Chi Xing

Github: Chi Xing | Website: OpenChi.Life | Scholar: Chi Xing | Email: chi.xing2002@outlook.com

Education

University of Edinburgh Edinburgh, UK

M.Sc Artificial Intelligence (Outstanding Thesis) 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 & Suzhou, China

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.

Selected Open Source Projects

ServerlessLLM (OSDI 2024) 500+ GitHub Stars

Core Contributor, Code Reviewer 2024/11 – 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). This makes model developers only focus on their model architecture and no longer concern themselves with system-level issues.
- Built a **multi-tenant LoRA-as-a-Service solution** enabling shared base-model instances and multi-tier NVMe SSD/ DRAM caching, substantially improving model-loading efficiency and GPU utilization.(#248, #221).

Casibase (Casbin Open-Source Community) 4k+ GitHub Stars

Core Contributor, OSPP (Open Source Promotion Plan) 2024 Mentor 2024/01 – 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).
- **Improved the core RAG workflow** by designing efficient text-splitting strategies, integrating lightweight knowledge-graph analysis, and implementing a hierarchical vector-retrieval pipeline that combines LLM-based semantic pruning with ANN similarity search, yielding significantly higher retrieval precision.(#1539, #778, #727).
- **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).
- **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).

Publication

Preference Alignment on Diffusion Model: A Comprehensive Survey for Image Generation and Editing 2025/02

Computer Science Review (Journal)

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

Work Experience

N8 CIR Liverpool & York, UK

Research Intern@Computational Biology Facility, AIBIO UK 2025 Poster 2024/06 – 2024/09

- **Built an automated evaluation pipeline** that extracted entities from model outputs and compared them with manually curated biomedical datasets (VEuPathDB), producing an objective scoring system for precision benchmarking.
- **Designed and implemented a summarization workflow** that reduced input size by 90%, improving inference speed (600 -> 300 min) and accuracy (0.42 -> 0.45 score) under constrained context windows.
- Benchmarked and optimized LLMs (e.g., Llama 3.1-70B/405B, DBRX, Mixtral-8x22B) for biomedical information extraction by quantizing models via Llama.cpp (4- to 8-bit) and deploying large-scale inference on HPC clusters (NVIDIA GH200, A100) using MPI-based multi-GPU parallelization to analyze hardware-specific efficiency trade-offs.

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**.

Academic Experience

ChaseGreen: Carbon-Aware Geospatial Shifting of LLM Training Workloads 2025/06 - 2025/09
Co-First Author (ICML2026 Under Review)

- **Design and built an efficient migration abstract API (cross multiple regions on AWS)**: a memory-to-memory, geo-distributed, multi-stream TCP migration API that attains >**99%** bandwidth utilization, reduced migration cost/time by **16.4%** vs. SOTA approaches.
- **Proposed a long-horizon, carbon-aware planner** based on Rolling Horizon Dynamic Programming that cuts total emissions by up to **59.2%** vs. heuristic baseline.
- **Evalution on methodology on real pre-training workloads**, which got exposure on Megatron-LM and DeepSpeed

RLCookbook 2025/04

- Implemented Dynamic Programming algorithms (Value Iteration and Policy Iteration) to solve Markov Decision Processes and verify policy convergence. **Built RL agents**, including ε-greedy exploration, Q-Learning, and on-policy every-visit Monte Carlo, trained on FrozenLake-8x8-v1 for performance comparison under stochastic transitions.
- **Designed and optimized Deep Q-Network agents** with replay buffers, target networks, and ε-scheduling strategies (linear / exponential decay), benchmarking against discrete tabular baselines on MountainCar-v0 and CartPole-v1.
- **Extended the framework to continuous control** by implementing Deep Deterministic Policy Gradient (DDPG) with actor-critic networks, Gaussian exploration noise, and soft target updates (τ-updates), achieving high-return performance in Racetrack (HighwayEnv).

MuralXGAN: Text-Guided Dunhuang Mural Image Inpainting Framework 2025/02 - 2025/04

- **Developed a text-guided restoration pipeline** using GPT-4o captions and a fine-tuned CLIP encoder to steer a UNet with cross-modal attention and an SN-PatchGAN discriminator.
- **Built a reproducible MuralDH training/eval stack;** generated irregular Perlin masks (10–60%) and random mask permutation to create pseudo ground truth with 1.49% overlap to real damage, improving generalization.
- Tech: PyTorch, CLIP fine-tuning, Cross-Attention, SN-PatchGAN, Perceptual Loss, PSNR/SSIM.

Predicting CT Slice Locations 2024/11
Machine Learning & Pattern Recognition

- Built a machine-learning system to **predict anatomical slice positions from CT features**, implementing linear/logistic regression and neural networks from first principles in NumPy and SciPy (minimal external ML libraries).
- **Designed a Gaussian-process Bayesian optimization loop** for NN hyperparameters and a comprehensive evaluation framework (validation/test splits, learning curves, error tracking).

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.

Skills

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