

ZIJUN WANG

✉ zwang745@ucsc.edu · 🌐 asillycat.github.io · 🐙 asillycat · 🔑 n5wjgV0AAAAJ

🎓 EDUCATION

University of California, Santa Cruz, Santa Cruz, United States 2024.08 – Present

PhD student (Expected graduation date: 06/2029)

Advised by Prof. Cihang Xie at VLAA LAB in Computer Science and Engineering Department

Research interest: AI Safety, Text Generation^{[1][2]}

Zhejiang University, Hangzhou, China 2020.09 – 2024.06

Bachelor of Engineering

Major in Computer Science and Technology, College of Computer Science and Technology

GPA: 3.92/4.00 **Credits:** 217.5 / 170.5

🎓 EXPERIENCE

TikTok Intern San Jose, CA

Data-TnS- Algo-Foundations & Intelligence Service 2025.06 – Present

- Under Supervision of Fengze Liu
- One co-authored paper is submitted to ICLR [Paper]
- Worked on Pretraining Foundation LLM
- Proposed a targeted data selection method (preparing for ICML)

Visiting Research Intern Santa Cruz, CA

VLAA LAB, UC Santa Cruz 2023.08 – 2024.08

- Under Supervision of Prof. Cihang Xie and Prof. Yuyin Zhou
- Worked on **Adversarial Attacks on LLMs & VLLMs**
- One paper is accepted by **TMLR 2025**. [Paper] [Code]
- One co-authored paper is accepted by **ECCV 2024**. [Paper] [Code]
- **Second Place** in both base & large model subtracks of Red Teaming LLM@**NeurIPS 2023**, Torjan Detection Challenge(**Team leader**). [Code]

🎓 SELECTED PUBLICATIONS

STAR-1: Safer Alignment of Reasoning LLMs with 1K Data

Zijun Wang, Haoqin Tu, Yuhan Wang, Juncheng Wu, Jieru Mei, Brian R. Bartoldson, Bhavya Kailkhura, Cihang Xie

Accepted by *Association for the Advancement of Artificial Intelligence (AAAI 2026 (oral))*

TL;DR: This paper introduces STAR-1, a high-quality, just-1k-scale safety dataset specifically designed for LRMs. Built on three core principles – diversity, deliberative reasoning, and rigorous filtering – STAR-1 aims to address the critical needs for safety alignment in LRMs. Experimental results show that fine-tuning LRMs with STAR-1 leads to an average 40% improvement in safety performance across four benchmarks, while only incurring a marginal decrease (e.g., an average of 1.1%) in reasoning ability measured across five reasoning tasks.

AttnGCG: Enhancing Adversarial Attacks on Language Models with Attention Manipulation

Zijun Wang, Haoqin Tu, Jieru Mei, Bingchen Zhao, Yisen Wang, Cihang Xie

Accepted by *Transactions on Machine Learning Research (TMLR 2025)*

TL;DR: This paper introduces an enhanced method that additionally manipulates models' attention scores to enhance the LLM jailbreaking. We term this novel strategy AttnGCG. Empirically, AttnGCG demonstrates consistent performance enhancements across diverse LLMs, with an average improvement of 7% in the Llama-2 series and 10% in the Gemma series. This strategy also exhibits robust attack transferability against both unseen harmful goals and black-box LLMs.

How Many Unicorns Are in This Image? A Safety Evaluation Benchmark for Vision LLMs

Haoqin Tu*, Chenhang Cui*, **Zijun Wang** *, Yiyang Zhou, Bingchen Zhao, Junlin Han, Wangchunshu Zhou, Huaxiu Yao, Cihang Xie (* represents equal contribution)

Accepted by *European Conference on Computer Vision (ECCV 2024)*

TL;DR: This work focuses on the potential of VLLMs in visual reasoning. Different from prior studies, we shift our focus from evaluating standard performance to introducing a comprehensive safety evaluation suite, covering both out-of-distribution (OOD) generalization and adversarial robustness.

AWARDS

- **Second Place** in both base & large model subtracks of Red Teaming LLM@**NeurIPS 2023**, Torjan Detection Challenge(**Team leader**). [Code]
- **National Scholarship (top 0.2% national-wide)** issued by Ministry of Education of the People's Republic of China, 2021
- **Provincial Government Scholarship (top 3%)** of Zhejiang Province, 2023
- **First-class Scholarship (top 3%)** of Zhejiang University, 2021 & 2023