邢 篪

https://openchi.life | 1191731324@qq.com | https://github.com/MartinRepo
Aug, 2002 | +86 182-4198-3999 | 爱丁堡, 英国

教育背景

爱丁堡大学 — **爱丁堡，英国**
人工智能硕士 (主攻自然语言处理方向) — *2024.09 – 2025.09*

利物浦大学 & 西交利物浦大学 — **利物浦，英国 & 苏州，中国**
计算机科学学士 (一等荣誉学位) — *2020.09 – 2024.07*

开源经历

ServerlessLLM (Github 400+ Stars) — **https://github.com/ServerlessLLM/ServerlessLLM**
社区成员，核心维护者（贡献+ 审阅） — *2024.11 – 至今*

- 支持 ServerlessLLM 在基于 SLURM 的高性能计算（HPC）集群中的部署，实现任务的高效调度与资源利用。
- 构建了大模型微调方案，支持按需加载多个 LoRA Adapter，提升微调并发能力与资源利用率。
- 深入理解 AI 基础设施工程实践，包括模型服务部署、事件驱动的 GPU 利用策略以及大模型推理系统的系统优化，不局限于传统模型训练任务。
- 在 Prof. Luo Mai 教授的指导下，扩展系统以支持 Serverless 场景下的大模型微调，涵盖参数高效微调（如 LoRA）及全量参数微调（如 RLHF、SFT、DPO）。

Casibase (Gihub 3100+ Stars) — **https://github.com/casibase/casibase**
核心贡献者，担任 2024 开源之夏项目导师 — *2024.01 – 2024.09*

- 扩展了 Casibase 对多种大语言/视觉模型（LLMs/VLMs）的支持，涵盖开源与商用模型，支持聊天与向量嵌入任务。
- 优化文本切分逻辑，提升 RAG 知识库中向量化嵌入的效率与效果。
- 参与全栈开发，后端采用 BeeGo，前端使用 React.js，基于 MVC 设计模式实现模块解耦与高可维护性。
- 提升系统多模态处理能力，优化多模态输出格式，处理多个关键性 Feature 和 Bug。
- 开发即时通讯系统，支持多智能体协同对话功能。
- 累计贡献近 9,000 行代码，涵盖核心功能开发与系统优化，并且有幸担任 2024 开源之夏的项目导师，指导学生参加社区工作。

实习经历

英国 N8 计算密集型研究卓越中心 — **英格兰北部八校联盟**
实习研究员，英国工程和自然科学研究委员会（EPSRC）资助 — *2024.06 – 2024.09*

- 专注于评估大语言模型在生物医学文献阅读任务中的表现，基于 Llama.cpp 对 Llama3.1-70B、Llama3.1-405B、DBRX 和 Mixtral-8x22B 等开源模型进行量化压缩处理。
- 构建客观评分系统，从模型输出中提取关键信息，并与人工提取数据进行相似度比对，做偏好对齐，用于模型性能评测。
- 设计摘要压缩方法，有效缩减输入长度，提升信息密集度，使得小上下文窗口的模型亦可完成长文阅读任务。
- 在多种硬件平台上高并发运行系统，包括 NVIDIA GH200、A100 以及 CPU/GPU 环境，并部署模型于高性能计算（HPC）架构中。

科大讯飞 — **苏州，中国**
研发部门，软件工程师 — *2022.06 – 2022.09*

- 参与“讯飞预见 • 警务超脑系统”中地址类数据的 NLP 标注与质检工作，提升位置相关自然语言任务的准确性。
- 对自动化系统初步标注的地址 POI（兴趣点）数据进行重新标注与精细化修正，显著提升数据质量，解决机器标注带来的低准确率问题。
- 自主开发数据清洗与预处理程序，显著提升标注效率，并获得项目负责人认可。
- 深入探索 NLP 中的实体关系抽取技术，包括 SPO（主谓宾）三元组抽取，并学习从结构化数据构建知识图谱的方法，增强对语义表示与知识建模的理解。

论文发表

Preference Alignment on Diffusion Model: A Comprehensive Survey for Image Generation and Editing — **IJCAI 2025**
https://arxiv.org/abs/2502.07829 — *2025.02*

- 贡献扩散模型偏好对齐的应用部分，并总结了一套应用范式，包括自动驾驶，机器人轨迹规划，医疗成像等领域。

黑客马拉松

2023 宝马黑客马拉松 — **沈阳，中国**
赛道第二名: HVB 再利用在生产中的节能方法 — *2023.08*

- **设计并实现电力调度算法**：基于动态规划思想，分析电价的峰谷规律、光伏发电量与太阳辐射强度等变量，推导出动态转移方程，成功求解电力调度的最优方案。
- **电池调度策略设计**：针对包含大量退役车载电池的储能柜系统，设计以贪心算法为核心的调度方案，有效延缓电池寿命衰减；通过理论验证，证明该策略满足贪心选择与最优子结构性质。
- **工程化部署**：使用 Docker 对调度系统进行工程化打包与部署，完成最终方案提交，并熟练掌握 Docker 基础操作。

技能 & 奖项

语言: 英语（无障碍交流）

奖项:

- 2023 年 ICPC ACM 竞赛（英国&爱尔兰赛区） 荣誉提名奖 （利物浦大学第一名）
- 2021 年数学建模竞赛（亚太赛区）三等奖 (前 25%)

技能:

- 常用: C/C++, Python, Pytorch, Go, JavaScript, Java, React.js, Git, Linux, Shell, Docker
- 没有很熟: Tensorflow, MPI, C#, R, SQL, MATLAB, Next.js, Node.js