Xinwei Niu

+44 (0)7947 145405 nxw2002@outlook.com linkedin.com/in/xinwei-niu github.com/xinwei-niu
xinwei-niu.vercel.app

Education

The University of Edinburgh **September 2024 – November 2025**
Master of Science in Artificial Intelligence **(with Distinction)** *Edinburgh, United Kingdom*

The University of Manchester **September 2021 – July 2024**
Bachelor of Science in Artificial Intelligence **(First Class Honours)** *Manchester, United Kingdom*

Experience

Summer Internship Program - Samsung Research China - Beijing **July 2023 - September 2023**

- NLP intern in language understanding lab, worked on MPMQA: Multimodal Question Answering on Product Manuals.
- Developed for a multimodal model based on transformer, experienced with data processing and model tuning processes.
- Facilitated in the development and evaluation phase, and utilised large multimodal datasets and deep learning frameworks Pytorch and HuggingFace.

Projects

Multimodal QA System with Mixture of Adapters for Scientific Documents
PyTorch, Huggingface Transformers, Huggingface PEFT **May 2025 – August 2025**

- Master's dissertation supervised by Dr. Emily Allaway.
- Designed Mixture of Adapters (MoA) methods to enhance zero- and few-shot generalization of multimodal LLMs.
- Implemented task-wise routing (jointly pretrained 5 LoRAs) and token-wise routing (task-specific LoRAs) for improved cross-task performance.
- Both methods outperformed closed- and open-weight baselines on seen and unseen scientific document tasks, with an improvement for seen tasks of by 13% on ROUGE and 1.2 on CIDEr, and improvement for unseen tasks by 14% on ROUGE and 1.4 on CIDEr.

4D Gaussian Splatting with Data-driven Depth Prior
PyTorch **February 2025 – June 2025**

- Group project for Machine Learning Practical course (team of 3).
- Implemented real-time 3D scene reconstruction using Gaussian point clouds, guided by data-driven depth and optical flow priors, and optimized with additional loss terms for depth and flow consistency.
- Optimized model performance for speed and accuracy, enabling efficient visualization and high-fidelity scene reconstruction of dynamic scenes.

Event Sequence Prediction with Linear-Time Sequencing State Space Model
PyTorch **September 2023 – June 2024**

- Bachelor final year project supervised by Dr. Mauricio Álvarez.
- Implemented Mamba-based Temporal Point Process (TPP) model using discretized state space models to predict next event type and timing.
- Compared with Transformer- and LSTM-based TPPs, achieving lower NLL and MSE for time prediction, and higher accuracy for event type prediction across standard datasets.
- Achieved faster training and inference due to Mamba architecture's efficiency.

Extracurricular

Independent Study on Efficient Methods for Generative Models **2023 – Present**

- Self-driven research on efficient generative models and optimization techniques.
- Studied advanced topics of efficient diffusion and autoregressive models, including efficient samplers, linear attention, linear RNNs (such as state-space models), and adaptive memory compression and eviction strategies.
- Conducted a literature review for the Informatics Research Review course about Efficient text-conditioned diffusion Models for image synthesis.
- Implemented fused kernels on experimental models using Triton to explore efficiency improvements and performance optimization.

AdaHack 2024 **October 2024**
Edinburgh, United Kingdom

- Participated in AdaHack Hackathon as the team lead.
- Implemented a market forecast system in a team of 4; applied PyTorch, data preprocessing, and rapid prototyping to deliver a working solution in 12 hours.

UKIEPC 2023 **October 2023**

Manchester, United Kingdom

- UK and Ireland Programming Contest.
- Attended as a group of three in the contest, received an Honorable Mention.

Private Coding Tutor **June – August 2022, June – August 2024**

Online

- Teaching a non-UK student currently study in Biology & Life Science Python.
- Covered Python intros, OOP principles, and standard Python libraries such as NumPy.
- Emphasized question-solving to enhance problem-solving skills.

Technical Skills

Programming Languages: Python (with beginner to intermediate level of Triton), Java, C, C#, PHP, SQL
Deep Learning Frameworks: Pytorch, Keras, PaddlePaddle
Machine Learning Libraries: Numpy, Scipy, Pandas, scikit-learn
Deep Learning Libraries: Huggingface Transformers, PEFT, Diffusers, TRL
Distributed Training: Accelerate, DeepSpeed
NLP Libraries: NLTK, spaCy
Languages: English (Proficient), Mandarin (First Language), Japanese (Beginner)