Jiawei ZHANG

jwz@uchicago.edu ◇ javyduck.github.io
(+1) 217-200-3511 ◇ Crerar 283, 5730 S Ellis Ave, Chicago, IL 60637

Education

University of Chicago — Advisor: *Prof.* Bo Li
Ph.D. in Computer Science, 4.0/4.0 — *Sept. 2024 – 2026/2027*

University of Illinois Urbana-Champaign — Advisor: *Prof.* Bo Li
Ph.D. in Computer Science, 4.0/4.0 — *Aug. 2023 – May. 2024*
M.S. in Computer Science — *May. 2023*

Zhejiang University — Hangzhou, China
Bachelor of Engineering — *Sept. 2017 – Jun. 2021*

Research Interest:

I study **Safe AGI**, with a focus on **making LLMs and LLM agents reliable at scale**. My research runs as a closed loop: **(i) scalable red-teaming** to elicit realistic, long-horizon failures. I participated in internal red-teaming evaluations of **OpenAI o1**, **Google DeepMind**, and **ElevenLabs TTS** with Virtue AI. In parallel, I study **agentic reliability** for LLM agents, focusing on agent-specific vulnerabilities (e.g., prompt injection and malicious action execution) as well as **environment security construction** for agentic systems; **(ii) defense & alignment** that transforms safety requirements into actionable safeguards, spanning **certified robustness** for worst-case guarantees and **interpretability-guided defenses** that translate mechanistic insights into generalizable mitigations; and **(iii) better safety-aware reasoning for autonomous driving** by incorporating richer world knowledge via reinforcement learning for vision–language–action models.

Together, I aim to build **scalable red-teaming pipelines for LLM agents** and develop **practical patches** that improve safety awareness **while maintaining scalability**. Currently, I am also an Anthropic Fellow for AI Safety Research.

Selected Publication (* denotes co-first authorship)

- **Jiawei Zhang**, Andrew Estornell, David D. Baek, Bo Li, Xiaojun Xu. Any-Depth Alignment: Unlocking Innate Safety Alignment of LLMs to Any-Depth. *International Conference on Learning Representations (ICLR) 2026.* [arXiv]
- Zhaorun Chen*, Xun Liu*, Mintong Kang, **Jiawei Zhang**, Minzhou Pan, Shuang Yang, Bo Li. ARMs: Adaptive Red-Teaming Agent against Multimodal Models with Plug-and-Play Attacks. *International Conference on Learning Representations (ICLR) 2026.* [arXiv]
- **Jiawei Zhang**, Yang Yang, Kaushik Rangadurai, Tao Liu, Minhui Huang, Yiping Han, Bo Li, Shuang Yang GraphQ-LM: Scalable Graph Representation for Large Language Models via Residual Vector Quantization. *[arXiv]*
- Mintong Kang*, Zhaorun Chen*, Chejian Xu*, **Jiawei Zhang***, Chengquan Guo*, Minzhou Pan, Ivan Revilla, Yu Sun, Bo Li. GuardSet-X: Massive Multi-Domain Safety Policy-Grounded Guardrail Dataset. *NeurIPS 2025.* [arXiv]
- **Jiawei Zhang**, Shuang Yang, Bo Li. UDora: A Unified Red Teaming Framework Against LLM Agents by Dynamically Leveraging Their Own Reasoning. *International Conference on Machine Learning (ICML) 2025.* [arXiv]
- **Jiawei Zhang**, Xuan Yang, Taiqi Wang, Yu Yao, Aleksandr Petiushko, Bo Li. SafeAuto: Knowledge-Enhanced Safe Autonomous Driving with Multimodal Foundation Models. *International Conference on Machine Learning (ICML) 2025.* [arXiv]
- Chejian Xu*, **Jiawei Zhang***, Zhaorun Chen*, Chulin Xie*, Mintong Kang*, Zhuowen Yuan*, Chenhui Zhang, Lingzhi Yuan, Yi Zeng, Peiyang Xu, Chengquan Guo, Andy Zhou, ..., Zidi Xiong, Zinan Lin, Dan Hendrycks, Dawn Song, Bo Li. MMDT: Decoding the Trustworthiness and Safety of Multimodal Foundation Models.*International Conference on Learning Representations (ICLR) 2025.* [arXiv]
- **Jiawei Zhang**, Chejian Xu, Yu Gai, Freddy Lecue, Dawn Song, Bo Li. KnowHalu: Hallucination Detection via Multi-Form Knowledge Based Factual Checking. *ICLR 2025 Workshop on Foundation Models in the Wild.* [arXiv]

