Zirui Pang

zpang112@connect.hkust-gz.edu.cn Google Scholar Website

Education

Hong Kong University of Science and Technology (Guangzhou) Jan. 2026 - Present
Ph.D. Student · Data Science and Analytics; Advisors: Zixin Zhong and Jiaheng Wei
Research Interests: LLM Unlearning, Efficient Agentic Reinforcement Learning, Embodied AI

University of Illinois Urbana-Champaign Aug. 2024 - Dec. 2025
M.S. · Predictive Analytics & Risk Management

Central University of Finance and Economics Sep. 2019 - Jun. 2023
B.S. · Data Science and Big Data Technology

Internship

Hong Kong University of Science and Technology (Guangzhou) Dec. 2024 - Dec. 2025
Research Assistant · Guangdong, China
Focused on developing more effective and safer LLM unlearning methods.

IRisk Lab, University of Illinois Urbana-Champaign Aug. 2024 - Dec. 2024
Research Assistant · Illinois, USA
Focused on developing and maintaining a risk assessment model for a company in Xiamen, China, by predicting customer repayment dates.

iFlytek Feb. 2022 - Apr. 2022
Assistant Software Development Engineer · Anhui, China
Focused on developing and maintaining Hudi data lake functionality for the Odeon platform.

Selected Publications

OFFSIDE: Benchmarking Unlearning Misinformation in Multimodal Large Language Models 2025
Hao Zheng, **Zirui Pang**, Ling Li, Zhijie Deng, Yuhan Pu, Zhaowei Zhu, Xiaobo Xia, Jiaheng Wei
ACL 2026 Findings

GUARD: Generation-time LLM Unlearning via Adaptive Restriction and Detection 2025
Zhijie Deng, Chris Yuhao Liu, **Zirui Pang**, Xinlei He, Lei Feng, Qi Xuan, Zhaowei Zhu, Jiaheng Wei
ICML 2025 Workshop

Research Projects

LLM Unlearning

LEAF-OPD: Learning to Forget from Privileged Contexts via On-Policy Distillation Research Leader
Overview

- Introduced on-policy distillation for targeted LLM unlearning, using a frozen teacher to supervise student-generated responses at the token level.
- Used privileged refusal instructions for forget queries and ordinary teacher supervision for retain queries; the student requires no privileged context at inference.
- Defined Supervision Signal metrics to diagnose weak teacher guidance and strengthened persistent refusal behavior through lightweight supervised fine-tuning.
- Evaluated on TOFU, RWKU, and MUSE-News, improving forgetting while preserving utility; reported a 34% gain over the second-best method in RWKU Forget Quality.

ReRULE: Replay What Matters: Off-Policy Replay for Efficient LLM Reinforcement Unlearning Research Leader
Overview

- Identified inefficient repeated sampling of easy prompts and underuse of difficult trajectories near the forget/retain boundary in reinforcement unlearning.
- Developed an off-policy replay buffer that stores low-reward rollout groups from early GRPO training and reuses them through importance-sampled updates.
- Derived a tighter hard-case convergence bound than on-policy RULE and evaluated replay on RWKU, MUSE-Books, and TOFU.
- Improved MUSE-Books Retain Quality from 46.3 to 56.2 with 5-11% additional training time; gains were smaller on the easier TOFU setting.

SGA: Label Smoothing Improves Gradient Ascent in LLM Unlearning Research Leader

Overview

- Addressed unstable gradient-ascent unlearning and the resulting degradation of general model utility.
- Proposed Smoothed Gradient Ascent, combining forget data with constructed normal responses through a tunable label-smoothing rate.
- Analyzed how smoothing mitigates divergent updates and provided theoretical guidance for choosing the smoothing rate.
- Validated on TOFU, Harry Potter, and MUSE-News, demonstrating improved forgetting-utility trade-offs over vanilla gradient ascent.

OFFSIDE: Benchmarking Unlearning Misinformation in Multimodal Large Language Models Research Leader

Overview

- Built a multimodal misinformation-unlearning benchmark from football transfer rumors, with 15.68K curated records covering 80 players.
- Designed four test sets to evaluate forgetting, generalization, utility, and robustness, including fine-grained targets and rumors embedded in images.
- Evaluated unlearning baselines and exposed weaknesses against visual rumors, prompt attacks, and relearning of removed misinformation.

Trustworthy AI

Reveal: When VLMs Meet Image Classification: Test Sets Renovation via Missing Label Identification Research Leader

Overview

- Developed REVEAL to identify both incorrect and missing labels in image-classification benchmarks, supporting more reliable model evaluation.
- Combined predictions from LLaVA, BLIP, Janus, and Qwen with label-curation tools and human annotation to address complementary labeling errors.
- Estimated model reliability and applied weighted aggregation, consensus filtering, and normalization to refine candidate label sets.
- Evaluated across six datasets, including CIFAR-10/100 and ImageNet, and analyzed when VLM-based renovation complements conventional label-cleaning methods.

Service

Teaching

Teaching Assistant, UFUG2601 C++ Programming, HKUST-GZ (Fall 2026)

Reviewer

ICLR

Volunteer

CCF YOCSEF Guangzhou Workshop 0905

Skills

Python, PyTorch, LLM post-training

Languages

Chinese (native), English