Jimmy Shong

650-681-7291 | jimmysh341@gmail.com | linkedin.com/in/jimmy-shong | github.com/Jiminator | Portfolio Website

Education

University of Illinois Urbana-Champaign — Champaign, IL
Masters of Science in Computer Science — *Aug. 2024 – May 2026*

New York University — New York, NY
Bachelor of Science in Computer Science, Minor in Mathematics and Cybersecurity — *Aug. 2020 – May 2024*

Research Experience

Student Researcher — Aug. 2024 – Present
University of Illinois Urbana-Champaign — *Champaign, IL*

- Leading a research project on reducing the search overhead of auto-parallelism systems used in distributed training
- Researching methods on improving the communication efficiency and fault-tolerance of LLM pre-training systems

Research Intern — May 2023 – May 2024
Massachusetts Institute of Technology — *Cambridge, MA*

- Implemented matrix multiplication optimizations for TinyChatEngine, a high-performance LLM inference library
- Built TinyVoiceChat, a tool that allows users to interface with a quantized LLM with their voice on edge devices
- Researched CoreML and Metal backends to improve TinyChatEngine's performance on Apple Silicon hardware

Unsupervised ML Research Lead — Sep. 2021 – May 2022
New York University — *New York, NY*

- Engineered a Python library that analyzed macrophage trajectories and predicted their diffusion states
- Presented our work to researchers at the University of Groningen who utilized our tool in their experiment
- Performed data visualization and analysis on the sleep-wake dynamics of mice for scientists at NYU Abu Dhabi

Projects

Improved Model Parallelism for Distributed LLM Training | *Python* — Aug. 2024 – Dec 2024

- Identified inefficiencies in Metis, a state-of-the-art auto-parallelizer for heterogeneous clusters
- Implemented a novel search strategy, achieving up to 2× speedup without compromising model accuracy
- Authored a draft research paper detailing the inefficiencies, proposed solution, and experimental results

Exploring SFT Methods for LLMs | *Pytorch, SLURM* — Apr. 2024 – May 2024

- Instruct-tuned Llama3-8B using parameter-efficient finetuning (PEFT) methods with Llamafactory
- Evaluated the training speed, memory usage, and model performance of these SFT methods

Google Suite Task Manager | *Flask, MongoDB, Ruff, mypy, Python PDM, CircleCI* — Jan. 2024 – May 2024

- Developed a task manager that aggregates all tasks created in both Google Tasks and other Google Suite products
- Implemented fundamental task management features currently available in Google Tasks

Efficient ResNet | *PyTorch* — Feb. 2024 – Mar. 2024

- Created a 4.6M modified ResNet model that achieves a 3% increase over the original 11.4M ResNet-18 model
- Performed ablation studies on multiple neural network optimizations including dropout, schedulers, and optimizers

NetArmor | *Python, TurboGears, FastAPI, SQLAlchemy* — Sep. 2023 – Dec. 2023

- Developed a tool for website owners to scan their sites for vulnerabilities and contact cybersecurity professionals
- Engineered a PostgreSQL database using SQLAlchemy to store all necessary data

Air Ticket Reservation System | *Flask, HTML, MySQL* — Oct. 2022 – Dec. 2022

- Developed a web-based application that allows airline staff to operate an airport flight transaction system
- Built a relational database built with to store necessary data and handle all transactions

Teaching Experience

Teaching Assistant Aug. 2024 – Present
University of Illinois Urbana-Champaign *Champaign, IL*

- Gave a lecture on transformers and FlashAttention in front of 90+ students taking CS 598: Systems for GenAI
- Hosted lab sections and office hours for STAT-107: Data Science Discovery to guide students on data science topics

Computer Science Tutor Jan 2023 – May 2024
NYU Polytechnic Tutoring Center *New York, NY*

- Answered questions from NYU CS students taking the required computer science courses
- Recorded explanation videos for answer keys of mock exams developed by the PTC

Instructor June 2022 – Jul 2022
BlueStamp Engineering *Palo Alto, CA*

- Taught practices and principles of engineering to a class of twenty high school students
- Led projects such as an intelligent door lock, a smart mirror, and Alexa home automation

Technical Skills

Languages: Python, C/C++, Java, SQL (MySQL), JavaScript, HTML/CSS
Frameworks: PyTorch, TensorFlow, CUDA, CoreML, Metal, MongoDB, FastAPI, Flask, TurboGears
Developer Tools: Git, Docker, CircleCI, Ruff, mypy, Python PDM, AWS, SLURM, Linux (Ubuntu, Kali), MacOS
Libraries: NumPy, pandas, Hugging Face, Matplotlib, SQLAlchemy