肖泽管

管理科学与工程博士生 | 大语言模型安全、对齐与鲁棒性

hainanxzg@gmail.com | 13539498124 (电话 / 微信同号)

教育背景

上海财经大学 2022 年 9 月 – 2026 年 7 月

计算机与人工智能学院，管理科学与工程博士 导师：陈云老师

暨南大学 2018 年 9 月 – 2021 年 7 月

信息科学技术学院，计算机软件与理论硕士

东北财经大学 2014 年 9 月 – 2018 年 7 月

统计学院，经济统计学学士

高考排名全省前 3%，达到大部分 211 分数线、少量 985 分数线。

科研经历

南方科技大学 2023 年 4 月 – 2026 年 4 月

访问学生 导师：陈冠华老师

1. 访问期间主要围绕大语言模型安全、对齐与鲁棒性开展系统研究。
2. 参与国家自然科学基金面上项目申请书撰写，负责部分研究内容与技术路线设计。

攻读博士期间代表性论文

一、大语言模型遗忘（2025 年 8 月 - 2026 年 2 月）

1. **Modeling LLM Unlearning as an Asymmetric Two-Task Learning Problem**
   第一作者 ACL 2026 会议（CCF-A）
2. **Representation-Guided Parameter-Efficient LLM Unlearning**
   第一作者 ACL 2026 Findings
3. **Rethinking Robust LLM Unlearning Against Relearning Attacks: The Minor Components in Representations Matter**
   第一作者 ICML 2026 投稿中，均分 3.67

二、鲁棒性与不确定性（2025 年 5 月 - 2025 年 8 月）

4. **Enhancing Uncertainty Estimation in LLMs with Expectation of Aggregated Internal Belief**
   第一作者 接收至 AAAI 2026 会议（CCF-A）
5. **Automatic Robustness Stress Testing of LLMs as Mathematical Problem Solvers**
   共同第一作者 ACL 2026 Findings

三、大语言模型对齐（2025 年 2 月 - 2025 年 5 月）

6. **Towards Bridging the Reward-Generation Gap in Direct Alignment Algorithms**
   第一作者 ACL 2026 Findings

四、大语言模型越狱（2023 年 9 月 - 2024 年 3 月）

7. **Distract Large Language Models for Automatic Jailbreak Attack**
   第一作者 接收至 EMNLP 2024 会议（CCF-B）
8. **SeqAR: Jailbreak LLMs with Sequential Auto-Generated Characters**
   共同第一作者 接收至 NAACL 2025 会议（CCF-B）