MATTHEW ZHANG

✉ matthew.zhang473@gmail.com   github.com/MatthewZhang473   linkedin.com/in/matthew-zhang-cambridge

Education

University of Cambridge Cambridge, UK

BA & MEng Engineering (Distinction & First Class Honours) *October 2021 – June 2025*

- Specialized in Information and Computer Engineering with a research focus on Bayesian ML.
- Consistently ranked in the top 5% of the cohort; awarded the Clough Scholarship annually for academic excellence.
- ***Key Modules:*** Deep Learning & Structured Data; Probabilistic Machine Learning; Computational Statistics & Machine Learning; Computer Systems & Concurrency.

Publication

Graph Random Features for Scalable Gaussian Processes

Accepted at ICLR 2026.

Matthew Zhang, Jihao Andreas Lin, Adrian Weller, Rich Turner, Isaac Reid

TL;DR: We enable graph Gaussian process inference in $\mathcal{O}(N^{3/2})$ time using graph random features (GRFs), making Bayesian optimization feasible on million-node graphs on a single computer chip.

https://arxiv.org/abs/2509.03691

Work Experience

Myrtle.ai Cambridge, UK

ML Engineer *November 2025 – Present*

- Pretrained and optimized Automatic Speech Recognition (ASR) models for low-latency, real-time streaming inference under tight compute constraints.
- Worked on production RNN-T ASR pipelines (caiman-asr).
- ***Tech Stack***: ASR, RNN-T, ML infra, pretraining, inference.

Machine Learning Group, University of Cambridge Cambridge, UK

Researcher *October 2024 – October 2025*

- Designed a scalable inference algorithm for GPs on graphs, reducing complexity from $O(N^3)$ to $O(N)$ and enabling inference on networks with 10M+ nodes.
- Integrated sparse linear algebra and randomized feature maps into the GP framework, achieving over 100× speedups versus standard implementations.
- Drafted a first-author manuscript for submission to ICLR.
- ***Research Areas***: Bayesian inference, graph machine learning, Monte Carlo methods.
- ***Tech Stack***: Python, PyTorch, CUDA, GPytorch, GPflow, NetworkX, SciPy, NumPy, MATLAB.

Microsoft Cheltenham, UK

Software Engineering Intern *July – September 2024*

- Researched and deployed a scalable graph ML algorithm for threat detection, processing data from 30M+ Azure cloud apps daily.

- Developed distributed ML pipelines using Azure ML to accelerate large-scale data processing.
- Implemented new threat indicator services in .NET/C#, with comprehensive unit tests to ensure robustness and reduce integration errors.
- ***Tech Stack***: C#, .NET, KQL, PySpark, PyTorch, Pandas, GraphSAGE.

Playfair Technologies Remote

Machine Learning Engineer *Nov 2022 – May 2023*

- Built end-to-end LLM/NLP pipelines for financial sentiment and industry classification; fine-tuned Hugging Face Transformers with long-context segmentation, improving accuracy by 20%+.
- Designed a multi-stage classification pipeline (segmentation → entity/ticker extraction → zero-shot industry labeling) delivering robust outputs with ~50% lower inference cost.
- Expanded datasets by 30× through augmentation (selective search, image transforms, noise injection) to enhance model robustness and generalization.
- ***Tech Stack***: Transformer-based LLMs (GPT-3, DistilRoBERTa), CNN/RNN, OpenCV, PyTorch.

Huawei Cambridge, UK

CPU Architecture / Software Engineering Intern *June – September 2022*

- Designed and implemented an auto-generation system for machine-readable ISA specifications by creating a parser to extract code snippets for a novel ISA semantics language, Sail.
- Deployed and managed a Kubernetes cluster to run containerized applications and streamline CI/CD pipelines.
- ***Tech Stack***: Python, Kubernetes, Docker, Jenkins, CI/CD, ISA, Linux, Git.

Selected Projects & Hackathons

LLM Fine-Tuning & Inference Optimization

- Fine-tuned **Qwen2.5-Instruct (0.5 B)** via LoRA adapters, systematically sweeping adapter rank, learning rate, and context length to maximize time-series forecasting performance under constrained compute budgets.
- Designed and executed post-training quantization + inference benchmarking, including 4-bit/8-bit quantization, measuring trade-offs in throughput, memory footprint, and accuracy to guide deployment decisions.
- ***Tech Stack***: PyTorch, Hugging Face Transformers, PEFT, TRL, CUDA, ONNX / vLLM

Cambridge GenAI Hackathon

- Developed a scalable SaaS platform leveraging multi-agent LLM frameworks; awarded 1st place out of 50+ teams.
- ***Tech Stack***: Multi-agent LLMs, AutoGen, HTML, JavaScript, CSS.

RISC-V Processor Optimization

- Optimized an RV32I processor on FPGA, achieved a 13.3% increase in maximum clock frequency.
- ***Tech Stack***: C/C++, FPGA, RISC-V Architecture, SIMD, Branch Prediction.

Hack Cambridge 2023

- Created a prototype to accelerate manga production by fine-tuning Stable Diffusion with the Manga101 dataset; placed 3rd out of 80+ teams.
- ***Tech Stack***: Stable Diffusion, ControlNet.