

AutoResearch 二期 | 项目证明速览

潘志超 · 计算机科学与技术 · 本科三年级

申请方向：把研究方法整理为可反复运行、可核验的实验任务

01 既有公开工程

LLM Evaluation Playbook：评测标准、数值容差和 Python 检查工具。

AI CLI Orchestrator：受约束的本地代理路由与执行证据。

GitHub：Zhi-Chao-PAN/llm-evaluation-playbook

GitHub：Zhi-Chao-PAN/ai-cli-orchestrator

02 可反复尝试的研究任务

加权 RRF / NFCorpus：参数边界、开发反馈、六次调用预算。

1 条历史轨迹 + 3 条新增真实代理轨迹；每条保留先行假设。

新增三条选择同一配置；公开测试 0.307044，低于预设六点

搜索 0.307294。冻结选择和完整稳定排序核验均随包提供。

03 论文方法复跑与失败分析

RoBERTa / MRPC：三方法 × 两预算 × 三种子，另做秩诊断。

20 次正式训练 + 1 次从头重复；逐条纠正前版分词输入错误。

12 轮 LoRA 的 MRPC 平均 F1 为 0.9074；PAWS 词序压力集

平衡准确率约 0.5009。保留常规指标未能外推的负面结果。

04 复核与范围

python reviewer_demo.py 可快速重算固定样例；完整预测、源码、

输入哈希、选优记录和独立核验均在同交付包。

电池 54 次训练用于机制审查：零违例来自共同后处理。

年级与个人贡献如实说明

目前大三，申请按招募说明的成果与实际能力酌情考虑。

本轮研究材料由 AI 辅助实现与运行；愿接受英文方法解释、

代码定位及限时复跑。Docker 容器运行尚待本机验收。

技术报告不等于论文发表，平台任务细节遵守保密要求。