

张骁

(+86) 17781537802 · SeanZhang03@outlook.com · 个人主页 · GitHub @OHHZZ

教育背景

西南石油大学 (双一流), 网络空间安全, 在读硕士研究生 2024.9 - 2027.6
排名 2/163(前 2%), 学业一等奖学金, 优秀团员/优秀学生/优秀青年志愿者
西南石油大学 (双一流), 网络空间安全, 工学学士 2020.9 - 2024.6
排名 7/53(前 15%), 优秀学生奖学金 (2 次)

技术能力

- 编程与框架: 熟悉 Python/C++, PyTorch, 具备强化学习算法与数据结构基础
- 工程化: 熟悉 Docker、CI/CD、GoogleTest、Git、Linux、MCP Server
- AI 工具: 深度使用 Claude Code、Cursor、Codex、Trae、OpenCode
- Agent / LLM: 熟悉 Agent 设计模式、RAG 系统、Prompt / Harness Engineering、Multi-Agent 协作
- 英语水平: 通过 CET4、CET6, 参与 2025 年成都世界运动会, 与国外团队成员协作超 15 场赛事

项目经历

Personalized Research Intelligence Agent 个性化科研情报推荐引擎 GitHub

技术栈: Python / Agentic RAG / Multi-Agent Workflow / Agent Trace / Embedding / LangGraph
个人主要贡献: 独立完成从需求分析到上线的全链路开发, 包括多 Agent 架构设计、5 个数据源 Connector 封装、LangGraph 状态机编排、LLM 增强模块、RAG 索引与检索

- 核心架构: 将复杂任务拆解后, 构建多 Agent 工作流, 各 Agent 只依赖 Pipeline State 通信, 无直接耦合, 同时用 LangGraph StateGraph 建立 SupervisorNode 在 Filter 后动态决策: 高优先级候选 ≥ 10 条时自动将 LLM 分析上限从 10 提升至 20, 候选类型单一时跳过无关工具, 减少无效调用
- RAG 系统: 实现 BM25(权重 32%) + Embedding(权重 68%) 混合召回的 RAG 检索层与 API 并发检索, 对用户当前选中条目设置 +0.35 boost, 保证上下文关联性, 相比单一 BM25, 将 Recall@50 从 61% 提升至 78%, Precision@10 从 52% 提升至 67%, 低质量内容过滤准确率约为 70%, 吞吐量相对串行检索提升 5 倍, 检验 LLM 回答内容, 每次回答均携带可解释的幻觉风险评估值
- 价值评估机制: 设计非热度驱动的 Research Value Score, 从 relevance、novelty、technical depth、activity、documentation 等多维度评估论文与 Repo; 在 300+ 人工标注验证集上, 相比 only citation / stars 将 Precision@10 从 42% 提升至 61%, NDCG@10 提升 20%, Long-tail Recall@20 提升约 10%

AgentTune Lab 面向 Agent 知识检索与文档问答场景的 LoRA 微调 GitHub

技术栈: Python / PyTorch / Hugging Face Transformers / PEFT / SFT / LoRA
个人主要贡献: 基于 Hugging Face Transformers 与 PEFT LoRA 构建面向 Agent 工具文档问答的轻量微调, 使用 1394 条 Lamini Docs QA 数据完成数据下载、格式清洗、8:2 train/eval 切分、LoRA 训练与评估

- 基座模型调整: 将基座模型从 70M 通用小模型升级为 0.5B 指令模型, 增强 Agent 工具文档问答场景下的指令跟随与文档理解能力; 在评估集上, 升级后模型的 avg token f1 相对提升约 44.0%
- Prompt 优化: 将通用 QA prompt template 改造为面向 Agent 工具/API 文档理解的 instruction prompt, 显式引入 Agent assistant 角色与 documentation question 任务语境; 在 279 条 held-out 样本上, LoRA 微调后 avg token f1 相对提升约 30.1%
- LoRA 超参数调优: 围绕训练步数、LoRA rank、learning rate 等参数设计多组消融实验, 相比 base 将 avg token f1 提升约 46.9%, 并将 exact match 提升 2 个百分点

科研成果

EnerVerse: 面向低功耗 IoT 终端的通用能效优化架构 ASPLOS 2027

- 学生一作, 在投四大体系结构顶会 ASPLOS(CCF-A)
- 基于 DRL 与 LLM 设计面向 MCU 芯片的嵌入式 IoT 通用能效优化架构, 由大模型仿真 Agent 自动转换工程语义为 DRL 输入, 将边缘设备任务执行能耗降低 43.46%–67.67%

竞赛获奖

- 中国国际大学生创新大赛金奖、挑战杯全国大学生创业计划竞赛金奖