Chi Xing

https://openchi.life martinchi7788@gmail.com https://github.com/MartinRepo Chi Xing

EDUCATION

University of Edinburgh **Edinburgh, UK**

MSc Artificial Intelligence *2024.09 – 2025.09*

- Focused on diverse machine learning frameworks, from basic neural networks to advanced architectures like Transformers and Diffusion Models, with research in reinforcement learning and keen interest in speeding up large model inference.
- Dissertation is working on accelerating and serverless-supported preference alignment techniques (such as LoRA fine-tuning, RLHF, DPO and SFT, etc.). This project is supervised by Prof. Luo Mai.

University of Liverpool **Liverpool, UK**

BSc Computer Science (first class with honors) *2020.09 – 2024.07*

- Interest Points: Algorithm Design, C++/C/C#, Optimisation, Machine Learning, AI Safety, Java, Web Development.
- Dissertation is focused on exploring various scheduling algorithms for modern smart grid. This project is supervised by Prof. Prudence Wong.

OPEN-SOURCE WORK

ServerlessLLM [500 Stars] **https://github.com/ServerlessLLM/ServerlessLLM**

Core Contributor, Reviewer *2024.11 – Present*

- Support ServerlessLLM deployment on SLURM-based HPC
- Contributed to a **scalable, cost-efficient LLM (LoRA) fine-tuning and serving solution**.
- Gained **real-world exposure to AI infra engineering** beyond just model training.

Casibase [3800 Stars] **https://github.com/casibase/casibase**

Core Contributor, 2024 OSPP Mentor *2024.01 – 2024.09*

- **Expanded Casibase's capabilities** by integrating support for various LLMs, including open-source and commercial models for chat and embedding tasks.
- **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.
- **Enhanced Casibase with multimodal support**, optimized output formatting, and bug fixes.
- **Optimized text splitting logic** to improve vectorized embedding for the RAG knowledge base.
- **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.
- **Designed a summarization method** to reduce input size, enabling the use of models with smaller context windows.
- **Developed an objective scoring system** that extracts key information from model outputs and evaluates their similarity to manually curation data for performance benchmarking.
- **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 Department *2022.06 – 2022.09*

- **Contributed to NLP data annotation and quality assurance** for address data in the "IFLYTEK Foresight" Police Super Brain System, focusing on improving the accuracy of location-based NLP tasks.
- **Re-labeled and refined address POI data (Point of Interest)** previously annotated by automated systems, enhancing data quality and addressing low-accuracy outputs generated by machine-based labeling.
- **Explored entity relationship extraction techniques in NLP**, including Subject-Predicate-Object (SPO) extraction, and learned methods for building knowledge graphs from structured data, enhancing understanding of semantic representation in NLP.

PUBLICATION

Preference Alignment on Diffusion Models: A Comprehensive Survey on Image Generation and Editing

- Contributed to the **Preference Alignment on DMs Applications** section, covering Medical Imaging, Autonomous Driving, Robotics, etc. Chi investigated and summarised a set of application paradigms.

SKILLS

- Various Programming Languages, advanced for me: **Python, C++, Go, C, JavaScript, Java**
- Deep Learning Frameworks: **Pytorch and Huggingface**
- Others: **vLLM**, **React.js, Node.js, AWS, Docker, SQL, Ray, Rust**