邢 箎

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 – 2025/12

- 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：碳排放感知下的 LLM 预训练工作负载地理空间迁移研究，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，强化学习，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 4k + 星

核心贡献者, OSPP (开源之夏) 2024 导师 2024/01 – 至今

- **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 核心工作流**：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

英国 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，项目海报署名作者 2024/06 – 2024/09

- **主导设计并开发了一套针对生物医学文献的 LLM 基准测试流程**，通过建立客观评分系统，自动化地将模型输出与专家数据进行比对，实现了对模型能力的评估。为前沿大模型（如 Llama 3.1-405B, Mixtral-8x22B）实施模型量化，使用 Llama.cpp 等工具显著降低了模型的显存占用和部署门槛。
- **创新地提出一种摘要压缩方法**，有效解决了长文本超出模型上下文窗口的问题。评估表明，该方法在处理超过 100 篇论文时，不仅降低了计算成本，还普遍提升了模型对文献的理解能力。
- **在多元化 HPC 硬件（GH200，A100，H100 集群）上执行了全面的性能评测**，系统性地比较了不同模型和硬件架构下的推理速度与精度，重构开源评估工具代码 (#3)，使研究社区能够复现和扩展该基准测试，持续评估未来新模型。

科大讯飞 苏州, 中国

软件工程师@核心研发平台 2022/06 – 2022/09

- 在科大讯飞“警务超脑”系统中，通过**细粒度的数据标注**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，提升了基于地理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, Rust, Java, Git, Linux, Shell
- 深度学习框架：Pytorch, Huggingface-Transformers, Deepspeed, Megatron-LM
- 分布式系统/计算：Docker, MPI, Ray