

Kevin Jiang

Dual Citizen (U.S. & Canada) | k69jiang@uwaterloo.ca

Kevin Jiang Portfolio

EXPERIENCE

Forward Deployed Engineer — RLWRLD — *San Francisco, California, United States* Aug 2026 — Present

- Captured multi-modal teleoperation data using Quest 3, Vive, and Manus gloves, pipeline-converting sensor streams and tactile flags into LeRobot dataset formats to fine-tune the RLDX-1 foundation model for specific tasks
- Partnered with the AWS Prototyping team to translate enterprise customer automation specs (heavy load handling, stamping, pick-and-place) into live feasibility proof-of-concepts within 2-week turnaround cycles
- Upgraded robot controller infrastructure by implementing mobile base teleoperation controllers, auto-homing routines, and state recovery features, ensuring zero-downtime performance during live executive demos

Software Engineer — Hamming AI (YC S24) — *Austin, Texas, United States* Dec 2025 — Apr 2026

- Shipped Red Teaming, an adversarial voice-agent evaluation suite with a 3-axis attack matrix (category × technique × voice modification) and a UCB1/Bayesian selector that adaptively targets gaps, orchestrated on Temporal workflows
- Designed a dual-mode floating chatbot: a discovery agent for in-app navigation and feature guidance, and an RCA mode using LLM tool-calls to pull filter-aware call data and return evidence-backed analysis with inline charts
- Hardened transcript-merge with provenance-aware source arbitration, turn-bounded alignment, and similarity-based artifact filtering; built an HTML diff harness to benchmark pipeline regressions against production samples

Robotics Engineer — Reazon Human Interaction Lab — *Akihabara, Tokyo, Japan* Jun 2025 — Aug 2025

- Designed a lightweight (<150 g) exoskeleton wrist controller in Fusion 360 using Dynamixel actuators for encoder feedback and haptic response, enabling remote teleoperation of an 8-DOF Damiao robotic arm for RL data collection
- Developed a custom wrist-mounted magnetic AprilTags tracking system using dual cameras, Python, and ROS 2 for real-time motion capture, improving tag recognition and bundle pose accuracy by 20% over lab's existing architecture
- Implemented modular nodes with optimized data publishing, allowing for compatibility across different camera setups
- Programmed bridge nodes to map encoder feedback into joint positions using linear interpolation for downstream inverse kinematic and TF calculations in quaternion space to drive universal URDF simulators such as Rviz2/ MuJoCo
- Built an interactive Three.js web interface using Flask and Blender bone rigging for visualization and calibration

PROJECTS

FlowTO | Python | FastAPI | React | TypeScript | cuGraph | PyG | PyTorch | Ollama

- Won 1st at NVIDIA Spark Hack for a city-scale Toronto traffic-prediction app running fully on-device on a DGX Spark
- Trained a PyTorch/PyG graph neural network for traffic-demand prediction (~90% vs. Toronto's dataset) with cuGraph
- Built a Nemotron copilot (Ollama) that turns natural-language intent into bylaw-checked tool calls via local RAG

KaitenCode | Rust | TypeScript | Tauri | React | xterm.js | SQLite

- Built a Tauri (Rust + React) desktop app for orchestrating AI coding agents as a trigger-driven kanban pipeline
- Embedded per-task xterm.js terminal with lazy PTY sessions, enabling shell access and CLI injection from agent triggers
- Exposed a MCP server with 20+ tools allowing external agents to manage workspaces, delegate tasks, and set up scripts

Slothing | Next.js | TypeScript | Drizzle | libSQL | TipTap | TailwindCSS

- Built an AI job-search platform with a semantic profile bank and vector similarity to tailor resume content to positions
- Designed a document editor with streaming AI rewrites, versioning, and PDF export on a provider-agnostic LLM client
- Architected a monorepo with Drizzle/libSQL, NextAuth, provider-agnostic LLM routing, and rate-limited AI endpoints

Expressive Animatronic Head | Python | PyTorch | ROS2 | Linux | llama.cpp | ESP32 | FreeRTOS | Fusion 360

- Integrated Silero (VAD), Whisper (STT), llama.cpp (LLM), and Zonos (TTS) for real-time speech, gaze, and movement sync
- Optimized CUDA toolkit performance to run multimodal pipeline fully locally on limited VRAM (WSL2/Ubuntu)
- Developed FreeRTOS-based multithreaded firmware on ESP32 for low-latency synchronized actuation of 12 servos

EDUCATION

University of Waterloo — BAsC in Computer Engineering

Sept 2024 - Present