Shi Jie Yu

yshijie1999@gmail.com | LinkedIn | GitHub | Website

Education

New York University (NYU) — **Sep 2024 – May 2026 (Expected)**
Master's in Computer Science **(GPA: 4.0/4)** — *New York, NY*

National University of Singapore (NUS) — **Aug 2020 – May 2024**
Bachelor of Science in Business Analytics with Honors (Distinction) — *Singapore*

Technical Skills

Languages: Python, SQL, C, C++, JavaScript, Typescript, Solidity, Java
Full Stack Development: React, Node.js, PostgreSQL, Vertex AI, Amazon Web Services, Kubernetes
Machine Learning: PyTorch, Transformers, Llama 3, DeepSpeed, Scikit-Learn, Tensorflow, LitGPT, LLaMA-Factory
Certifications: Google Cloud Platform Certified Professional Machine Learning Engineer

Experience

Founding AI Engineer @ Tensorplex Labs, Singapore — **Apr 2024 – Present**

- Spearheaded continued pre-training of **Llama 3 70B** language models on **billions** of tokens' worth of web-scraped corpora using state-of-the-art techniques like **Fully Sharded Data Parallel (FSDP)** and **task arithmetic**.
- Conducted post-training **(SFT and RLHF)** on frontend interface datasets using **Qwen2.5 Coder** language models, achieving **2x performance gains on interface generation tasks** and robustness against general benchmarks.
- Evaluated language models on general benchmarks using the lm-evaluation-harness framework. Contributed to the project by implementing **MMLU-Pro** and **GSM-Plus**, two influential benchmarks with **18k+** combined downloads.
- Led research and implementation of a novel and efficient way to perform model merging on LoRA modules; **authored a research paper** on the proposed method, Parameter-Efficient Checkpoint Merging via Metrics-Weighted Averaging.

Quantitative Research Consultant @ WorldQuant BRAIN, Singapore — **Nov 2023 – Apr 2024**

- Placed **12th out of 500+** participants in WorldQuant's NUS Alphathon 2023; thereafter, onboarded as a quantitative research consultant to develop **over 30 alphas (predictive signals)** for the US market via market-neutral strategies.

Data Analyst Intern @ Autodesk, Singapore — **May 2022 – Oct 2022**

- Designed interactive Looker dashboards for **3 Autodesk product lines** (ReCap Desktop, ReCap Viewer, ReCap Pro), yielding actionable business intelligence capable of informing **strategic decisions at the director level**.
- Engineered, scheduled, and monitored **20+** daily and weekly Big Data extract, transform, and load workflows on **terabyte-scale** data with SQL, Jenkins, and Apache Airflow for analytics with Qubole and monitoring with Mixpanel.
- Initiated a machine learning project to predict EC2 peak memory usage using ensemble methods (Random Forests, XGBoosts, LightGBM) and deep neural networks, **reducing the cloud cost of AWS EC2 deployments by 50%**.

Projects & Extracurriculars

Research Assistant / Teaching Assistant @ NYU — **Spring 2025 - Present**

- Ideated, designed, and implemented QBERT, a novel BERT Architecture that leverages Quaternion modules in Attention and Feedforward layers, achieving a **75% reduction** in parameters while improving performance on benchmarks.
- Managed a **graduate-level operating systems class of 30** by facilitating weekly consultation sessions, designing grading rubrics and model solutions, and grading and delivering detailed feedback on assignments.

Open-source Partner @ Lightning-AI/litgpt, Lightning AI — **Fall 2024 – Present**

- **Top contributor** to an LLM training and deployment project with **12k+ GitHub stars**; responsible for **80% of new features** since November 2024.
- Implemented state-of-the-art LLMs, including Phi-4 and Qwen2.5 series models; added general enhancements like introducing new architectures and reducing overhead in pretraining and other compute intensive operations.

Full Stack Developer @ Rugged — **Spring 2025**

- Designed and implemented a **REST API** for querying live blockchain data from Gecko Terminal, and seamlessly integrated its endpoints into a Next.js frontend; leveraged ElizaOS to build an LLM agent for real-time crypto insights.

Director @ Fintech Society Machine Learning Department, NUS — **Fall 2023 – Spring 2024**

- Managed a **60-member machine learning club**, creating learning and networking opportunities for members via workshops and hands-on projects; managed **over 10 ML projects** in diverse domains, including NLP, LLMs, and XAI.

Co-founder and Head of Engineering @ Surf — **Fall 2022 – Spring 2024**

- Spearheaded development of the first LLM-powered smart contract IDE and auditing platform; **incubated by NUS Venture Initiation Program with 10,000 SGD funding**.