Jiawei ZHANG

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

Education

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

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

Zhejiang University (ZJU) — Hangzhou, China
Bachelor of Engineering (Excellent Class) — *Jun. 2021*

Research Interest:

My current research predominantly centers on **trustworthy large language models (LLMs)**. I'm particularly interested in enhancing their trustworthiness by mitigating issues like hallucination, using external knowledge sources as leverage. While my foundation in **robustness, privacy, fairness, and explainability** remains intact, my renewed focus aims at the integration of these principles into the development and understanding of LLMs, thereby ensuring they align more closely with human values and expectations.

Selected Publication

- **Jiawei Zhang**, Chejian Xu, Yu Gai, Freddy Lecue, Dawn Song, Bo Li. KnowHalu: Hallucination Detection via Multi-Form Knowledge Based Factual Checking. [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.* [paper]
- **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.* [paper]
- **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.* [paper]
- **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.* [paper]
- 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 (NIPS) 2022.* [paper]
- Linyi Li, **Jiawei Zhang**, Tao Xie, Bo Li. Double Sampling Randomized Smoothing. *International Conference on Machine Learning (ICML) 2022.* [paper]
- **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.* [paper]

Research Experience

Detecting the Hallucination for LLM — *Sept. 2023 – Present*
Research Assistant | Secure Learning Lab, UIUC — *Advised by* Prof. *Bo Li*

- Developed methods to evaluate the self-consistency, knowledge consistency, and logical consistency of text generated by LLMs.
- Leverage a retrieval system to externally cross-check claims made in LLM responses against trusted knowledge bases, ensuring the validity of generated content.

Safety-Critical Driving Scenario Generation Based on LLM *Oct. 2022 – March. 2023*

Research Assistant | Secure Learning Lab, UIUC *Advised by* Prof. *Bo Li*

- Aim to enrich the safety-critical testing scenarios in SafeBench [link] for Autonomous Vehicles.
- Train an adversarial agent (vehicle/pedestrian/bicyclist) via specifically designed multi-agent reinforcement learning to cause the unexpected collision of the ego vehicle.

Enhance Robustness via Diffusion Models and Local Smoothing *Jun. 2022 – Oct. 2022*

Research Assistant | Secure Learning Lab, UIUC *Advised by* Prof. *Bo Li* & *Postdoc. Huan Zhang (CMU)*

- Prove that the "one-shot" denoising of DDPM can approximate the mean of the generated posterior distribution by continuous-time diffusion models, which is an approximation of the original instance under mild conditions.
- Propose a local smoothing technique based on the diffusion models, achieve the **SOTA 43.6%** certified accuracy on CIFAR-10 under ℓ_2 radius 1.0 and the **SOTA 53.0%** certified accuracy on ImageNet under the ℓ_2 radius 1.5.

Certifiably Robust Learning with Reasoning via Variational Inference *May. 2022 – Sep. 2022*

Research Assistant | Secure Learning Lab, UIUC *Advised by* Prof. *Bo Li* & *Prof. Ce Zhang (ETH Zürich)*

- Propose a scalable and certifiably robust learning with reasoning pipeline CARE, which is able to integrate knowledge rules to enable reasoning ability for reliable prediction
- Propose an efficient Expectation Maximization (EM) algorithm to approximate the reasoning based on Markov Logic Network (MLN) via variational inference using Graph Convolutional Network (GCN).
- Extensive experiments on different datasets show that the proposed method achieves significantly higher certified robustness than SOTA baselines, for example, the certified accuracy could be improved from 36.0% (SOTA) to 61.8% under ℓ_2 radius 2.0 on AwA2.

Boundary Blackbox Attack via Projection Based Gradient Estimation *Jun. 2020 – May. 2021*

Research Intern | Cooperate with Ant Financial *Advised by* Prof. *Bo Li*

- Propose the first theoretical framework to analyze boundary blackbox attacks with general projection functions.
- Characterize the key characteristics and trade-offs for a good projective gradient estimator.
- Propose Progressive-Scale based projective Boundary Attack via progressively searching for the optimal scale in a self-adaptive way under spatial, frequency, and spectrum scales.
- The extensive experiments show that our method outperforms the state-of-the-art boundary attacks on MNIST, CIFAR-10, CelebA, and ImageNet against different blackbox models and an online API (MEGVII Face++).

Industry

Nuro *May 2024 – Aug 2024*

Machine Learning Research Intern, Mountain View *Advised by Aleksandr Petiushko*

- Fine-tuning a Multimodal LLM to provide high-level action descriptions and low-level control signals for autonomous driving based on video input.
- Proposing a new RAG training mechanism for multimodal retrieval and using a Markov logic network to achieve post-safety verification. This leverages first-order traffic rules to improve the safety of the high-level actions provided by the LLM.
- Proposing a new CE loss to make the numerical values predicted by the LLM more accurate and stable.

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.

Teaching

CS 307 - Modeling and Learning in Data Science (Spring 2022)

- Teaching Assistant with *Prof.* Bo Li and *Prof.* David Forsyth