Zhenghang Zhao

Zz3410@columbia.edu | (917) 886-4627 | GitHub | New York, NY

EDUCATION

Columbia University, Fu Foundation School of Engineering — New York, NY
Master of Science in Computer Engineering — *September 2025 – Present*

University of Georgia, Franklin College of Arts & Sciences — Athens, GA
Bachelor of Science in Computer Science — *January 2023 – May 2025*

Relevant Coursework: Data Structures, System Programming, Software Engineering, Web Programming, CUDA C for GPU, Algorithms, Computer Networks, Computer Architecture, Mobile Application Development, Machine Learning, Natural Language Processing, Math for Deep Learning, Blockchain Technology

EXPERIENCE

LLM-Augmented Automated Bug Repair Research — Remote
Graduate Student Researcher, supervised by Prof. Juan Zhai, UMass — *January 2026 – Present*

- Extending the **RepairAgent** automated bug repair framework on the **Defects4J benchmark** (835 bugs; baseline agent repaired 162) by integrating a specification generation pipeline to provide semantic repair guidance beyond the base agent's capabilities
- Designed and implemented a spec generation module within the existing RepairAgent workflow, prompting **GPT-4o / GPT-4o-mini** to derive pre/post-condition specifications from Javadocs and test suites, then injecting generated specs to guide the repair agent toward correct patches
- Iterated on LLM prompt design to produce accurate, structured specifications; identified that invoking spec generation as a dedicated **LLM call** (rather than a tool wrapper) produced higher-quality outputs better aligned with the repair agent's reasoning process
- Demonstrated early evidence of improvement: the spec-augmented agent successfully repaired bugs that the unmodified RepairAgent failed on from the **Defects4J 835-bug** corpus — full evaluation ongoing

LLM Medical QA Research — Remote
Undergraduate Researcher, supervised by Prof. Liu Zhi, Columbia University — *August 2024 – February 2025*

- Fine-tuned **Meta Llama 3.1B** for medical question-answering using **QLoRA**, training across **9,510 steps over 3 epochs** on a remote dual RTX 3090 workstation via SSH — reducing training loss from **3.37 to 1.18** with **∼117M trainable parameters**
- Curated a combined training dataset of **∼107K samples** from 3 public medical sources (ChatDoctor 96K, MedQA-USMLE 10K, PubMedQA 1K) using **Pandas**; identified that PubMed-only fine-tuning produced overly clinical outputs and resolved this by blending ChatDoctor data to achieve natural doctor-style conversational responses
- Evaluated output model quality using **Hugging Face benchmark datasets and metrics**, then published the final **16.08GB merged model** to Hugging Face Hub for open-access community use and continued fine-tuning
- Built a fault-tolerant training pipeline with **checkpoint management** to recover from GPU memory crashes, tuning batch size and hyperparameters to fit dual RTX 3090 GPU constraints across the full training run

Chinese Academy of Sciences Institute of Automation — Remote
Deep Learning Intern — *August 2024 – September 2024*

- Surveyed and benchmarked **15+ deep RL algorithms** (Double DQN, Dueling DQN, Prioritized Experience Replay) and authored a technical report comparing performance tradeoffs for autonomous decision-making tasks
- Implemented DQN from scratch in **PyTorch**, training an agent on OpenAI Gym CartPole-v1 across **10,000 episodes** with a **50,000-transition experience replay buffer** and **∼3.2M total gradient updates**, with training data generated on-the-fly via agent-environment interaction
- Tuned hyperparameters (epsilon-greedy decay, learning rate, replay buffer capacity, batch size) across algorithm variants; visualized training curves in **Matplotlib** and delivered weekly progress presentations to research mentor

PROJECTS

AI Expense Guard | *In Development* — Oct 2025 – Feb 2026

- Architected a full-stack personal finance app with a mobile client and **FastAPI** backend supporting async communication, with multi-user auth and account management
- Built an AI financial advisor module using **LLM prompt engineering** and **Retrieval-Augmented Generation (RAG)** — user transaction history is embedded into **Pinecone** (vector database) and retrieved at inference time to generate personalized, context-aware financial advice
- Designed a dual-database persistence layer combining **PostgreSQL** for structured records (income, spending categories, monthly budgets) and **Pinecone** (vector database) for semantic retrieval, with **Firebase** as a local on-device cache for offline access
- Integrated **Celery + Redis** as a message queue to handle async tasks such as AI inference and financial analysis, decoupling long-running operations from the request lifecycle for improved reliability and scalability
- Containerized backend services with **Docker**, provisioned AWS infrastructure using **Terraform**, and automated build and deployment pipelines via **GitHub Actions** CI/CD

TECHNICAL SKILLS

Languages:	Python, Java, C/C++, JavaScript/TypeScript, HTML/CSS
AI/ML:	PyTorch, TensorFlow, HuggingFace Transformers, QLoRA, LLM, RAG, OpenAI API, NumPy, Pandas, Matplotlib
Backend & Infrastructure:	FastAPI, Celery, Redis, Docker, AWS, Terraform, GitHub Actions
Databases:	PostgreSQL, Firebase, Pinecone
Tools & Concepts:	Git, Linux, REST APIs, distributed systems, CI/CD