Ziyuan (Ryan) Yan

San Jose, CA | Ziyuan.yan2000@gmail.com | ryan-yan.site | linkedin.com/in/ziyuan-yan | github.com/yanzai-4

EDUCATION

Northeastern University – San Jose, CA | *M.S. in Artificial Intelligence* — Expected Dec 2028

San José State University – San Jose, CA | *B.S. in Software Engineering* · GPA: 3.42 / 4.0 — Aug 2018 – May 2023

EXPERIENCE

Kwon Lab, Northeastern University – San Jose, CA — Oct 2026 – Present

Graduate Student Researcher (Advisor: Prof. MiYoung Kwon) — Unity, C#, Python, PyTorch

- Building a VR and tablet version of the lab's **gaze-contingent training**, which helps people with central vision loss learn to see with a new retinal spot (a "pseudofovea") in place of the damaged fovea.
- Making consumer VR headsets research-grade by pre-training eye-movement models on lab data and fine-tuning on headset data.

RoboForce – Milpitas, CA — Mar 2025 – Jul 2026

Full Stack Engineer (RLHF Platform) — React, Rust, Python, WebSocket, GCP

- Built the **RLHF platform** for solar-panel maintenance robots: a live map of every robot and panel with position and status, where operators track and command robots in real time.
- When a robot is unsure, an operator takes over to finish the complex task, and the takeover is saved as **training data**.
- Brought robot-to-operator updates **under 500 ms** both ways, so operators catch anomalies without delay.

Anuttacon – Mountain View, CA — May 2025 – Apr 2026

AI Trainer (LLM Post-training) — SFT, reward modeling, red-teaming

- Built SFT and reward data for **LLM game characters** with long-term memory and personality that steer the story.
- **Red-teamed** LLMs with jailbreaks and labeled hallucinations by type, giving the team failures to train against.

Werfen – Mountain View, CA — Aug 2024 – Mar 2025

Full Stack Developer (Genomics Software & AI Tools) — C, C++, Python, Qt, Linux, Jenkins

- Cut genomic-pipeline processing time by ∼**10%** by removing redundant data loading and stopping work on invalid samples.
- Built a **RAG assistant** that answers product questions from the documentation, cutting manual lookups for users and support staff.
- Built a **CI/CD pipeline** that automatically builds, deploys, versions, and tests each day's code and emails a results report every morning, catching regressions early.

Cradi MedTech – Los Angeles, CA — Jul 2023 – Aug 2024

Full Stack Developer (Team Lead) — Java, Spring Boot, Vue, Flutter, AWS

- Led 8 engineers to launch a patient-monitoring platform built to HIPAA standards, linking home medical devices, video visits, and NLP-organized patient records (**200 active users**).
- Scaled the distributed backend to **10,000+ requests/sec** with Redis caching, load balancing, synced databases, and backups.

PROJECTS

Yuzu: Multi-Agent AI Workplace – 1st Place, Track 2, The Executable World Hackathon — Sep 2026

Java, React, LLM Agents, MCP, LoRA Fine-tuning

- Built solo in one day a team chat for 8 AI agents where only relevant agents respond: each agent splits chat from work on shared context, so off-topic talk never disrupts its task and instructions **redirect it instantly**.
- Secured every risky agent action with an **AI harness and hard-coded rules**, blocking outside attacks and rule-breaking.
- Improved an open-source model's accuracy by **64%** through fine-tuning, and built the evaluation that picked the production model.

Autonomous Delivery Robot – Senior Project, San José State University — Sep 2022 – May 2023

ROS, C++, Python, CUDA, Jetson TX2

- Wrote the robot's software, from hardware drivers to self-driving control, localization, obstacle avoidance, and path planning, taking an inherited robot to **fully autonomous** delivery across campus.
- Followed Google Maps route waypoints one by one across campus while dodging obstacles and **moving pedestrians**.
- Fused camera and motion-sensor data to localize the robot **within 1 cm** outdoors; planning ran at 50 Hz on its GPU.

SKILLS

Languages: Python, Java, C/C++, TypeScript/JavaScript, Rust, C#, SQL
AI & LLMs: Multi-agent systems, tool calling, MCP, structured output, RAG, LLM evaluation, red-teaming, SFT/RM data
Model Training: PyTorch, Transformers, full fine-tuning, LoRA, transfer learning, mixed precision, TensorFlow
Robotics: ROS, SLAM, EKF/UKF sensor fusion, motion planning, CUDA, Jetson, real-time telemetry
Systems & Cloud: Spring Boot, React, microservices, WebSocket, MySQL, MongoDB, Redis, AWS, GCP, Docker, CI/CD