pipeline (raw data to LLM input)
- **LLM pre-training** across 64 GPUs on 8 nodes with JAX distributed framework and sharding mechanism

Machine Learning Engineer Intern, *Tesla*, Palo Alto, CA — May 2023 – Sep 2023

- Implemented state-of-the-art deep learning models and algorithms for trajectory prediction, showcasing the efficacy of this method over baselines with proof-of-concept experiments in different scenarios.

Machine Learning Research Intern, *Bosch Center for AI*, Pittsburgh, PA — Jun 2022 – Dec 2022

- Developed of a "universal" adversarial defense using diffusion model that is robust to both ℓ_p (digital) and patch (physical) adversarial attacks against images. Demonstrated the effectiveness and potential of the defense through extensive experiments, which resulted in a patent.

OPEN-SOURCE SOFTWARE

- **lmms-finetune:** Lightweight codebase for fine-tuning various multimodal (vision) LLMs (**356 Stars**)
- **VLM-Visualizer:** Visualizing the attention of vision LLMs (LLaVA) (**256 Stars**)
- **OpenOOD:** Large-scale, unified evaluation platform for out-of-distribution detection (**1k+ Stars**)

TECH STACK

Deep Learning Framework and Library: PyTorch, JAX, Flax, transformers, diffusers

Programming Language and Tool: Python, Bash, Bazel (Blaze), Docker, Git

PUBLICATIONS

SELECTED CONFERENCE AND JOURNAL PAPERS

- Min-K%++: Improved Baseline for Detecting Pre-Training Data from Large Language Models
  - *ICLR'25 Spotlight* | [paper] [code] [project page]
- OpenOOD v1.5: Enhanced Benchmark for Out-of-Distribution Detection
  - *Journal of Data-Centric Machine Learning Research, NeurIPS'23 DistShift Workshop Oral* | [paper] [code]
- Which Agent Causes Task Failures and When? On Automated Failure Attribution of LLM Multi-Agent Systems
  - *ICML'25 Spotlight* | [paper] [code]
- Unsolvable Problem Detection: Evaluating Trustworthiness of Vision Language Models
  - *ACL'25, ICLR'24 R2FM Workshop* | [paper] [code]
- Mixture Outlier Exposure: Towards Out-of-Distribution Detection in Fine-Grained Environments
  - *WACV'23* | [paper] [code]
- Privacy Leakage of Adversarial Training Models in Federated Learning Systems
  - *CVPR'22 The Art of Robustness Workshop Oral* | [paper] [code]
- DVERGE: Diversifying Vulnerabilities for Enhanced Robust Generation of Ensembles
  - *NeurIPS'20 Oral* | [paper] [code]