Shi Jie Yu

sy4468@nyu.edu | linkedin.com/in/sjyuxyz | github.com/ysjprojects | sjyu.ai

Education

New York University (NYU) — **Sep 2024 – May 2026 (Expected)**
Master's in Computer Science — *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, Scala, Go, SQL, JavaScript, Typescript, C, C++, Solidity, Java
Full Stack Development: React, React Native, Node.js, PostgreSQL, Bazel, AWS (EC2, SQS), Kubernetes, Kafka
Machine Learning: PyTorch, vLLM, Transformers, TRL, DeepSpeed, Verl, Unsloth, LitGPT, Nemotron
Certifications: Google Cloud Platform Certified Professional Machine Learning Engineer

Experience

Software Engineer Intern @ Gemini, New York, NY — **Jan 2026 – Present**

- Architected an end-to-end, event-driven email pipeline using **Scala/Play**, PostgreSQL, and SQS to generate and deliver highly personalized monthly reward recaps for Gemini Credit Card holders at scale **(500k MAU)** via SendGrid.
- Built a real-time shipment tracking experience across **web (React) and mobile (React Native)**; integrated **Shippo webhooks** into an asynchronous Scala backend to automate notifications across 9 distinct shipment states.
- Developed a high-throughput **Kafka** producer/consumer pipeline in **Go (franz-go)** within a Kubernetes-deployed microservice to process live sports events and serve real-time predictions with minimal latency.

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

- Engineered custom frameworks based on **Verl and NVIDIA-NeMo/RL** for agentic reinforcement learning to train LLMs with **GRPO** and other state-of-the-art RLVR techniques in multi-turn environments (Alfworld, Webshop, etc.).
- Conducted post-training on interface generation tasks using **Qwen2.5 Coder** LLMs and synthetic data curated via distributed human feedback, achieving **2x performance gains on interface generation tasks**.
- Evaluated LLMs on general benchmarks using the lm-evaluation-harness framework. Contributed to the open-source 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.
- 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**.

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 – Fall 2025**

- 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 inference project with **13k+ GitHub stars**; created **30+ pull requests**.
- Implemented state-of-the-art LLMs, including Phi-4 and Qwen3 series models; added general enhancements like introducing new architectures and reducing overhead in LoRA fine-tuning and other compute intensive operations.

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

- Led a **60-member machine learning interest group**, creating learning and networking opportunities for members via workshops and hackathons; managed **over 10 projects** in diverse domains spanning classical ML and LLMs.

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**.