

Anurag Gowda

gowdaanurag@gmail.com • anuraggowda.com • github.com/AnuragGowda • 443-625-7376

SUMMARY

Backend/platform engineer with production experience building GenAI infrastructure at Visa, including RAG systems, LLM evaluation services, Redis vector search, Celery-based job orchestration, and internal developer platforms. Strong CS/math background with additional systems experience in C networking, distributed lookup, and ML from scratch.

WORK EXPERIENCE

Visa

Austin, TX

Software Engineer -- GenAI Platform

June 2025 -- Present

- Built a production RAG semantic cache for Visa Assist, a globally deployed GenAI product used across 18K+ clients in 186 countries, using Redis vector search, cosine similarity, and embedded SME-curated Q/A pairs.
- Improved answer accuracy from 81% to 96% on internal evaluation sets and reduced hallucination rate by 50% by routing semantically similar queries to validated expert-written responses.
- Led backend development of an internal LLM evaluation platform, designing APIs, async job flows, metric execution pipelines, and LangSmith experiment integrations for team-wide model evaluation.
- Reduced representative evaluation job runtime from approximately 3 hours to under 5 minutes by fixing inefficient library usage, parallelizing metric execution with Celery, and hardening execution with retry backoff, structured logging, and error handling.
- Building a no-code fine-tuning hub integrated with Visa's internal GPU platform, supporting training job creation, evaluation gates, model approval workflows, and inference handoff.

Learn Prompting

Remote

Part-Time Software Engineer / Contractor

August 2023 -- May 2025

- Led migration of learnprompting.org from Webflow to a Next.js, React, and Tailwind application, reducing page load time by 40% while supporting a 574% increase in unique daily visitors.
- Built responsive course interfaces and internal Node.js tooling for content generation workflows, improving course creation efficiency by 3x and contributing to a 312% increase in mobile traffic.
- Shipped a full-stack HackAPrompt application for the Voice and AI Conference in one week using Next.js, Firebase, Tailwind, Auth0, and OpenAI API, supporting 10,000+ prompt queries and 51 queries/second peak load.

PROJECTS

C Networking Projects | C, TCP/UDP, Protobuf, Wireshark

- Built TCP/UDP client-server protocols, a reverse-engineered TCP chat server, and a BitTorrent-like CLI supporting pause, resume, and seeding.
- Implemented Chord distributed lookup from scratch using ring-based routing logic and Protobuf message serialization.

NumPy Neural Network Classifier | Python, NumPy | [Blog](#)

- Implemented a three-layer MNIST classifier from scratch, deriving ReLU, softmax, cross-entropy, and backpropagation manually; achieved 97% accuracy without external ML libraries.

EDUCATION

University of Maryland, College Park

Aug. 2022 -- May 2025

B.S. in Computer Science; Double Major in Mathematics; Dean's List

Graduated in 3 years

- Selected Coursework: Algorithms, Data Structures, Operating Systems, Computer Networks, Machine Learning, Linear Algebra, Real Analysis, Topology, Computational Game Theory.

TECHNICAL SKILLS

Languages: Python, TypeScript, JavaScript, C, C++, Java, SQL, Bash

Backend/Infra: FastAPI, Celery, Node.js, Docker, Kubernetes, Redis, PostgreSQL, REST APIs, Linux

AI Systems: RAG, embeddings, Redis Vector Search, LangChain, LangSmith, DeepEval, OpenAI API, LLM-as-judge evaluation

Frontend/Data: React, Next.js, Tailwind, Firebase, MongoDB, GitHub Actions