Anirudh Herady

602-793-8590 | aherady@asu.edu | linkedin.com/in/anirudhherady | github.com/avh17 | avhportfolio.lovable.app

Education

Arizona State University — Tempe, AZ
M.S. Computer Science – GPA: 3.78/4.0 — *Aug. 2024 – May 2026*

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

Technical Skills

Languages: Python, JavaScript, TypeScript, C/C++, SQL, Bash
Cloud & DevOps: AWS (Lambda, EC2, S3, SQS, ECR), Docker, CI/CD (Github Actions), Vercel, Render, Linux
Backend Engineering: FastAPI, Node.js, Flask, Microservices Architecture, RESTful APIs, GraphQL
AI & Machine Learning: RAG (Retrieval-Augmented Generation), LangChain, Vector Embeddings, LLMs (GPT, Gemini), Transformers, Hugging Face
Databases: PostgreSQL, MongoDB, Redis, SQLite, SQLAlchemy ORM
Frontend: React.js, Next.js 14, Tailwind CSS, Streamlit

Professional Experience

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

- Engineered and shipped a full-stack Location Intelligence Chatbot within an 8-week timeline, enabling non-technical users to query complex geospatial data without analyst support
- Architected a Retrieval-Augmented Generation (RAG) pipeline combining database retrieval with LLM synthesis; improved response relevance by 60% by fine-tuning context injection strategies
- Integrated OpenStreetMap (OSM) data and location APIs to power real-time Point of Interest (POI) search, achieving 95% accuracy across 2,000+ locations in 50+ cities

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

- Designed and deployed 8 FastAPI microservices and 20+ REST/GraphQL endpoints to support a multi-tenant video streaming platform, ensuring high availability for concurrent media consumption
- Optimized deployment pipelines by implementing Docker multi-stage builds, successfully reducing container image sizes by 60% and accelerating build/deployment velocity
- Designed normalized PostgreSQL schemas for data integrity and MongoDB collections with compound indexes, optimizing complex query performance for high-volume user data
- Implemented robust RBAC (Role-Based Access Control) utilizing JWT and OAuth2 to manage authorization across 3 distinct user permission levels
- Conducted code reviews for 20+ pull requests and led onboarding sessions for junior developers to maintain code quality standards

Projects

Distributed Face Recognition System | *AWS (Lambda, EC2, SQS, S3), Python, Docker*

- Designed and implemented two distinct architectures to process video frames for face recognition, analyzing trade-offs between latency, concurrency, and operational cost
- Built a Python-based autoscaler for the EC2 implementation that dynamically provisions instances (scaling 0-15 nodes) based on SQS queue depth metrics to handle traffic spikes
- Achieved ¡3s latency and 99%+ accuracy under a load of 100+ concurrent requests by utilizing a two-stage ML pipeline (MTCNN for detection, ResNet for recognition)
- Decoupled ingestion and processing layers using SQS to ensure system reliability during burst traffic

StoreIt: Scalable Cloud Storage Solution | *Next.js 14, TypeScript, Appwrite*

- Engineered a full-stack file management system featuring secure authentication and efficient file storage using Appwrite
- Built a responsive, component-driven UI with 20+ reusable React components, including dynamic data visualization charts for storage usage analytics