Zengxiao He

+1 (650)-785-9266 | zengxiao@stanford.edu | San Francisco, CA | linkedin.com/in/zengxiao-he

Education

Stanford University, Stanford, CA Sept. 2024 - Jun. 2026
M.S in Electrical Engineering, GPA: 3.8/4.0 | AI & Machine Learning
Teaching Assistant: CS 295 Software Engineering (Prof. Sara Achour)

Central South University, Changsha, China Sept. 2020 - Jun. 2024
B.Eng. in Software Engineering, GPA: 3.9/4.0

Experience

Oracle, Redwood City, CA Jun. 2025 - Sept. 2025
Software Engineer Intern – AI/ML

- Developed an **AI-powered** automation agent end-to-end, from architecture to deployment that automated three core workflows in **Oracle's supply chain** platform, replacing manual, error-prone processes.
- Built automation pipelines using **Node.js**, **Puppeteer**, **Slack API**, and **OCI Queues**, integrating **Llama-4** for natural language root-cause generation and deploying via **Docker** with **CI/CD** pipelines and **90%+ unit test coverage**.
- **Cut** backport processing **time by 40%** and improved workflow visibility, leading to faster release cycles.

Glowia AI, Stanford, CA Nov. 2025 - Mar. 2026
Co-Founder & Full-Stack Engineer – Conversational AI for MedSpa

- Built an automated customer engagement agent using **TypeScript** and **Python (FastAPI)** that processed Instagram DM conversations and converted leads into booked appointments, handling the full inquiry-to-deposit workflow.
- Developed a **React** dashboard for MedSpa operators to monitor conversation flows, booking conversion rates, and agent performance metrics in real time.
- Integrated multiple LLM providers with configurable prompt pipelines and **function-calling APIs** for context-aware response generation, intent classification, and automated scheduling actions.

Stanford University, Stanford, CA Dec. 2024 - Apr. 2026
Research Assistant — Prof. Tom Lee, School of Engineering

- Built an associative memory system for LLM agents inspired by Hebbian learning, replacing standard **RAG** with graph-based spreading activation to surface contextually relevant memories across multi-hop connections.
- Implemented a multi-stage retrieval pipeline in **Python** (**PyTorch**, **NetworkX**) combining embedding similarity, YAKE keyword extraction, and time-decay scoring, eliminating per-node LLM calls and reducing encoding cost to near zero.
- Evaluated on LongMemEval-S benchmark (500 multi-turn QA items), improving associative recall over embedding-only baselines through dual-pathway ranking of semantic and graph-activated results.

Sumixer, Shenzhen, China Apr. 2024 - Sep. 2024
Full-Stack Engineer Intern

- Built an AI-powered dispatch optimization platform for local home services using **Python (FastAPI)**, **React**, and **PostgreSQL**, reducing provider idle **time by 35%** in simulated workloads.
- Implemented a weighted scoring engine combining geospatial distance, skill matching, and workload balancing to **assign 200+ daily orders across 30 providers**, with LLM-assisted ranking for edge cases.
- Developed a real-time operator dashboard with live map visualization, KPI tracking, and manual override, processing dispatches in under 2 seconds per assignment.

Skills

Programming: Python, TypeScript, JavaScript, Java, C/C++, SQL, HTML/CSS, Go
AI/ML: LLM Integration, Prompt Engineering, Function Calling, Agent Architecture, RAG, Fine-tuning
Frameworks: React.js, Node.js, FastAPI, Flask, Django, Next.js, Angular, Vue, Spring Boot, Spring MVC, MyBatis
Developer Tools: Docker, AWS, Azure, Git, Linux, CI/CD, Nginx, Maven, Gradle
Databases: PostgreSQL, MongoDB, MySQL, Redis, DynamoDB