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

- Cumulative GPA: 3.84/4.0 (90.43/100), Class Rank: 1/26
- Selected Awards: China National Scholarship (2024); WHU Scholarship for Academic Excellence (2024, 2023& 2022); WHU Merit Student (2024, 2023 & 2022); Advanced Individual in Scientific Innovation (2023)

PUBLICATIONS

(denotes joint first-authors.)*

[1] Pi, R.*, **Zhang, J.***, Zhang, J., Pan, R., Chen, Z., & Zhang, T. *"Image Textualization: An Automatic Framework for Generating Rich and Detailed Image Descriptions"*. Accepted by **NeurIPS 2024**.

[2] Pi, R.*, Han, T.*, **Zhang, J.***, Xie, Y., Pan, R., Lian, Q., Dong, H., Zhang, J., & Zhang, T. *"MLLM-Protector: Ensuring MLLM's Safety without Hurting Performance"*. Accepted by **EMNLP 2024**.

[3] Shum, K.*, Xu, M.*, **Zhang, J.***, Chen, Z., Diao, S., Dong, H., Zhang, J., & Raza, M. O. *"FIRST: Teach a Reliable Large Language Model through Efficient Trustworthy Distillation"*. Accepted by **EMNLP 2024**.

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

[5] Yao, D.*, **Zhang, J.***, Harris, I. G., & Carlsson, M. *"FuzzLLM: A Novel and Universal Fuzzing Framework for Proactively Discovering Jailbreak Vulnerabilities in Large Language Models"*. Accepted by **ICASSP 24**.

[6] **Zhang, J.***, Rong, X.*, He, K., & Ye, M. *"Clients As Navigators: Towards Native Federated Continual Learning"*. **Under Review of CVPR 2025**.

[7] Pi, R.*, **Zhang, J.***, Han, T., Zhang, J., Pan, R., & Zhang, T. *"Personalized Visual Instruction Tuning"*. **Under Review of ICLR 2025**.

[8] Zhang, J.*, **Zhang, J.***, Li, Y.*, Pi, R., Pan, R., Liu, R., Zheng, Z., & Zhang, T. *"Bridge-Coder: Unlocking LLMs' Potential to Overcome Language Gaps in Low-Resource Code"*. **Under Review of ARR**.

RESEARCH EXPERIENCES

MIT Media Lab & HKUST NLP Group Oct. 2024-Present

Supervisor: Prof. Paul Liang and Dr. Yi R. (May) Fung

➢ **PC-Bench: How Multimodal Large Language Models Perceive then Connect Fine-grained Visual Cues (on going…)**

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**

- Proposed the Image Textualization (IT) framework that automatically generates detailed image descriptions, leveraging the coarse descriptions from MLLMs, the fine-grained perception of visual experts enhanced by depth maps and object masks, and the reasoning power of LLMs to synthesize accurate, detailed descriptions.
- Verified that the MLLM, benefiting from training on the image textualization-curated descriptions, acquires improved capability to generate detailed and extensive image descriptions with less hallucination.
- Using IT, we curate a large-scale high-quality image description dataset termed IT-170K to facilitate future research.

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

- Identified vulnerabilities in multi-modal LLMs (MLLMs) arising from malicious visual inputs and challenges in safety alignment using supervised fine-tuning. The continuous nature of image signals limits scenario coverage, and a lack of image-text pairs increases the risk of catastrophic forgetting during safety updates.
- Introduced MLLM-Protector, a novel defense approach that mitigates jailbreak vulnerabilities using a divide-and-conquer strategy, with a detector at the multimodal input level and a detoxifier at the text-based output level, while preserving model performance on general tasks.

➢ **Personalized Visual Instruction Tuning**

- Identified limitations in MLLMs that fail to conduct personalized dialogues targeting at specific individuals, exhibit "face blindness", limiting their application in tailored settings.
- Introduced Personalized Visual Instruction Tuning (PVIT), a novel data curation and training framework that enables MLLMs to perceive personalized inputs as in-context prefixes, allowing them to generalize to any individual without requiring additional fine-tuning or model modifications.

- Released the first and the largest personalized visual instruction tuning dataset PVIT-3M and validated its effectiveness, demonstrating substantial personalized performance of the MLLM enhancement after tuning.

➢ **Bridge-Coder: Unlocking LLMs' Potential to Overcome Language Gaps in Low-Resource Code**

- Identified the NL (natural language)-PL (programming language) Gap as the main factor limiting LLM performance in low-resource PLs (LRPLs), preventing LLMs from fully leveraging their inherent capabilities.
- Introduced Bridge-Coder, which first leverages LLMs' intrinsic abilities in high-resource PLs to generate high-quality data for LRPLs, then applies two-step alignment—first assist and then direct— to bridge the NL-PL gap of LRPLs.
- Demonstrated Bridge-Coder's effectiveness through extensive experiments across various LRPLs, providing valuable insights for future research in this underexplored field.

Multimedia Analysis and ReaSoning (MARS) Group, Wuhan University Jun.2024-Oct.2024

Supervisor: Prof. Mang Ye

➢ **Clients As Navigators: Towards Native Federated Continual Learning**

- Introduced a client-centric perspective with two key observations—Client Expertise Superiority and Client Forgetting Variance—highlighting the necessity of positioning clients as navigators in Federated Continual Learning (FCL).
- Proposed CAN, a native FCL approach that leverages the unique role of clients in FL. By using expert-driven data synthesis for old knowledge and client-specific adaptive replay buffer, CAN ensures more efficient and targeted replay.
- Achieved SOTA performance in extensive experiments on CIFAR100, TinyImagenet, and Imagenet100.

Data Security Lab, Wuhan University Sept. 2023- Oct.2024

Supervisor: Prof. Kun He

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

- Introduced Adaptive Quantity Allocation that divides the data in the replay buffer into targeted recall data and spaced repetition data for different tasks, ensuring optimal allocation for tasks with different levels of forgetting.
- Designed Quality-Focused Data Selection to ensure a uniform distribution features of data, effectively representing feature centers and guaranteeing the inclusion of data that best encapsulates each task's features in the buffer.
- Achieved an average accuracy of 37.93% on split-CIFAR10, surpassing the best baseline by 6.52%, with a 6.03% improvement on the poorest-performing task, also outperformed all baselines on split-MNIST and split-CIFAR100.

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**

- Leveraged the fuzzing technique to uncover jailbreak vulnerabilities in LLMs by decomposing a jailbreak prompt into three 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, achieving an impressive success rate of up to 85.18% of prompts produced by FuzzLLM across 6 open-sourced and 2 commercial LLMs.

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

Independent Project

- Discovered that LLMs suffer from miscalibration, where the predicted confidence does not align with the actual likelihood of correctness, emphasizing a key challenge beyond accuracy in building truly trustworthy models.
- Proposed FIRST, a knowledge distillation approach that efficiently transfers a small portion of knowledge from the teacher model, and recalibrates it through 'trustworthy maximization' to produce a more reliable model.
- Validated through extensive experiments that models trained with FIRST consistently achieved high accuracy and well calibration across both in-domain and out-of-domain settings.

SKILLS

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