Zhanqiu (Jack) Guo

LinkedIn: linkedin.com/in/zhanqiu-guo-b932b624b | Phone: 917-361-4116 | Email: zhanqiu.guo321@gmail.com | Homepage: cs.cmu.edu/~zhanqiug

EDUCATION

Carnegie Mellon University, Pittsburgh, PA
M.S. in Machine Learning — Aug. 2025 – Dec. 2026 (expected)

New York University, New York, NY
B.S. in Computer Science, Mathematics — Aug. 2021 – May 2025

PROFESSIONAL EXPERIENCE

TikTok, *Machine Learning Engineer Intern* — San Jose, CA, May 2026 – Aug. 2026
- Contributed to TikTok Shop Ads Ranking's **Large Recommendation Model**, integrating MixFormer-based feature interaction to jointly model sequential, non-sequential, and multimodal signals for e-commerce ads ranking.
- Developed components for **Semantic ID-based generative recommendation**, exploring multi-level item representations and attention-based integration strategies for large-scale recommendation models.

Depository Trust & Clearing Corporation, *Technology Research & Innovation Intern* — Jersey City, NJ, Jun. 2025 – Aug. 2025
- Engineered an **asynchronous web scraping** and **batch processing pipeline** for newsletter generation, reducing latency and costs by 40×.
- Built a synthetic data generator with Snowflake backend for prompt- and schema-based generation and developed Treasury yield forecasting models with **XGBoost quantile regression**.

Teragonia, *AI/ML Engineer Intern* — New York, NY, Feb. 2025 – May 2025
- Engineered a **document intelligence** pipeline for legal and financial PDFs using layout-aware **OCR**, **recursive retrieval**, automated answer verification, and LLM-to-Excel integration.
- Generalized the system across diverse document types, outperforming 10+ vendor solutions based on client feedback.

Coleman Research, *Machine Learning & Generative AI Intern* — New York, NY, Jun. 2024 – Aug. 2024
- Engineered **citation linking** for an LLM-based expert profile summarizer to improve provenance and traceability.
- Built a **hybrid RAG** system with **Elasticsearch**, improving generation speed by 30% while reducing hallucinations.

PROJECT EXPERIENCE

GlassTorch: Interpretable and Efficient Training in a Custom DL System, *Lead Developer* — Nov. 2025
- Built a PyTorch-like deep learning framework with tensor ops, **autograd**, custom **CUDA** kernels, and module-level runtime profiling.
- Reduced training and inference cost through **dynamic freezing** and **optimizer-aware pruning** with progressive sparsity schedules.

A Task Scheduling and Monitoring System, *Lead Developer* — Nov. 2024
- Developed a **Spring Boot**, **JDBC**, and **React**-based task scheduling system for ARM microcontrollers, with optimized, thread-aware CPU utilization and a real-time monitoring dashboard.
- Engineered fault-tolerant microservices architecture using **PriorityBlockingQueue** for scheduling, **PostgreSQL** for history tracking, and **Java Socket** for real-time failure alerts and CPU/memory monitoring.

RESEARCH EXPERIENCE

Embodied AI and Vision-Language-Action Research, supervised by Prof. Chenyan Xiong — Dec. 2025 – Present
- Developed a distribution-aware **VLM mid-training** and **VLA fine-tuning** framework with a VLM-VLA **proximity estimator**, yielding stronger downstream embodied policy learning; co-authored **EmbodiedMidtrain** (under review at **COLM 2026)**.

Web Agents and Human-AI Collaboration, supervised by Prof. Graham Neubig and Prof. Jeffrey P. Bigham — Sep. 2025 – Feb. 2026
- Built data processing pipeline for **fine-tuning LLM** as personalized intervention-aware **web agents** based on user interaction styles; co-authored **CowCorpus** (under review at **EMNLP 2026)**.

Dynamic Robot Memory for Open-World Mobile Manipulation, supervised by Prof. Lerrel Pinto — Jul. 2024 – Feb. 2025
- Built and deployed an online **spatio-semantic memory** system for **mobile manipulation** that maintains **dynamic 3D point-cloud** memory and supports **open-vocabulary** object localization, search, and recovery in changing environments; co-authored **DynaMem (ICRA 2025)**, with 70% pick-and-drop success on non-stationary objects and more than 2× improvement over static systems.

Clinical LLMs for Radiology Report Understanding, Supervised by Prof. Artie Shen — Jul. 2024 – Dec. 2024
- Engineered a **knowledge distillation** pipeline to compress GPT-4o-style reasoning into a Llama-3-8B model using 31k clinical report pairs, combining **confidence learning** with **LoRA** fine-tuning for **noisy-label filtering**.

SKILLS

Programming Languages: Python | C/C++ | Java | JavaScript | SQL | MATLAB | R | Rust
Frameworks & Packages: PyTorch | TensorFlow | Spring Boot | Spark | Flask | React | CUDA | LangChain | LlamaIndex | DeepSpeed | Hugging Face
Tools & Platforms: Linux | Git | Docker | Kubernetes | Wireshark | Azure AI | GCP | AWS | CI/CD | Maven | Gradle
Awards: Best Paper Award, CoRL Workshop on Lifelong Learning for Home Robot (2024) | Honorable Mention Award, Mathematical Contest in Modeling (2022)