Jianshu Zhang

+(86) 15361592875 jianshu.zhang@whu.edu.cn Personal Homepage

EDUCATION

Wuhan University (WHU) Sept. 2021-Jun. 2025 (Expected)

B.Eng. in Information Security

- Cummulative GPA: 3.84/4.0 (90.43/100), Class Rank: 1/26
- Junior-Year GPA: 3.96/4.0 (92.22/100), Deparment Rank: 1/146
- Selected Awards: **China National Scholarship** (2024, the highest honor for undergraduates in China, 0.2% national-wide); WHU Scholarhip for Academic Excellence (2024, 2023& 2022); Merit Student (2024, 2023 & 2022);

PUBLICATIONS

*Co-first authors contributed equally to this work.

Pi, R. *, **Zhang, J.** *, Zhang, J., Pan, R., Chen, Z., & Zhang, T. (2024). Image textualization: An automatic framework for generating rich and detailed image descriptions. 2024 Annual Conference on Neural Information Processing Systems (**NeurIPS 2024**). Accepted.

Pi, R.*, Han, T.*, **Zhang, J.***, Xie, Y., Pan, R., Lian, Q., Dong, H., Zhang, J., & Zhang, T. (2024). MLLM-Protector: Ensuring MLLM's safety without hurting performance. 2024 Conference on Empirical Methods in Natural Language Processing (**EMNLP 2024**). Accepted.

Shum, K.*, Xu, M.*, **Zhang, J.***, Chen, Z., Diao, S., Dong, H., Zhang, J., & Raza, M. O. (2024). FIRST: Teach a reliable large language model through efficient trustworthy distillation. 2024 Conference on Empirical Methods in Natural Language Processing (**EMNLP 2024**). Accepted.

Yao, D. *, **Zhang, J.***, Harris, I. G., & Carlsson, M. (2024). FuzzLLM: A novel and universal fuzzing framework for proactively discovering jailbreak vulnerabilities in large language models. 2024 IEEE International Conference on Acoustics, Speech and Signal Processing ***(ICASSP 24)*** (pp. 4485-4489). doi: 10.1109/ICASSP48485.2024.10448041

Zhang, J. *, Fu, Y. *, Peng, Z. *, Yao, D., & He, Kun. (2024). CORE: Mitigating Catastrophic Forgetting in Continual Learning through Cognitive Replay. *2024 Cognitive Science (**CogSci 2024**)*. [Oral]

IN SUBMISSION

Zhang, J.*, Rong, X.*, He, K., & Ye, M. Client as navigator: Emphasizing the role of clients in federated class-continual learning. *Submitted to the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2025)*

Pi, R.*, **Zhang, J.***, Han, T., Zhang, J., Pan, R., & Zhang, T. PVIT: Personalized visual instruction tuning. *Submitted to the International Conference on Learning Representations (ICLR 2025)*

RESEARCH EXPERIENCES

Tong Zhang's Research Group, UIUC (Remote) Feb. 2024-Present

Supervisor: Prof. Tong Zhang;

➢ **Image textualization: An automatic framework for generating rich and detailed image descriptions**

- Leveraged multimodal LLM to create the Reference Description, which, despite lacking details and containing hallucinations, provides the basic structure not only for the visual information but also for the linguistic expression.
- Implemented an object extraction process from both textual and visual inputs and applied a hallucination filter to determine which objects to omit or include in the descriptions.
- Enhanced image content textualization accuracy by using depth maps and object masks for precise annotation of objects to be added, and employed an LLM to recaption the Reference Description with fine-grained visual textualization.
- Verified that the LLaVA-7B, benefiting from training on the image textualization-curated descriptions, acquires improved capability to generate detailed and extensive image descriptions with less hallucination.

➢ **MLLM-Protector: Ensuring MLLM's safety without hurting performance**

- Identified vulnerabilities in Multimodal LLMs related to malicious image inputs and the challenges of addressing these with Supervised Fine-Tuning (SFT).
- Introduced a novel defense approach, termed MLLM-Protector, to address the alignment task via a divide-and-conquer strategy, serving as a plug-and-play module applicable to any MLLM.
- Developed and released Safe-Harm-10K, a dataset for training harm detectors and detoxifiers.
- Empirically demonstrated that MLLM-Protector effectively mitigates harmful outputs from malicious inputs while maintaining model performance.

Data Security Lab, Wuhan University Sept. 2023-Present

Supervisor: Prof. Kun He

➢ **Overcoming Catastrophic Forgetting in Continual Learning via Cognitive Replay**

- Introduced the Adaptive Quantity Allocation strategy that divides the data in the replay buffer into Targeted Recall Data and Spaced Repetition Data for tasks with varying forgetting rates, facilitating focused recollection and ensuring minimum effective data for spaced repetition, respectively.
- Designed an algorithm named Quality-Focused Data Selection to ensure uniform distribution of data in the buffer within the feature space and effective representation of feature centers, thus guaranteeing the inclusion of representative data that best encapsulates the features of each task within the buffer.
- Achieved an average accuracy of 37.93% on split-CIFAR10, surpassing the best baseline method by 6.52%; enhanced the poorest-performing task by 6.03% compared to the top baseline; and outperformed all baselines on both the split-MNIST and split-CIFAR100 datasets.

UCInspire Summer Research Program, UC Irvine (Remote) Jul. 2023-Sept. 2023

Supervisor: Prof. Ian G. Harris

➢ **A Fuzzing Framework to Uncover Jailbreak Vulnerabilities of LLMs**

- Utilized the fuzzing technique to automatically uncover jailbreak vulnerabilities in LLMs.
- Decomposed a prompt into 3 basic components that could be combined to craft numerous test samples and devised jailbreak base classes and templates capable of merging into more powerful combo attacks (with greater diversity in class granularity)
- Proved the effectiveness and comprehensiveness of our testing framework through comparisons with existing jailbreak prompts and jailbreaking methods.
- Evaluated the attack efficiency of prompts produced by FuzzLLM across 6 open-sourced and 2 commercial LLMs, achieving an impressive success rate of up to **85.18%.**

FIRST: Teach A Reliable LLM through Efficient Trustworthy Distillation Mar.2024-Jun. 2024

Independent Project

- Discovered that LLMs exhibit "concentrated knowledge" and "tuning-induced mis-calibration," providing insights into developing trustworthy models.
- Proposed a framework called FIRST, which maximizes the effectiveness and trustworthiness of a relatively small portion of knowledge transferred from the teacher by "trustworthy maximization" to obtain a reliable student model.
- Validated through extensive experiments that models trained with FIRST consistently achieved high trustworthiness across various settings.

INTERNSHIP

Tencent, Shenzhen, China Jun.2023-Sept.2023

Intern at IEG (Interactive Entertainment Group)

- Assisted cross-functional teams to deliver a secure gaming platform for internal and external gaming companies.
- Facilitated the integration of game security features for client companies by managing requirement evaluation, resource allocation, and version control throughout the development, testing, and deployment phases
- Identified and addressed operational challenges within the Agile development framework, such as optimizing requirement clarity and improving communication channels for timely synchronization of staff schedules

SKILLS

- Programming: Python, C, HTML, CSS, PHP, Javascript, SQL
- Tools: VS Code, PyTorch, Jupyter Notebook, LATEX, Git