Himanshu Joshi

Senior Software Engineer — AI/ML & Full-Stack

joshi.himanshu09@gmail.com

PROFESSIONAL SUMMARY

Senior Software Engineer with **10+ years of experience** delivering AI/ML, data science and full-stack web and mobile solutions in enterprise environments. Proven track record of leading AI-powered systems that improved candidate screening accuracy by **30%** and reduced processing costs by **25%**. Skilled in architecting scalable AWS solutions, building RAG and agentic LLM pipelines, and driving end-to-end product development from data ingestion through to the interface.

TECHNICAL SKILLS

Languages	Python, SQL, C, JavaScript (ES6+), TypeScript
Generative & Agentic AI	Agentic AI, LangChain, LangGraph, RAG, Prompt Engineering, LLM Integration, OpenAI (GPT-4 / GPT-3.5), Google Gemini, Anthropic, Vector Databases, Chroma DB, Pinecone, LangSmith, MLflow, Fine-tuning, Whisper, AssemblyAI
Machine Learning	Model Evaluation, Feature Engineering, Random Forest, FinBERT, TensorFlow, PyTorch, Scikit-learn, Pandas, NumPy, spaCy, FER
Backend & APIs	FastAPI, Flask, SQLAlchemy, REST APIs, Node.js, Express, WebSockets
Databases	PostgreSQL, MongoDB, Redis, Prisma
Frontend	React, Next.js, Redux, Hooks, Tailwind CSS, HTML5, CSS3
Mobile	React Native, Android Studio, Xcode, Firebase, Push Notifications, Stripe
Cloud & DevOps	AWS Lambda, Amazon S3, AWS SageMaker, AWS MediaConvert, AWS Rekognition, Amazon CloudWatch, Docker, Jenkins, CI/CD Pipelines, Git, GitHub, Bitbucket

PROFESSIONAL EXPERIENCE

EvluateAI — Team of 16

Senior Software Engineer

Python · AWS Lambda · AWS SageMaker · OpenAI GPT-3.5 · Gemini Pro · LangChain · Whisper · FER · AssemblyAI · AWS MediaConvert

- Led the **end-to-end AI workflow** for video interview evaluation, integrating FER for emotion recognition and Whisper for speech-to-text transcription.
- Developed **multimodal response evaluation** combining GPT-3.5 and Gemini Pro against explicit scoring rubrics, ensuring consistent and defensible candidate assessments.
- Implemented dynamic question generation via LangChain, reducing manual interviewer time by **40%**.
- Fine-tuned LLMs for domain-specific scoring accuracy, reducing false positives and negatives.
- Designed a **serverless AWS architecture** delivering **20%** cost savings and high scalability.
- Enhanced model performance by **35%** through caching, asynchronous processing and optimised inference endpoints.
- Implemented **GDPR and EEOC-compliant** security protocols across the platform.

Carecircle HelpA — Team of 5

AI Engineer / Backend Developer

Python · LangChain · OpenAI GPT-4 · ChromaDB · FastAPI · PostgreSQL · AWS

- Designed and deployed a **Retrieval-Augmented Generation (RAG)** pipeline for secure querying of patient data and clinical literature.

- Implemented **multi-agent orchestration** in LangChain — one agent retrieving patient data, a second generating summaries, a third validating output for compliance before release.
- Built vector embeddings of patient notes in ChromaDB for contextual recall and personalised care insights.
- Developed the RESTful backend in FastAPI, hosted on AWS Lambda, S3 and RDS for scalability and low latency.
- Collaborated with healthcare domain experts to tune prompt templates for clinical accuracy.
- Reduced time to retrieve and summarise patient data by **~60%** and improved automated medical note accuracy by **35%** against the template baseline.

Augmentt

Senior Full-Stack JavaScript Developer

React · JavaScript · Express · HTML5 · CSS3 · AWS S3 · Stripe · FCM

- Developed independent **micro-frontends** and orchestrated them within a container application.
- Built reusable components and front-end libraries, and translated designs and wireframes into production code.
- Optimised components for performance across a wide range of web-capable devices and browsers.
- Implemented notification functionality with FCM, Stripe payment integration, and S3 upload/download using pre-signed URLs.

Resume Parser

Solo

Lead Developer

Python · Flask · PostgreSQL · GPT-3.5 · PyPDF2 · python-docx · spaCy · Pyresparser

- Built an automated **resume parsing engine** extracting structured data from PDF and DOCX documents.
- Integrated NLP libraries for high-accuracy parsing across varied multi-format layouts.
- Developed APIs for PostgreSQL storage and analysis, and used GPT-3.5 to generate skill insights and job-fit recommendations.

QSkipper

Team of 4

Frontend & Backend Developer

React Native · Next.js · Firebase · Redux · Stripe · Google Places API · AWS S3

- Developed user-facing features in React Native and an admin panel in React.
- Integrated Stripe payments, AWS S3 uploads and Google Places autocomplete.
- Managed app releases end to end for Google Play and the Apple App Store.

Capzul

Team of 5

Frontend & Backend Developer

React Native · Next.js · Google Maps API · Redux · WebSocket

- Developed mobile and web front ends on a shared library of reusable components.
- Integrated push notifications and map-based features, and implemented **real-time chat over WebSockets**.

WaterFrontConcerts

Team of 5

Frontend Developer

React · Next.js · Node.js · MongoDB · AWS S3

- Developed web applications with authentication and notification systems.
- Integrated AWS S3 for secure file storage and created reusable UI components and API interceptors.

PERSONAL PROJECTS

TradeCharts

tradecharts.tech

Multi-broker charting & trading workspace

- Built a charting and trading workspace running on the user's own broker account, normalising **six broker APIs** (Angel One, FYERS, Upstox, Dhan, CoinDCX, Bybit) behind a single internal interface.

- Designed a **server-side alert engine** evaluating user-defined conditions against live streaming market data, removing the browser-dependency failure mode of client-side alerts.
- Implemented an AI assistant that authors custom indicators, annotates charts and configures alerts through natural language, with a **confirmation-card pattern** ensuring no action executes without explicit user approval.
- Developed a strategy backtesting engine and paper-trading environment for validating strategies before committing capital.

VectraSearch

Built solo

Multi-tenant vector search SaaS

Node.js · Express 5 · Prisma · PostgreSQL · Pinecone · Redis · BullMQ · React

- Built a three-tier vector search platform end to end — API, customer dashboard and admin console — spanning **118 REST endpoints, 31 data models and 53 UI screens**.
- Designed a **hybrid retrieval pipeline** fusing PostgreSQL full-text ranking with Pinecone vector similarity through Reciprocal Rank Fusion, with per-endpoint keyword/vector weighting.
- Implemented Algolia-style typo tolerance from scratch using Damerau-Levenshtein distance over a per-tenant catalog vocabulary with length-based typo budgets.
- Shipped a RAG chatbot on the same index with a provider-agnostic LLM layer over OpenAI, Anthropic and Gemini, and usage-based billing on a credit ledger with date-versioned pricing.

Safewaters — Environmental statistics platform surfacing water levels, flow rates and safety information for hydropower facilities across the United States.

Encora — Sustainable food-takeout marketplace with separate admin, vendor and customer portals covering the full order lifecycle.

EDUCATION

B.Tech — Bachelor of Technology

2012