Anirudh Heredy

602-793-8590 | aheredy@asu.edu | linkedin.com/in/anirudhheredy | github.com/avh17 | anirudhvhportfolio.vercel.app

Professional Summary

Software engineer specializing in scalable backend systems and AI applications. Worked on production microservices for video streaming and location chatbot solutions. Excel in fast-paced, high-impact technical environments.

Education

M.S. Computer Science Aug. 2024 – May 2026
Arizona State University, Tempe, AZ *GPA: 3.85/4.0*

B.Tech. Computer and Communications Engineering Jul. 2018 – Jul. 2022
Manipal Institute of Technology, Manipal, India *GPA: 7.27/10*

Technical Skills

Languages: Python, SQL, Bash
Backend Engineering: FastAPI, RESTful APIs, JWT, OAuth2, GraphQL, Microservices Architecture
Cloud & DevOps: AWS (Lambda, EC2, S3, SQS, ECR), Docker, CI/CD
AI & Machine Learning: RAG (Retrieval-Augmented Generation), Large Language Models(LLMs), LangChain, LangGraph, Vector Embeddings, PyTorch
Databases: PostgreSQL, SQLite, MongoDB

Professional Experience

Software Intern Jul. 2025 – Aug. 2025
Potters Tech *Remote*

- Designed and built the backend for a location-aware Large Language Model(LLM) chatbot, enabling natural language queries with proximity search, ranking, and category-based filtering over geospatial data.
- Reduced hallucinations by implementing a Retrieval-Augmented Generation (RAG) pipeline grounded in OpenStreetMap data, ingesting 2,000+ POIs into PostgreSQL with PostGIS for sub-200 ms spatial queries.
- Built multi-step LLM workflows using LangChain and LangGraph to support conversational state and multi-turn query refinement across user sessions.

Associate Software Developer Aug. 2022 – Sept. 2023
Valtech India *Bengaluru, India*

- Built the backend for a multi-tenant video streaming CMS, designing a configuration-driven architecture that reduced new client onboarding effort by enabling customization through JSON-based settings.
- Developed 15+ low-latency RESTful APIs and GraphQL queries using FastAPI; secured all administrative endpoints with JWT-based authentication, OAuth2, and role-based access control across tenant environments.
- Designed and managed data storage across PostgreSQL and MongoDB (10+ entities), and improved local setup time by 50% by containerizing backend services with Docker.

Projects

AI Study Buddy Nov. 2025

- Supported document ingestion, AI-driven study plans, and contextual QA by building backend services using FastAPI to deliver 5+ REST endpoints under a 24-hour hackathon timeline.
- Extracted structured topics from course material and generate personalized study guidance by integrating 3 external services (Canvas LMS, Supermemory API, and Claude API).
- Handled multiple document uploads and queries per session while ensuring reliable data flow between backend and frontend components by managing data persistence using SQLite.

Serverless Face Recognition (Lambda/FaaS) Apr. 2025

- Processed video frame streams in real-time via SQS event-driven architecture.
- Reduced infrastructure management overhead to zero by architecting a fully serverless face recognition pipeline using AWS Lambda, SQS, and ECR with automatic scaling.
- Enabled seamless client integration by exposing face detection via Lambda Function URLs accepting base64-encoded frames and returning recognition results through SQS response queues.

Elastic Face Recognition (Serverful/IaaS) Mar. 2025

- Built a production-ready distributed ML inference system deployed across auto-scaling EC2 instances with S3 persistence for input images and classification results.
- Enabled on-demand scalability from 0 to 15 concurrent instances by designing a custom autoscaling controller that dynamically provisions EC2 app-tier instances based on SQS queue depth.
- Achieved real-time face recognition inference by using PyTorch deep learning models deployed on custom EC2 AMIs.