- Bowen Jin, Chulin Xie, **Jiawei Zhang**, Kashob Kumar Roy, Yu Zhang, Zheng Li, Ruirui Li, Xianfeng Tang, Suhang Wang, Yu Meng, Jiawei Han. Graph Chain-of-Thought: Augmenting Large Language Models by Reasoning on Graphs. *Findings of the Association for Computational Linguistics (ACL) 2024.* [arXiv]
- **Jiawei Zhang**, Chejian Xu, Bo Li. ChatScene: Knowledge-Enabled Safety-Critical Scenario Generation for Autonomous Vehicles. *Conference on Computer Vision and Pattern Recognition (CVPR) 2024.* [arXiv]
- **Jiawei Zhang**, Tianyu Pang, Chao Du, Yi Ren, Bo Li, Min Lin. MMCBench: Benchmarking Large Multimodal Models against Common Corruptions. [arXiv]
- **Jiawei Zhang**, Zhongzhu Chen, Huan Zhang, Chaowei Xiao, Bo Li. DiffSmooth: Certifiably Robust Learning via Diffusion Models and Local Smoothing. *32th USENIX Security Symposium 2023.* [arXiv]
- **Jiawei Zhang**, Linyi Li, Ce Zhang, Bo Li. CARE: Certifiably Robust Learning with Reasoning via Variational Inference. *IEEE Conference on Secure and Trustworthy Machine Learning (SatML) 2023.* [arXiv]
- Zhuolin Yang*, Zhikuan Zhao*, Boxin Wang, **Jiawei Zhang**, Linyi Li, Hengzhi Pei, Bojan Karlas, Ji Liu, Heng Guo, Ce Zhang, Bo Li. Improving Certified Robustness via Statistical Learning with Logical Reasoning. *Advances in Neural Information Processing Systems (NeurIPS) 2022.* [arXiv]
- Linyi Li, **Jiawei Zhang**, Tao Xie, Bo Li. Double Sampling Randomized Smoothing. *International Conference on Machine Learning (ICML) 2022.* [arXiv]
- **Jiawei Zhang***, Linyi Li*, Huichen Li, Xiaolu Zhang, Shuang Yang, Bo Li. Progressive-Scale Boundary Blackbox Attack via Projective Gradient Estimation. *International Conference on Machine Learning (ICML) 2021.* [arXiv]

Industry Research Experience

NVIDIA — *Oct 2025 – Present*
ML Research Intern, Santa Clara — *Advised by Dr. Boris Ivanovic and Prof. Marco Pavone*

- Exploring improved reinforcement learning methods for reasoning to reduce minADE.
- Developing enhanced reasoning capabilities for autonomous driving by leveraging world models.

ByteDance Seed Research — *June 2025 – Sept 2025*
Responsible AI Research Intern, San Jose — *Advised by Dr. Xiaojun Xu and Dr. Hang Li*

- Developed a scalable, model-agnostic defense for LLMs based on safety-signature tokens, effective against deep prefill attacks with context lengths exceeding 3,000 tokens.
- Demonstrated robustness to in-the-wild jailbreaks (AutoDAN, PAIR) while maintaining 0% false positive rate on standard benchmarks (MMLU, MATH, BBH, AlphaEval).

Meta AI — *Sept 2024 – June 2025*
GenAI Research Collaborator (External) — *Advised by Dr. Shuang Yang*

- Built an RVQ-based graph tokenization module that converts continuous node features into compact discrete tokens aligned with LLM embeddings, enabling scale-free tokenization on 100k+ node graphs.
- Matched or exceeded leading baselines on ogbn-arxiv, Cora, and PubMed while cutting storage from gigabytes to megabytes; released reproducible code and ablations.

Nuro AI — *May 2024 – Aug 2024*
Machine Learning Research Intern (Pathfinder), Mountain View — *Advised by Dr. Aleksandr Petiushko*

- Fine-tuned a multimodal LLM for video-conditioned autonomous driving to output high-level action plans and low-level control signals, unifying perception, reasoning, and control.
- Developed a multimodal retrieval-augmented training pipeline and a Markov Logic Network that encodes first-order traffic rules to verify and improve the safety of LLM-proposed actions.
- Proposed a position-dependent cross-entropy loss (PDCE) to stabilize and improve numeric control predictions when represented as text.

Sea AI Lab — *May 2023 – Aug 2023*
Machine Learning Research Intern, Singapore — *Advised by Dr. Tianyu Pang & Dr. Chao Du*

- Conducted evaluations on cross-modal models (e.g., Stable Diffusion, Whisper) to assess their consistency under a range of common data corruptions.
- Developed a rigorous benchmark for assessing the self-consistency of these models. The benchmark was designed to provide a comprehensive understanding of model behavior, incorporating a wide range of scenarios and inputs to measure their resilience and accuracy.

Workshops and Competitions

- **CLAS 2024: The Competition for LLM and Agent Safety, NeurIPS 2024**
- **Secure and Safe Autonomous Driving (SSAD) Workshop and Challenge, CVPR 2023**

Teaching

CS 307 - Modeling and Learning in Data Science (Spring 2022)
- Teaching Assistant with *Prof.* Bo Li and *Prof.* David Forsyth

SELECTED HONOR & AWARDS

Anthropic Fellow for AI Safety Research	*Jan. 2026*
Graduate Conference Travel Award	*March. 2023*
Accumulative Research Innovation & Entrepreneurship Index: 1st Department-wide	*June. 2021*
Meritorious Winner, MCM COMAP's Mathematical Contest in Modeling	*Apr. 2020*
First Prize, China Harbour Scholarship	*Jan. 2020*
First Prize, The Chinese Mathematics Competitions, Zhejiang Province	*Nov. 2018*

Professional Service

Session Chair — ICLR 2025
Top Reviewer — NeurIPS 2025
Program Committee— ICML, NeurIPS, ICLR, CVPR, ECCV, AISTATS, AAAI, COLM, ACL, JIMR, etc.