Chi Xing

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

Education

University of Edinburgh Edinburgh, UK
M.Sc Artificial Intelligence 2024/09 – 2025/09

- Focused on various machine learning architectures, ranging from basic neural networks to advanced modern architectures (Transformer, Diffusion Model, Visual Auto-Regressive Model, etc.)
- Dissertation is working on Efficient Geo-spatial Shifting Mechanism for Distributed Deep Learning Training Workloads. This project is supervised by Prof. Luo Mai.

University of Liverpool & Xi'an Jiaotong-Liverpool University Liverpool, UK
B.Sc Computer Science (1st class with honors) 2020/09 – 2024/07

- Research Interest Points: Algorithm Design, C++/C/C#, Machine Learning, Trustworthy AI, Web Development.
- Thesis is focused on scheduling algorithms for modern smart grid. This project is supervised by Prof. Prudence Wong.

Open Source Projects

ServerlessLLM 500+ Stars on Github
Core Contributor, Reviewer 2024/11 – Present

- Developed proficiency in utilizing Huggingface Transformers and vLLM for large-scale distributed inference system.
- Architected an **end-to-end serverless PEFT LoRA fine-tuning solution** to provide on-demand, cost-effective model customization within the ServerlessLLM ecosystem.
- Developed a **multi-tenant serverless serving solution for LoRA adapters** based on Ray, achieving up to 4.4x faster loading speeds than safetensors by leveraging the multi-layer checkpoint loading mechanism.

Casibase 3.8k+ Stars on Github
Core Contributor, Mentor in OSPP (Chinese GSoC) 2024 2024/01 – Present

- **Optimized text splitting logic (RAG workflow)** to improve vectorized embedding for the knowledge base.
- **Expanded Casibase's capabilities** by integrating support for various LMs, including open-source and commercial models for chatting and embedding tasks. **Also enhanced Casibase with multimodality support**.
- Full-stack developement: backend services in **BeeGo** and frontend interfaces in **React.js**, applying the MVC design pattern to ensure loose coupling and maintainable code.
- **Developed an instant messaging system** for multi-agent functionality.

Work Experience

N8 CIR Liverpool & York, UK
Research Intern@Computational Biology Facility 2024/06 – 2024/09

- **Focused on benchmarking various LLMs** for reading biomedical literature, utilizing Llama.cpp to quantize open-source models such as Llama3.1-70B, Llama3.1-405B, DBRX, and Mixtral-8x22B.
- **Developed an objective scoring system** that extracts key information from model outputs and evaluates their similarity to manually extracted data for performance benchmarking.
- **Designed a summarization method** to reduce input size, enabling the use of models with smaller context windows.
- **This work also involved comparing model performance** across different hardware platforms, including NVIDIA GH200, A100, and CPU/GPU references, and deploying LLMs on high-performance computing (HPC) architectures.

IFLYTEK Suzhou, China
SDE@R&D 2022/06 – 2022/09

- Enhanced the accuracy of location-based NLP tasks within the IFLYTEK "Police Super Brain" system by **conducting meticulous data annotation** and quality assurance, **correcting machine-labeled address POIs**, and applying foundational knowledge of **entity relationship extraction**.

Publication

Preference Alignment on Diffusion Model: A Comprehensive Survey for Image Generation and Editing Feb, 2025

- Preference Alignment on DMs Application section, investigated and summarised a set of application paradigms.