Jianshu Zhang

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

EDUCATION

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

- **School of Cyber Science and Engineering** GPA: 3.83/4.0 (90.1/100)
- **Awards:** 2022-2023 Scholarship for Academic Excellence, 2022-2023 Merit Student, 2021-2022 Scholarship for Academic Excellence; 2021-2022 Merit Student.

PUBLICATIONS

*Co-first authors contributed equally to this work.

- **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. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (**ICASSP 24**)* (pp. 4485-4489). [Poster]
- **Zhang, J**. *, Fu, Y. *, Peng, Z. *, Yao, D., & He, Kun. CORE: Mitigating Catastrophic Forgetting in Continual Learning through Cognitive Replay. *2024 Cognitive Science (**CogSci 2024**)*. **[Oral]**

PAPERS IN SUBMISSION

- Pi, R. *, **Zhang, J.** *, Zhang, J., Pan, R., Chen, Z., & Zhang, T. (2024). Image Textualization: An Automatic Framework for Creating Accurate and Detailed Image Descriptions. *In submission to NeurIPS*.
- 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. *In submission to EMNLP.*
- Shum, K*, Xu, M. *, **Zhang, J.***, Chen, Z., Diao, S., Dong, H., Zhang, J., Raza, O. (2024). FIRST: Teach A Reliable Large Language Model Through Efficient Trustworthy Distillation. *In submission to EMNLP.*

RESEARCH EXPERIENCES

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

Supervisor: Prof. Tong Zhang

Project 1: An Automatic Framework for Creating Accurate 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.

Project 2: Ensuring the Safety of Multimodal LLMs without Compromising 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.
- Develop and release 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.

Trustworthy Distillation of Large Language Models for Enhanced Reliability Mar. 2024-Jun. 2024

Independent Research & Project Leader

- Discovered that LLMs exhibit “concentrated knowledge” and “tuning-induced mis-calibration,” providing insights into developing trustworthy models.
- Proposed FIRST, a framework that 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.

Data Security Lab, Wuhan University Sept. 2023-Present

Supervisor: Prof. Kun He

Project: 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

Project: A Fuzzing Framework to Uncover Jailbreak Vulnerabilities of LLMs

- Leveraged the traditional fuzzing technique to 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%.**

SELECTED PROJECTS

Big Data Analysis based on arXiv Dec. 2023

- Utilized Hadoop and Spark to analyze and process approximately 600,000 articles from arXiv database
- Analyzed features based on time (yearly, monthly, intervals between first and last submissions) and explored relationships and influences among different categories.
- Applied BERT for semantic feature extraction and PCA for dimensionality reduction, implemented K-means clustering to achieve a more fine-grained classification of a vast array of articles.

Basketball Statistics Analysis System May. 2023

- Developed a comprehensive basketball statistics analysis system that enables efficient analysis of game data and team information, incorporating a wide range of functionalities for both users and administrators.
- Implemented front-end and back-end on Linux, ensuring seamless integration and smooth user experience.
- Analyzed the security and reliability of the system, employed hash code to store sensitive user data, applied the correlation function to prevent SQL injection, and incorporated mechanisms for system data rollback and backup.

Similarity and Difficulty Level Judging Program for Math Problems Apr. 2023

- Developed a classification strategy for elementary school math application problems, enabling automatic measurement of problem similarity and assessment of difficulty.
- Trained a K-nearest neighbor classification model using the output of a weighted feature vector-based support vector machine, which ultimately achieved an 88% accuracy rate in classification.
- Won the **First Prize (Top 4.5% Nationwide)** in the Huazhong Contest in Mathematical Modeling

SKILLS

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