HERAMB VIJAY DEVBHANKAR

Senior Software Engineer | Backend, Distributed Systems, Cloud-Native Data Platforms, Generative AI

Relocating to India from USA | +91-8097204338 | heramb.hvd@gmail.com | linkedin.com/in/heramb-devbhankar

SUMMARY

Senior Software Engineer with 5+ years of experience, currently building scalable backend systems and cloud-native data platforms for real-time analytics and quant models at Apollo Global Management. Delivered **60%+ lower API latency**, **70% faster data pipelines** and **60% lower LLM inference cost** using Python, FastAPI, SQL Server, Redis, Temporal, Azure (AKS), AWS and Kubernetes. Builds Generative AI systems with LLM agents, Model Context Protocol (MCP) and text-to-SQL.

EDUCATION

Master of Science (MS) in Computer Science, University of Southern California, Los Angeles, CA — **Jan 2021 - Dec 2022**

Bachelor of Technology (B.Tech), Dr. Babasaheb Ambedkar Technological University, India — **May 2019**

EXPERIENCE

Senior Software Engineer | **Apollo Global Management**, El Segundo, CA — **Jan 2025 - Present**

- Cut API response latency by **60%+** under heavy load by building high-performance backend microservices and REST APIs in Python (FastAPI) on SQL Server, with Protobuf serialization, query optimization and Redis caching.
- Reduced daily pipeline runtime by **70% (13 hrs to 4 hrs)** and monthly cycles from **4 days to 1 day** by designing fault-tolerant, asynchronous workflow orchestration with Temporal and Azure Blob Storage for multi-stage data ingestion.
- Lowered LLM inference latency and token cost by **60%** by re-architecting Generative AI infrastructure with Model Context Protocol (MCP), tool calling and text-to-SQL agents, using MapReduce-style parallel chunking, grounding and eval guardrails.
- Architected a cloud-native data platform processing high-dimensional data streams (**78K x 370 matrices per run, 25M+ rows/year**), building high-throughput indexing and retrieval layers for real-time analytics and quant models.
- Deployed distributed services on Azure Kubernetes Service (AKS) with Docker, Terraform and Helm; built CI/CD and GitOps pipelines (GitHub Actions, Flux CD) for high availability and continuous deployment to production.

Software Development Engineer | **SAG-AFTRA Foundation**, Los Angeles, CA — **May 2023 - Oct 2024**

- Led architecture and end-to-end delivery of a **Ruby on Rails monolith to microservices migration** (Python, FastAPI, React, Swift iOS), improving scalability, modularity and deployment velocity.
- Engineered production backend and AWS cloud infrastructure with CI/CD, load balancing, async processing and data-access optimizations, improving system reliability, performance and release frequency.

Software Development Engineer | **Eye Labs Inc**, Santa Clara, CA — **Sep 2022 - Apr 2023**

- Developed facial landmark detection and gaze-tracking algorithms adopted as the **production baseline for FDA clinical trials**.
- Built an iOS app in Swift (SwiftUI, UIKit, Apple Vision) that detects eye anomalies via the front camera, with custom camera and image-processing modules integrated through UIViewControllerRepresentable.

Research Assistant | **University of Southern California**, Los Angeles, CA — **Apr 2021 - Aug 2022**

- Built computer vision pipelines to track movement and gene expression in free-moving Drosophila from GFP video; trained LSTM and CNN models to classify motion trajectories with high temporal accuracy.
- Trained StarGAN models for synthetic face-recognition data generation, improving generalization with less real-world data.

Software Development Engineer | **IOTric**, India — **Aug 2019 - Dec 2020**

- Built a web app that extracts text from distorted invoice/receipt images via object detection, text detection and OCR.

PROJECTS

Spoken2written, Open-Source Python Package (PyPI) | Python, spaCy, NumPy | pypi.org/project/spoken2written

- Published an open-source NLP library on PyPI (pip install spoken2written) that converts spoken-style text to written form.

Tactile Image Generation for the Visually Impaired (Stanford University) | Python, OpenCV, TensorFlow, Flask

- Built a Flask REST API converting RGB images to tactile graphics using instance segmentation and edge detection.

TECHNICAL SKILLS

Languages: Python, SQL, JavaScript, Swift, C++, C#, Ruby

Backend & Data: FastAPI, Flask, Node.js, Ruby on Rails, REST APIs, Microservices, SQL Server, Redis, Protobuf, Temporal

Cloud & DevOps: Azure (AKS, Blob Storage), AWS, Docker, Kubernetes, Terraform, Helm, CI/CD, GitHub Actions, Flux CD, GitOps

AI/ML: Generative AI, LLMs, AI Agents, Model Context Protocol (MCP), Text-to-SQL, TensorFlow, OpenCV, CNN, LSTM, GANs, Computer Vision, NLP, spaCy

Frontend & Mobile: React, Angular, SwiftUI, UIKit

Core: Distributed Systems, System Design, Data Structures & Algorithms, OOP, High Availability, Performance Optimization