

Daniel Guo

Pittsburgh, PA | danielg3@andrew.cmu.edu | (929) 530-7872 | [LinkedIn](#) | [Personal Website](#)

EDUCATION

Carnegie Mellon University
New York University

M.S. in Electrical and Computer Engineering
B.S in Computer Science

Aug. 2026 - Jan. 2028
Sep. 2022 - Jan. 2026

PROFESSIONAL EXPERIENCE

Alibaba Group

AI Engineering Intern

Jun. 2026 - Aug. 2026

- Consolidated Q&A and development workflows across **Alibaba's promotion pricing infrastructure** into an **AI platform**, exposing shared domain knowledge and tools as Skills across **Qwen, Taobao**, and other product surfaces
 - Delivered a **multi-agent platform**, orchestrating Q&A, requirements tickets, PRD generation, technical design, and **TDD AI coding** via Skills and handoffs; ensured consistency through **distributed locking, transactions, and idempotent processing**, while decoupling artifacts via **OSS**, achieving a 76% acceptance rate for first-draft PRDs
 - Built an LLM-assisted support **data flywheel** that transformed expert-resolved cases into **reusable knowledge** through LLM-based extraction and quality control, BGE-M3 embeddings, and DBSCAN clustering, while adding API-backed Agent Skills for unmet capabilities; increased self-service resolution from **65% to 86%**
 - Prototyped an offline, trace-based RAG evaluation framework over queries, measuring **token-level context efficiency** and **ROUGE-L** answer coverage to identify redundant retrievals and low-value context; guided knowledge-base and retrieval-configuration improvements, reducing average response latency from **70 to 40 seconds**

Shein

Machine Learning Engineering Intern

Mar. 2026 - Jun. 2026

- Built a coding agent evaluation and governance framework for Java codebases to address the mismatch between public benchmarks and internal code and provide reproducible metrics for model and Skill iteration
 - Curated a **SWE-bench-style** benchmark by screening historical MRs; used agents to parse commit diffs, generated an average of six **F2P/P2P unit tests** per instance, and created problem statements, yielding **2,100 benchmark instances**
 - Devised a **closed-loop, multi-agent benchmark** framework that diagnosed statements and test failures from agent execution traces, automatically refined statements and tests, and **distilled failure patterns** into reusable **prompt constraints and few-shot examples**; increased the **valid-instance rate to 87%** and **reduced average calibration iterations by 32%**
 - Architected an agent eval platform using **Graph-Engineering**, with **YAML-defined DAG** workflows executed through **FastAPI** and **E2B sandboxes**; integrated **pass@k/pass^k** metrics with **7 strategies**, enabling evaluation and continuous regression testing for **100+ Skills** across models and prompt configurations

Z.ai

Software Engineer Intern

Dec. 2025 - Mar. 2026

- Built a **Claude Agent SDK**-based **agent harness** for Mercedes-Benz's QNX/ASPICE SWE.6 testing, orchestrating Writer-Reviewer-Repair agents and custom **MCP and tools** through case-level test generation, review, repair, and validation
 - Constructed a **multi-agent** orchestration framework with a layered Orchestrator-Writer-Reviewer-Repair architecture and **asynchronous collaboration** mode to reduce context rot and improve case completion from **30% to 87%**
 - Developed an agent control framework with isolated contexts, role-scoped permissions, tool/memory constraints, structured handoffs, and Git-backed checkpoints, **achieving an 88% recovery rate** for long-running task failures in large monorepos
 - Benchmarked agent tools for QNX test-log analysis and refined tool names, descriptions, and error responses to improve stability and context, reducing the tool-selection error rate to 7% and per-call context length by 54%

Goldman Sachs Gao Hua Securities

Machine Learning Engineering Intern

Jun. 2024 - Sep. 2024

- Built an LLM-powered financial research platform with LangChain-orchestrated RAG for grounded market analysis, using crawlers to collect data before processing it into offline BGE-M3 embeddings indexed in Milvus
 - Engineered an **OCR-based** section extraction workflow for certain filing sections that embedded only high-value content, lowering indexing cost while achieving **96%** extraction accuracy and improving downstream RAG retrieval relevance
 - Optimized metadata-based pre-filtering, a hybrid retrieval pipeline combining **BM25** and **semantic embedding search**, and **Cohere Rerank** for final-stage reranking, improving **Recall@20** on a multi-source financial corpus from **62% to 84%**

PUBLICATIONS

[Glo-VLMs: Leveraging VLMs for Fine-Grained Diseased Glomerulus Classification](#) [SPIE Medical Imaging](#) (First author)

[MetaboliteChat: A Unified M-LLM for Metabolite Analysis and Functional Insights](#) Technical Reports (First author)

[From Classification to Cross-Modal Understanding: Leveraging VLMs for Fine-Grained Pathology](#) MELBA (First author)

RESEARCH EXPERIENCE

Cornell University

Research Assistant (Advisor: Prof. [Mert Sabuncu](#) & [Ruining Deng](#))

Mar. 2025 - Present

- Built a **Qwen2.5-VL** pipeline to curate ~70K image-text pairs, automating image classification and caption rewriting
- Established a few-shot benchmark for glomerulus classification, comparing CLIP, PLIP, and CONCH through full fine-tuning, LoRA, adapters, and classifier; evaluated Accuracy/AUC/F1, image-text alignment, similarity gap, Silhouette score, and ECE

UC San Diego

Research Assistant (Advisor: Prof. [Pengtao Xie](#))

Jun. 2025 - Sep. 2025

- Built a pipeline to curate metabolite data and generated 1.08M QA pairs, using DeepSeek-8B to reduce truncation by 30%; tuned Vicuna-13B with GNN/CNN and an MLP, achieving ~20x higher BLEU and ~1.7x higher text similarity than GPT-4o

PROJECTS

- Qwopus-Reasoning**: Post-trained Qwen through **SFT** on reasoning and math datasets, **GRPO** on DAPO-Math, and **MTP** integration, improving GPQA accuracy by 16% and inference throughput by 19% over the base Qwen

SKILLS

- Python, Java, SQL, JavaScript, PyTorch, RAG, LangChain/Graph, Kubernetes, LoRA, GRPO, PPO, DPO, Agentic RL, Spring Boot, FastAPI, Flask, Node.js, React, MongoDB, Redis, AWS, MyBatis, MinIO, RabbitMQ, Kafka, GCP, Azure, CI/CD