Jingyang Zhang

zhjy227@gmail.com | +1 (984) 245-5792 | Webpage | Google Scholar | Github | LinkedIn

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

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

- B.S. in Electronic Engineering

PROFESSIONAL EXPERIENCE

Research Scientist - Post-Train, AI Safety @ *Virtue AI* — Aug 2025 – Current

- Owned multiple **end-to-end post-train pipelines (SFT + RL)** for various guardrails (prompt, action, code vulnerability), including data curation, reasoning trace construction, training, evaluation, and efficiency improvements
- Led **LLM safety evaluation and red-teaming** for real-world enterprise agents, delivering actionable failure analysis across large-scale deployments (Fortune 500-level customers and frontier labs)
- Designed and deployed **agent emulation and adversarial testing frameworks**, including co-developing an internal framework featuring large-scale **real-world sandbox environments** to study emergent behavior, misuse risks, and failure modes in tool-using LLM agents

Research Scientist - Foundation Models, Pre-Train @ *Sciforium* — Jan 2025 – Aug 2025

- Led design of a **purpose-built native multimodal LLM** treating **text, image, video, and audio as first-class modalities** — unified byte-level representation with no modality-specific encoders, targeting coherent cross-modal understanding and generation
- Designed **model architectures and efficient components** (non-complex RoPE, optimized KV cache, FP8 GEMM) enabling scalable training and inference across all modalities
- Built **end-to-end pretraining pipelines** with scalable data transformation from raw multimodal inputs (text, image, video, audio) to unified byte representations
- Conducted **distributed pretraining across 64 GPUs on 8 nodes** using JAX

PUBLICATIONS

SELECTED CONFERENCE AND JOURNAL PAPERS

- 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]
- 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]
- 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]

OPEN-SOURCE SOFTWARE

- **OpenClaw-RL:** Training agents simply by talking through OpenClaw setup (**4k+ Stars**)
- **lmms-finetune:** Lightweight codebase for fine-tuning various multimodal (vision) LLMs (**368 Stars**)
- **VLM-Visualizer:** Visualizing the attention of vision LLMs (LLaVA) (**284 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