

邢 麓

chi.xing2025@gmail.com | github.com/MartinRepo | openchi.life

教育背景

爱丁堡大学

爱丁堡, 英国

人工智能硕士 (在读)

2024/09 – 2025/09

- 专注于多种深度学习架构, 从基础神经网络到各种先进架构 (如 Transformers, Diffusion Models, VAR 等)。
- 毕业论文研究方向: 分布式深度学习训练负载中的高效地理空间转移机制, 该项目由 [Prof Luo Mai](#) 指导。

利物浦大学 & 西交利物浦大学

利物浦, 英国

计算机科学学士 (荣誉一等学位)

2020/09 – 2024/07

- 研究兴趣: 算法设计与优化, C++/C/C#, Java, 计算机系统, 高性能计算, 机器学习, 多智能体协同
- 毕业论文研究方向: 智能电网的调度算法设计与开发。该项目由 [Prof. Prudence Wong](#) 指导。

开源项目

ServerlessLLM

Github 500+ 星

核心贡献者, 代码审阅者

2024/11 – 至今

- 熟悉使用 Huggingface Transformers 和 vLLM 构建大规模分布式推理系统。
- 在 ServerlessLLM 生态系统中, 设计并实现了一套端到端的服务器无感知 PEFT LoRA 微调解决方案, 以提供按需、高性价比的模型定制服务 (#251, #189)。
- 基于 Ray 开发了一套针对 LoRA 适配器的多租户服务器无感知服务解决方案, 通过利用多层检查点加载机制, 加载速度比 safetensors 格式提升高达 4.4 倍 (#248, #221, [博客](#))。

Casibase

Github 3.8k+ 星

核心贡献者, OSPP (开源之夏) 2024 导师

2024/01 – 2024/09

- 增强了平台的核心多模态能力: 实现了各种多模态大模型的深度集成, 支持图像理解、生成及图文混合对话的端到端功能。通过拖拽上传、URL 解析等方式优化了用户体验 (#925, #895, #717, #716)。
- 扩展并优化了大模型支持: 集成多种行业领先模型, 并设计实现了模型提供商多路复用 (Multiplexing) 机制, 允许系统根据负载和成本动态选择模型 (#785, #783, #703)。
- 改进了 RAG 核心 workflow: 通过设计新的文本分割策略, 显著提升了知识库的向量化质量和检索相关性 (#778, #727)。
- 负责全栈开发与性能优化: 使用 Go (BeeGo) 和 React.js 进行全栈开发。独立负责的功能包括: 实时计费与用量统计 (#898, #735)、富文本渲染 (LaTeX, 代码高亮) (#775, #776) 及前端性能优化, 有效提升了消息渲染速度和系统稳定性 (#777, #954)。

实习经历

英格兰 N8 联盟高性能计算研究中心

利物浦 & 约克, 英国

研究实习生@计算生物研究平台

2024/06 – 2024/09

- 专注于对多种大语言模型进行基准测试, 以评估其阅读生物医学文献的能力。利用 Llama.cpp 对 Llama3.1-70B, Llama3.1-405B, DBRX, 和 Mixtral-8x22B 等开源模型进行量化。
- 开发了一套客观评分系统, 该系统从模型输出中提取关键信息, 并与生物医学专家手动提取的数据进行相似度评估, 以实现大模型的性能基准测试。
- 设计了一种摘要压缩方法来总结论文, 以减小输入大小, 从而支持评测上下文窗口较小的模型。
- 该工作还涉及在不同硬件平台 (包括 NVIDIA GH200、A100 集群) 上比较模型和加速器性能, 并在高性能计算 (HPC) 架构上部署大语言模型。

科大讯飞

苏州, 中国

软件开发工程师@核心研发平台

2022/06 – 2022/09

- 在科大讯飞 “警务超脑” 系统中, 通过进行细致的数据标注和质量保证、修正机器标注的地址兴趣点 (POI), 并应用实体关系抽取的基础知识, 提升了基于位置的自然语言处理任务的准确性。

论文发表

[Preference Alignment on Diffusion Model: A Comprehensive Survey for Image Generation and Editing](#)

2025/02

- 贡献了偏好对齐在扩散模型中的应用章节, 包括自动驾驶, 医疗和具身智能等领域, 调研并总结了一套扩散模型应用范式。

技能

- 编程语言 & 工具: C/C++, Python, Go, JavaScript, Java, Git, Linux, Shell
- 深度学习框架: Pytorch, Huggingface-Transformers,, Deepspeed
- 分布式系统/计算: Docker, MPI, Ray