Anirudh Herady

602-793-8590 | aherady@asu.edu | linkedin.com/in/anirudhherady | 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

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

Manipal Institute of Technology — Manipal, India
B.Tech. Computer and Communications Engineering — *Jul. 2018 – Jul. 2022*

Technical Skills

Languages: Python, SQL, Bash
Backend Engineering: FastAPI, RESTful APIs, JWT, OAuth2, GraphQL, Microservices Architecture
Cloud & DevOps: AWS (Lambda, EC2, S3, SQS, ECR), boto3 AWS SDK, 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 the backend for a location-aware Large Language Model(LLM) chatbot, supporting natural language queries including proximity search, rating-based ranking, category filtering.
- Reduced hallucinated responses by grounding outputs in OpenStreetMap data using a Retrieval-Augmented Generation (RAG) system.
- Achieved sub 200 ms radius-based spatial queries by ingesting 2000+ OpenStreetMap POIs into PostgreSQL using PostGIS extension.
- Enabled conversational state and multi-turn query refinement across sessions by building multi-step LLM workflows using LangChain and LangGraph.

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

- Built the backend for a multi-client video streaming content management system(CMS).
- Reduced new client onboarding effort by designing a configuration-driven architecture using JSON-based client settings.
- Delivered 15+ low-latency RESTful APIs, GraphQL queries and well-documented endpoints using FastAPI.
- Secured 100% of CMS and administrative endpoints and enabled role-based access control across tenant environments by implementing JWT-based authentication with OAuth2.
- Managed data storage using PostgreSQL and MongoDB, modeling and querying 10+ relational entities while supporting flexible, schema-less client configurations.
- Improved local setup time by 50% and ensured consistent environments across development and staging by containerizing backend services using Docker.

Projects

Elastic Face Recognition (Serverful/IaaS) | *AWS (EC2, SQS, S3), boto3 AWS SDK, Python, PyTorch*

- 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.

Serverless Face Recognition (Lambda/FaaS) | *AWS (Lambda, SQS, ECR*, boto3 AWS SDK, Python, PyTorch, Docker

- 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.