Aman Zaveri

San Francisco, CA | a2zaveri@uwaterloo.ca | azaveri.com | linkedin.com/in/amanzav | github.com/amanzav

SKILLS

AI & Evals: LangGraph, PyTorch, LLM evals, regression testing, fine-tuning, RAG, MCP, OpenAI API, Anthropic SDK

Platform & Data: Python, TypeScript, SQL, FastAPI, PostgreSQL, pgvector, Redis, Kafka, MotherDuck, AWS, Docker, Kubernetes

EXPERIENCES

Vooma Sep 2026 – Present
AI Platform Engineer *San Francisco, CA*

- Rebuilt Vee's LangGraph architecture from separate chats into **one ongoing conversation**; persistent user profiles, time-based clustering for compaction, and searchable history across **100K messages**.
- Extracted common workflows from nightly runs; **Jev selects the workflow for each message**, cutting execution to **~⅓ the steps**, while parallel tool calls reduce runtime **by over half**.
- Helped fine-tune **Qwen3-ASR-1.7B** on **200 hours** of dispatcher audio using human-corrected pseudo-labels + noise augmentation; **60% lower serving cost** and **10% WER versus Deepgram's 16%**.
- Replaced manual prompt edits with configurable procedures and **one-click failure labeling**; agents automatically repair configs and rerun **500 snapshot/simulation cases** across **30 SOPs** to verify fixes and catch regressions.
- Built a MotherDuck **ETL pipeline** for usage data and nightly Jev/LLM evals; retention cohorts surfaced failing workflows, guiding fixes that raised **30-day retention from 22% to 44%**.

Boomerang Inc. Jan 2026 – Jul 2026
AI Engineering Intern *New York, NY*

- Solo-built our local software factory which shipped **2.2x more tickets per cycle**.
- Swapped our gradient-boosted ranker for a two-stage neural retriever (PyTorch + FastAPI) on **12M candidate-to-job matches**, 4x more good hires per search (**8% → 31% acceptance**).
- NLP search over **2M alumni embeddings** (Postgres + pgvector) replaced 300+ filters. Recruiters type "founders who left FAANG in 2023"; fine-tuned on misses. **Outreach 4x'd**.
- Shoved Claude into every dev step. In-app bug widget (Kotlin backend, React UI) runs RAG over Notion + Granola, triages, drafts the fix PR. **70% of bugs close without a human**, **on-call alerts down 80%**.
- Sped up our **HR sync 17x** with Redis caching, 52 min → 3 min on 200K-employee orgs. Knocked **60% off our OpenAI bill** via few-shot field normalization + delta-only calls.

Ford Motor Company Sep 2024 – Dec 2024
Embedded Software Engineering Intern *Waterloo, ON*

- Built a **C++ graph-based route ranker** for the 2027 F-150 and Mach-E lineup (**380K vehicles/year**), reducing unnecessary detours by **41%**.
- Reworked **C++ TPMS interrupts** with deferred packet parsing and event-driven dispatch, cutting end-to-end puncture-alert latency **1.4s → 612ms** across **1.2M vehicles**.

Systems Software Engineering Intern May 2025 – Aug 2025

- Built a **Python/FastAPI** firmware-debugging assistant over **8M Kafka events**, returning root causes and supporting traces; reduced investigations from **45 to 4 minutes**.
- Built a **Python hardware-in-the-loop harness** for CAN replay and fault injection; automated **300 regression cases** covering dropped messages, stale readings, and ECU resets.

Transpire Jan 2024 – Apr 2024
Full-Stack Engineering Intern *Toronto, ON*

- XGBoost ranker on 18 months of conversion data, Playwright scraper for 20+ daily Toronto events, sentence-transformer embeddings. **22 new client signups in Q1**.
- Killed our 14-day manual onboarding with a self-serve Next.js site, Typeform intake, Python infra-assessment, auto-filled Notion templates. **14 → 4 days**, **pipeline 3x'd**.

PROJECTS

AI Tennis Coach azaveri.com/tennis

- Fine-tuned **YOLOv8n** on **~1,200 tennis frames** for player/ball detection at **~84 FPS on an M2 Pro**.
- Combined Kalman tracking + OpenCV homography into **~6 KB rally traces** for Claude coaching across **14 matches**.

EDUCATION

University of Waterloo Sep 2023 – Apr 2028
BASc in Mechatronics Engineering, AI Specialization (GPA: 85%)