Zheng (John) Jiang

AI Developer | Python/FastAPI/vLLM | Robotics and Multimodal AI
transpxvs@gmail.com | +1 437-889-9013 | Authorized to work in Canada | No sponsorship required
LinkedIn: linkedin.com/in/zheng-jiang-921aa51b2 | Portfolio: zheng-john-jiang-portfolio.pages.dev | GitHub: github.com/parker557/portfolio

SUMMARY

Computer Science graduate focused on practical AI application tooling: model-serving APIs, vision-language demos, robotics perception workflows, and full-stack prototypes. Recent work includes FastAPI/vLLM serving for Qwen-family models, RealSense depth input with SAM 3 tracking, Qwen3-VL batch detection scripts, and Jetson AGX Orin testing. I am also comfortable with the parts around the model: Docker/Linux setup, request formats, logs, timing checks, rendered outputs, and handoff notes.

EDUCATION

The Hong Kong Polytechnic University - Bachelor of Science (Honours) in Computing, Aug 2025
Relevant coursework: Data Structures and Algorithms, Operating Systems, Computer Networks, Database Systems, Software Engineering, Computer Systems Security.

TECHNICAL SKILLS

Languages: Python, JavaScript, TypeScript, Java, C, C++, SQL
AI/ML: PyTorch, Hugging Face, vLLM, Qwen2.5-VL, Qwen3-VL, LangGraph, RAG, LoRA, multimodal AI, SAM 3, V-JEPA 2.1, pandas, NumPy
Backend, Systems, DevOps: FastAPI, Node.js, REST APIs, React, Firebase, Docker, Git, Nginx, Linux, OpenAI-compatible APIs, ESXi, RAID, Jetson AGX Orin, JetPack 6.2, RK3588/RKNN, Intel RealSense

PROFESSIONAL EXPERIENCE

Beijing Qianjue Technology Co., Ltd. - AI Developer Experience | Beijing, China
2025 - 2026

Built robotics AI demos with Python, FastAPI, Gradio, vLLM, and Qwen-family models for instruction following, visual question answering, and multimodal perception. Packaged Qwen2.5-VL-7B and Qwen3-VL behind OpenAI-compatible APIs in Docker/Linux, then wrote benchmark and batch-detection scripts for JSON, CSV, rendered outputs, and timing checks. Connected RealSense D435I color/depth streams with SAM 3 tracking for masks, bounding boxes, 2D/3D anchors, and object-state outputs. Tested edge workloads on Jetson AGX Orin and RK3588/RKNN devices.

Dream Technology Solutions - Co-Founder and Infrastructure Lead | Hong Kong and Remote
May 2020 - Sep 2025

Co-founded a registered Hong Kong technology startup during university and handled client requirements, technical scoping, server delivery, and support. Maintained Dell R730XD, ESXi, RAID 5, Docker, Nginx, reverse proxy, and network-tunneling environments, including deployment fixes and handoff notes. Delivered BeamNG and Minecraft server setups on Zap Hosting, including mod installation, version testing, troubleshooting, and platform-support coordination. Used Codex, Antigravity, Hermes OpenClaw, and Coze for reviewed internal automation experiments.

Microsoft - Large Model Intern | Beijing, China
Jul 2024 - Aug 2024

Supported LLM deployment experiments and cloud server maintenance for model testing. Helped prepare model-test environments, check basic deployment steps, and keep notes clear enough for follow-up work.

WeDragon Technology Limited - Software Engineer Intern | Hong Kong
May 2024 - Sep 2024

Developed AI course materials for beginner machine-learning, LLM, and Python learners, including slide decks, Jupyter labs, exams, and supporting code. Revised materials from feedback and helped with project-site maintenance, backend issue checks, and documentation.

PROJECTS

Multi-Agent Property Management System, Capstone Project
The Hong Kong Polytechnic University | Dec 2024 - Apr 2025

- Built a multi-agent property management prototype with LangGraph, DeepSeek API, and Node.js to coordinate user requests, data processing, and agent workflows.
- Designed an edge-cloud architecture for anomaly detection, decision support, and scalable interaction flows.

RentConnect Secure SaaS Platform
Academic and Startup Hybrid Project | Feb 2024 - Apr 2024

- Built a full-stack lease-record and data-management application with React and Firebase.
- Added authentication, AES-256-based data handling, and layered frontend, backend, and cloud architecture for confidential records.

CERTIFICATIONS

NVIDIA Fundamentals of Deep Learning; Git; Python, Java, C, and C++; Cybersecurity and Computer Networks.