齐瑞泽

AI 安全 · 大模型内容安全 · 产品研究方向
19547421021　wsQRZ@iCloud.com

教育经历

南京农业大学（211）　农学 · 本科　2022.09 — 2026.06

核心研究项目

大语言模型操作性边界的语义越权与元认知对齐　独立研究 · 2026.3

红队实证研究　|　7 模型 · 4 攻击向量 · 140 轮次测试　|　Python 全流程自动化

- 自主设计「低约束角色扮演」红队框架，以银行客服场景为语义载体，评估 Claude、GPT-5、Gemini、DeepSeek 等 7 款主流模型在越权指令压力下的内容安全合规边界
- 构建 4 类攻击向量（权威身份伪造、情感施压、指令走私、编码混淆），覆盖输入层的多维度安全护栏压力测试
- 设计 5 项观测指标，分别对应输入意图识别与输出内容合规两个维度，基于 140 轮次对抗数据建立可复现的量化评估方法论
- 核心发现：推测「有用性过拟合」可能导致输出合规失效——高风险模型虚假执行确认率与沦陷率高度吻合（DeepSeek 两项均达 35%）；Claude-Sonnet-4.6 全向量保持最高防御稳定性
- 以 AI 辅助编程方式实现完整研究 pipeline：Python 自动化调用多模型 API、自动打分、数据可视化，并将研究成果封装为 HTML 交互式阅读器

实践经历

英特利普（深圳）科技有限公司　猎头助理 · Tech 部门　2025.06 — 2025.09　杭州

- 解读区块链、前端/全栈等技术方向 JD，制定精准人才搜索策略，快速建立对前沿技术岗位核心要求的判断力并进行人选寻访
- 通过 LinkedIn、脉脉、Boss 等渠道寻访候选人，对 React、Node.js、云服务等技术栈进行初步评估与面谈，提高推荐成功率
- 汇总候选人推荐报告，协助客户推进面试流程与录用决策，锻炼结构化信息整合与跨方沟通能力

美团优选　校内 BD　2025.01 — 2025.04

- 0 到 1 搭建 13 个 200 人微信社群（覆盖全校 21%），单群最高转化率 95%，单日拉新峰值 500 人
- 设计「日签单竞赛+阶梯奖励」激励机制；联动菜鸟驿站优化提货流程，实现准时送达

校内外卖平台　创业者　2023.03 — 2023.07

- 100+ 用户问卷提炼核心痛点，首月复购率 60%，单日订单峰值 90 单，服务覆盖 80%宿舍区
- 基于配送数据优化路线，平均送达误差从 25 分钟压缩至 15 分钟，同步降低人力成本

自我评价

非科班背景自驱转型 AI 安全方向，擅长借助 AI 工具快速完成从 0 到 1 的复杂任务；对于新技术和 AI 有强烈好奇心与学习能力，持续跟进 Anthropic、Google 等前沿安全框架及中国 AI 监管动态；具备结构化分析思维，能将研究发现转化为可量化的评估方法论。

技能与工具

LLM 工具：熟练使用 Claude、GPT、DeepSeek 等模型；具备系统性 prompt 工程经验，可独立设计多轮对抗测试集

AI 辅助编程：以 AI 辅助编程方式完成完整研究 pipeline（API 调用 · 自动化测试 · 数据分析 · 可视化 · HTML 前端封装）

AI 安全认知：红队测试方法论 · 深入学习中国 AI 合规监管框架

实习意向

2026 届应届生 · 每周到岗 5 天 · 意向方向：LLM 内容安全 / AI 安全策略产品