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.)
- Thesis is focused on Carbon-Aware Geospatial Shifting of LLM Training Workloads, which 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+ GitHub Stars
Core Contributor, Code Reviewer Nov 2024 – Present

- Proficient in building large-scale distributed inference systems using Hugging Face Transformers and vLLM.
- Designed and implemented an **end-to-end serverless PEFT LoRA fine-tuning solution** within the ServerlessLLM ecosystem to provide on-demand, cost-effective model customization services (#251, #189).
- Developed a **multi-tenant serverless serving solution for LoRA adapters** using Ray, achieving up to a 4.4x faster loading speed compared to the safetensors format by leveraging a multi-tier checkpoint loading mechanism (#248, #221).

Casibase 4k+ GitHub Stars
Core Contributor, OSPP (Open Source Promotion Plan) 2024 Mentor Jan 2024 – Present

- **Enhanced the platform's core multi-modal capabilities**: Deeply integrated various large multi-modal models to enable end-to-end functionalities for image understanding, generation, and mixed-media dialogue. Optimized user experience with features like drag-and-drop uploads and URL parsing (#925, #895, #717, #716).
- **Expanded and optimized LM support**: Integrated multiple industry-leading models and engineered a model provider multiplexing mechanism, allowing the system to dynamically select models based on load and cost (#785, #783, #703).
- **Improved the core RAG workflow**: Significantly boosted the quality of knowledge base vectorization and retrieval relevance by designing novel text-splitting strategies (#778, #727).
- **Led full-stack development and performance optimization**: Utilized Go (BeeGo) and React.js to independently deliver features including real-time billing & usage statistics (#898, #735), rich text rendering (LaTeX, code highlighting) (#775, #776), and front-end optimizations that enhanced message rendering speed and system stability (#777, #954).

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.

Hackathon

2023 BMW Hackathon — Shenyang, China

2nd Place, HVB Reuse for Energy Saving in Production Channel — 2023.08

- **Designed a power scheduling algorithm**: formulated a dynamic programming model based on electricity price fluctuations, photovoltaic generation, and solar radiation intensity; derived the dynamic transition equations and successfully solved for the optimal scheduling strategy.
- **Developed a battery dispatch strategy**: proposed a scheduling method based on greedy algorithms for energy storage cabinets utilizing retired automotive batteries, effectively mitigating battery degradation; theoretically validated the strategy by proving its greedy choice and optimal substructure properties.
- **Engineered and deployed the system**: containerized the scheduling solution using Docker, completed final system submission, and gained proficiency in basic Docker operations.

Skills

- Programming Language & Tools: C/C++, Python, Go, JavaScript, Rust, Java, Git, Linux, Shell
- Deeplearning Framework：Pytorch, Huggingface-Transformers, Deepspeed, Megatron-LM
- Distributed System/Computing：Docker, MPI, Ray