个人简历——齐巍巍

15090236598 weiweiqi@zju.edu.cn AutumnleavesQaQ

教育背景

本科：哈尔滨工业大学（校本部） | （计算学部）计算机科学与技术专业 | 学分绩 88.7 / 100

博士(在读)：浙江大学 |（计算机学院）网络空间安全专业

专业课程：线性代数（98）、C++程序设计（97）、算法设计与分析（93）、自然语言处理（92）、人工智能安全（94）、人工智能算法与系统（96）

科研经历

基于马尔可夫链自适应策略组合的LLM越狱研究 *AAAI-2026（CCF-A 一作）*

- 大型语言模型目前易受越狱攻击，而现有的攻击技术依赖静态或单一策略 。我们设计并实现了一个名为 MAJIC 的自适应越狱框架 ， 通过构建一个包含优化策略和创新策略的“伪装策略池” ， 并将策略的顺序选择过程建模为马尔可夫链 ， 最后引入基于Q-learning启发的轻量化实时更新机制 ， 我们在多种安全对齐模型上取得了优异的攻击效果。

大语言模型安全对齐与机制可解释性研究 *ACL-2026（CCF-A 一作在投）*

- 针对大语言模型安全机制不透明的问题 ， 我们设计了安全重要参数识别框架以量化模型中参数对LLM安全能力的影响 。研究揭示了不同模型架构下安全关键参数是稀疏且差异分布的。据此，本文设计了两种轻量化安全微调范式，实现在仅更新极少量参数的情况下 ， 显著增强模型安全性和预防下游任务微调导致的安全能力衰减 。

项目经历

基于昇腾服务器的DeepSeek-R1-Safe模型千卡训练 *项目成员，2025.06*

- 根据千卡集群大规模训练背景，配置了1024卡昇腾服务器的训练环境，完成固件、驱动、CANN toolkit 及 Conda 环境的自动化部署与调试任务。基于 LLM 安全对齐问题，设计安全数据筛选方案，构建专用安全数据集，并参与实施了模型的 SFT 训练过程，以增强模型在安全方面的能力。在 MMLU、GSM8K、CEVAL 及自建安全 Benchmark 环境下，对比基线模型DeepSeek-R1，在安全任务上，安全能力大幅提升，在通用任务上，性能下降均未超过 1%。

竞赛奖项

哈工大校级大创年度项目二等奖 *项目组长，2020.08*

第十二届全国大学生数学竞赛三等奖 *个人竞赛，2020.12*

第十三届全国大学生数学竞赛一等奖 *个人竞赛，2021.11*

第十三届全国大学生数学竞赛（决赛）二等奖 *个人竞赛，2023.03*

第二届全国大学生奥林匹克数学竞赛（夏季赛）优秀奖 *个人竞赛，2023.05*

全国大学生数学建模竞赛黑龙江赛区一等奖 *竞赛组长，2021.11*

2021年美国大学生数学建模竞赛H奖 *竞赛组长，2021.02*

2022年美国大学生数学建模竞赛H奖 *竞赛组长，2022.01*

荣誉奖项

2019.09 优秀学兵

2020.04 社会实践优秀个人项目奖

2020.05 抗疫志愿服务奖

2020.05 优秀团干部（笃实践行奖）

2020.06 2019 年度人民奖学金

2020.09 2020 年度人民奖学金

2020.12 优秀学生干部

2021.05 抗疫志愿服务奖

2021.12 优秀学生（实学实干奖）

2025.10 优秀研究生（学术创新奖）