Zeguan Xiao

PhD Candidate in Management Science and Engineering | LLM Safety, Alignment, and Robustness

hainanxzg@gmail.com | 13539498124 (Phone / WeChat)

Education

Shanghai University of Finance and Economics — Sept 2022 – Jul 2026
School of Computing and Artificial Intelligence, PhD in Management Science and Engineering — Advisor: Yun Chen

Jinan University — Sept 2018 – Jul 2021
School of Information Science and Technology, MS in Computer Software and Theory

Dongbei University of Finance and Economics — Sept 2014 – Jul 2018
School of Statistics, BS in Economic Statistics
Ranked in the top 3% province-wide in the national college entrance examination, reaching the admission thresholds of most Project 211 universities and some Project 985 universities.

Research Experience

Southern University of Science and Technology — Apr 2023 – Apr 2026
Visiting Student — Advisor: Guanhua Chen
1. Conducted systematic research on LLM safety, alignment, and robustness during the visiting period.
2. Contributed to the writing of an NSFC General Program grant proposal, taking responsibility for part of the research content and technical roadmap design.

Selected Publications During PhD Studies

1. LLM Unlearning

This line of work focuses on two questions: how to preserve general capabilities while unlearning in LLMs, and how to improve the robustness of LLM unlearning. Specifically, I have (1) formulated LLM unlearning as an asymmetric two-task learning problem and designed a gradient-synthesis-based algorithm to improve retention performance; (2) studied direct parameter-update modeling under the LoRA framework and proposed a representation-guided initialization and regularization method to better preserve capabilities; and (3) analyzed internal representations in unlearning, showing that poor robustness stems from over-reliance on dominant representation components, and proposed a more robust unlearning method based on minor representation components.

1. Modeling LLM Unlearning as an Asymmetric Two-Task Learning Problem
   First Author Under review at ACL 2026, average score: 3.8 (Strong Accept)
2. Representation-Guided Parameter-Efficient LLM Unlearning
   First Author Under review at ACL 2026, average score: 3.0 (Findings Accept)
3. Rethinking Robust LLM Unlearning Against Relearning Attacks: The Minor Components in Representations Matter
   First Author Under review at ICML 2026

2. Robustness and Uncertainty

4. Enhancing Uncertainty Estimation in LLMs with Expectation of Aggregated Internal Belief
   First Author Accepted to AAAI 2026 (CCF-A)
5. Automatic Robustness Stress Testing of LLMs as Mathematical Problem Solvers
   Co-first Author Under review at ACL 2026, average score: 3.0 (Findings Accept)

3. LLM Alignment

6. Towards Bridging the Reward-Generation Gap in Direct Alignment Algorithms
   First Author Under review at ACL 2026, average score: 3.0 (Findings Accept)

4. LLM Jailbreak

7. Distract Large Language Models for Automatic Jailbreak Attack
   First Author Accepted to EMNLP 2024 (CCF-B)
8. SeqAR: Jailbreak LLMs with Sequential Auto-Generated Characters
   Co-first Author Accepted to NAACL 2025 (CCF-B)