

SOHAIL HARESH GIDWANI

Los Angeles, CA

+19736525842 | sohailgidwani15@gmail.com | <https://sohailgidwani.app/>
<https://linkedin.com/in/sohail-gidwani/> | <https://github.com/SohailGidwani>

EDUCATION

University of Southern California

Masters of Science, Computer Science, GPA: 3.5/4.0

• **Coursework:** Analysis of Algorithms, Information Retrieval & Web Search Engines, NLP, ML for Data Science

Aug 2025 - May 2027

Los Angeles, CA

University of Mumbai

B.E., Computer Engineering, CGPA: 9.05/10

• **Coursework:** Artificial Intelligence, Machine Learning, Advanced DBMS, Data Structures & Algorithms, Software Engineering, Big Data Analytics

Aug 2019 - May 2023

Mumbai, India

WORK EXPERIENCE

Keck School of Medicine, USC

Research Assistant - Multi-Modal AI for Alzheimer's Disease

- Architected a multi-modal deep learning pipeline for Alzheimer's disease prediction using T1 MRI, DTI imaging, and clinical data across 2,363 ADNI subjects, achieving 72.7% balanced accuracy on 3-class diagnosis and 93.1% on binary classification (CN vs Dementia).
- Designed missing-modality fusion via cross-attention with modality dropout, enabling robust inference when imaging data is incomplete (39.4% DTI coverage); nearly doubled preclinical amyloid detection sensitivity from 29% to 56% between model iterations.
- Built end-to-end experimentation infrastructure: two-stage training (CLIP contrastive pre-training → multi-task fine-tuning), modality ablation studies across 7 combinations, and confidence calibration analysis on ~160M parameter models.

Oct 2025 - Present

Los Angeles, CA

Insaito, Inc.

Senior Software Engineer - I

- Led architecture of an AI agent builder platform enabling custom workflow creation with OAuth integrations for 100+ third-party apps; built core infrastructure for open-source LLM deployment and Model Context Protocol (MCP) servers.
- Designed and deployed serverless backend services for agent orchestration, supporting concurrent multi-step workflows with tool-calling and context management.

May 2025 - Jul 2025

Remote

IIFL Finance Ltd.

Full Stack Software Developer

- Built an internal RAG chatbot using Python, Flask, Qdrant vector DB, and Azure OpenAI; integrated with Zoho ticketing system to automate employee support workflows. (Certificate of Achievement)
- Engineered an AI-powered Gold Loan Image Audit system using GroundingDINO and Swin-Transformer for automated fraud detection, reducing potential loan fraud by 15%.
- Designed CapitalGenie, an automated support system leveraging GPT-4o and internal APIs to diagnose user issues and generate personalized responses, accelerating resolution time by 70%.

Jun 2023 - May 2025

Mumbai, India

SKILLS

Programming: Python, TypeScript, JavaScript, Java, C/C++, SQL

AI/ML: PyTorch, TensorFlow, Keras, Scikit-Learn, LangChain, NLP, Computer Vision, OpenCV, LLMs, CNN, LSTM, Transformer, Vector DBs, LLMops, MCP

Full-Stack: React, Next.js, Node.js, Express, Hono, FastAPI, Flask, RESTful APIs, HTML5, CSS3, Tailwind CSS, WebRTC

Databases & Cloud: MySQL, PostgreSQL, MongoDB, Oracle, SQLAlchemy, Azure, AWS, Docker, Git, Linux, Serverless

Methodologies: Agile, Test-Driven Development, CI/CD, Code Reviews

PROJECTS

- **Knowledge Hub (Semantic Search & Study Assistant):** Built a Dockerized document management system with Flask and PostgreSQL+pgvector that ingests PDFs, images, and handwritten notes via OCR (OpenCV, PyMuPDF, Tesseract). Implements hybrid search combining full-text and vector similarity (Sentence-Transformers all-MiniLM-L6-v2) with confidence-aware ranking, intelligent chunking (300–700 tokens with overlap), and a RAG-powered Q&A pipeline backed by a local LLM via Ollama.
- **Image Feature Detection & Captioning:** Developed an end-to-end image captioning pipeline using VGG-16 for feature extraction paired with both LSTM (BLEU 0.65) and Transformer (BLEU 0.80) decoders, demonstrating the impact of attention mechanisms on caption quality. Deployed via Streamlit for real-time inference.
- **ScribeGlobe (Full-Stack Blogging Platform):** Built a Medium-style publishing platform with a React/Vite/TypeScript frontend and a serverless Hono backend deployed on Cloudflare Workers for edge computing. Features user authentication, markdown editing with real-time preview, and PostgreSQL persistence.
- **Project-TechUpdates:** Built a tech-news aggregation tool using Python scrapers, Azure OpenAI for AI-powered article categorization, Qdrant vector DB for semantic deduplication, and a TypeScript frontend delivering ad-free, categorized headlines.

EXTRACURRICULAR ACTIVITIES

Won Tech-A-Thon at IIFL Finance. Runner-up at Crack The Code (Jai Hind, 2018).

Organized “Blind Code,” a sub-event of ASCENT 2022 (ISTE).

Courses: 0-100 Cohort 2.0 (Harkirat Singh), TensorFlow for Deep Learning.