Tianle Gu

gtl23@mails.tsinghua.edu.cn | Github

Education

Tsinghua University, M.Eng. in Electronic Information (GPA: 3.87/4.0) Sept. 2023 – Jun. 2026

- Supervised by Prof. Yujiu Yang, National Scholarship * 1

Hunan University, B.Eng. in Computer Science and Technology Sept. 2019 – Jun. 2023

- GPA: 3.83/4.0 (Top 2%), National Scholarship * 2

Publications

1. **Tianle Gu**, Zeyang Zhou, Kexin Huang, et al. "*MLLMGuard: A Multi-dimensional Safety Evaluation Suite for Multimodal Large Language Models*" (Accepted by NeurIPS 2024)

- (1) We designed a multi-dimensional safety evaluation suite for MLLMs, which includes a bilingual dataset, an inference toolkit, and a lightweight evaluator. (2) We trained GuardRank, a fully automated lightweight evaluator using the annotated dataset, achieving full automation of the evaluation process.

2. **Tianle Gu**, **Zongqi Wang**, Kexin Huang, et al. "*Invisible Entropy: A Safe and Efficient Paradigm for Low-entropy Watermarking*" (Accepted by EMNLP Main Conference (Oral))

- We introduced IE, a novel watermarking framework for low entropy text without origin LLMs.

3. **Tianle Gu**, Kexin Huang, Zongqi Wang, et al. "*Probing the robustness of large language models safety to latent perturbations*" (Submitted to ACL 2026)

- We proposed a novel adversarial attack and defense framework that systematically uncovers and mitigates latent vulnerabilities in large language models.

4. **Tianle Gu**, **Kexin Huang**, Ruilin Luo, et al. "*From Evasion to Concealment: Stealthy Knowledge Unlearning for LLMs*" (Accepted by ACL 2025 Findings)

- We proposed a streamlined and stealthy knowledge unlearning algorithm that enhances forgetting quality while maintaining model utility, preserving NLU and NLG capabilities, and demonstrating resilience to MIA.

5. **Tianle Gu**, Kexin Huang, Lingyu Li, et al. "*From Sparse Decisions to Dense Reasoning: A Multi-attribute Trajectory Paradigm for Multimodal Moderation*" (Submitted to ICML 2026)

- We proposed UniMod, a paradigm shift from sparse binary decisions to dense reasoning trajectories, enabling fine-grained, data-efficient multimodal safety moderation.

6. **Zongqi Wang**, **Tianle Gu**, **Baoyuan Wu**, et al. "*MorphMark: Flexible Adaptive Watermarking for Large Language Models*" (Accepted by ACL 2025 Main)

- We proposed an adaptive, model-agnostic method that resolves the trade-off in LLM watermarking.

Note: Bold indicates first or co-first author.

Projects

ValuePRM – Core Contributor Technical Report

- Led the training of ValuePRM using response-level data to evaluate and verify value-aligned behavior in MLLMs.

ChatZoo Star★ 80+ GitHub

- Developed ChatZoo, an open-source tool for local development and evaluation of multiple LLMs.

CoLLiE Star★ 400+ (Accepted by EMNLP-demo 2023) GitHub

- Implemented soft prompt techniques for better task adaptation.

OpenRT Star★ 200+ GitHub

- Introduced a standard red teaming framework for quick evaluation.

Internship

Shanghai AI Lab, LLM Distributed Training Engineer (Internship), Supervised by Dr. Hang Yan Mar. 2023 – Aug. 2023

Shanghai AI Lab, LLM Evaluation and Alignment Researcher (Internship), Supervised by Dr. Yan Teng Jan. 2024 – Jun. 2026