

Akmal Anvarov

New York, NY · akmal@anvarov.dev · (512) 210-6689 · anvarov.dev · github.com/anvarov · linkedin.com/in/anvarov

SUMMARY

Backend-focused software engineer with production experience building document search over OCR'd content at a mortgage servicer. Built and deployed retrieval-augmented search on the modern stack — the same problem, solved with embeddings and vector search instead of keyword indexing. Python and TypeScript, PostgreSQL, AWS.

SKILLS

Languages Python · TypeScript · JavaScript · SQL

Backend Node.js · Express · REST APIs · Server-sent events · React

Data PostgreSQL · pgvector · AWS OpenSearch · Full-text and vector search

AI/ML Retrieval-augmented generation · Embeddings · Tool use / function calling · LLM APIs (Anthropic, OpenAI)

Infrastructure Docker · AWS (EC2, S3, Secrets Manager, SSM, OpenSearch) · nginx · Caddy · GitHub Actions CI · Vitest

PROJECTS

Garage Copilot — Retrieval-Augmented Repair Assistant

TypeScript, Node, PostgreSQL + pgvector, OpenAI embeddings, Anthropic API, Docker · 2026

github.com/anvarov/garage-copilot · garage.anvarov.dev

A question-answering assistant for DIY vehicle repair, grounded in a licensed document corpus with citations back to source.

- Implemented semantic retrieval in PostgreSQL using pgvector, returning ranked passages joined to their source metadata so every answer carries a verifiable citation — a question phrased as a symptom retrieves the right procedure with no keyword overlap
- Added a tool-calling loop so the model queries the NHTSA recall API mid-answer, with the prompt and error path distinguishing a *failed lookup* from a genuine *no recalls found* — so it reports uncertainty rather than implying a vehicle is safe
- Streamed LLM responses over server-sent events with client-disconnect handling, so abandoned requests stop generating instead of billing for tokens nobody reads
- Deployed on AWS EC2 behind Caddy with automatic TLS; secrets loaded from AWS Secrets Manager into the container environment at start and never written to disk; no inbound SSH, with access via SSM Session Manager
- Designed the corpus around licensing constraints — public-domain government data and manufacturer documentation under attribution — rather than scraping forum content, with attribution enforced at the schema level

EXPERIENCE

Career Break

March 2023 – 2026 · Family responsibilities overseas; worked outside engineering. Returned in 2026 with a deliberate focus on backend and applied AI.

Aspen Capital — Full Stack Software Engineer

March 2022 – March 2023 · New York, NY

Mortgage servicing platform. Team of 10 engineers.

- Built the search layer for a document digitisation programme that moved the company from manual, paper-based record retrieval to full-text search across thousands of OCR'd mortgage documents — reducing document lookup from a physical file search to a single query
- Implemented full-text indexing and query handling on AWS OpenSearch over documents stored in S3, working alongside one other engineer on the search component of a larger scanning and OCR pipeline
- Stabilised a legacy Python banking web application suffering from unexplained runtime errors, an outdated database, and unmaintained dependencies — rewriting backend components to make failures diagnosable and the codebase safe to change

Opentree Education — Apprentice Software Engineer

December 2021 – March 2022 · Remote

Meditation web application, shipped to production. Team of six, structured apprenticeship with regular 1-on-1s.

- Built front-end features in React as part of a six-person team that rotated across the whole stack rather than splitting by speciality — every engineer worked on both the interface and the server side, and the product reached real users
- Configured and ran the application behind nginx, covering the web-server layer end to end

EDUCATION & CERTIFICATIONS

App Academy — Software Engineering Immersive · 2020

Tashkent Institute of Finance — B.S., Finance · 2012–2016

AWS Certified Developer – Associate · 2022