SOHAIL HARESH GIDWANI

Los Angeles, CA • sohailgidwani15@gmail.com • LinkedIn • Github • Portfolio • +19736525842

Passionate software engineer focused on AI technologies and full-stack development, with problem-solving and collaborative skills.

EDUCATION

Masters of Science in Computer Science | University of Southern California *Aug 2025 – May 2027*

- **Coursework:** Analysis of Algorithm, Information Retrieval, and Web Search Engines

B.E. Computer Engineering | University of Mumbai *Aug 2019 – May 2023*
CGPA: 9.05/10

- **Coursework:** Artificial Intelligence, Machine Learning, Advanced DBMS, Data Structures & Algorithms, Operating Systems, Software Engineering, Cloud Computing, Object-Oriented Programming, Big Data Analytics, Computer Networks, Cryptography & System Security, Blockchain.

WORK EXPERIENCE

Student Worker / Research Assistant | Keck School of Medicine, USC *Oct 2025 - Present*

- Working on Vision Language Models (VLMs) for multimodal tasks, which involve data set preparation, model training, and evaluation.
- Supporting research that connects visual and textual information for healthcare-related applications.

Senior Software Engineer - I | Insaito, Inc. *May 2025 - July 2025*

- **AI Agent Builder Platform:** Led the architecture and development of an AI agent builder platform, enabling users to create advanced AI agents with custom workflows and deep third-party integrations. Built core infrastructure for open-source LLM deployment, OAuth for 100+ apps, and Model Context Protocol (MCP) servers.

Full Stack - Software Developer | IIFL Finance Ltd. *June 2023 - May 2025*

- **Custom Data Chatbots (RAG):** Built an internal employee support chatbot using NLP, Python, and Flask. Integrated with Qdrant vector database, Azure OpenAI service, and Zoho ticketing system. This AI-powered solution significantly reduced the number of support tickets raised by employees, streamlining internal processes. (Certificate of Achievement)
- **Gold Loan Image Audit App:** Engineered AI-powered application using models like GroundingDino, Swin-Transformer, enhancing fraud detection and reducing potential loan fraud by 15%
- **CapitalGenie:** Designed and implemented an automated user support system that responded to queries via email and other channels. Leveraged IIFL's internal APIs to fetch user information, diagnose issues, and utilized GPT-4o to generate structured, personalized responses. This solution accelerated the resolution of the issue by 70% compared to manual processes.

SKILLS

- **Programming:** Python, TypeScript, JavaScript, Java (intermediate), C/C++ (Basic), SQL
- **Full-Stack:** React, Next, Node.js, Express, Hono, FastAPI, Flask, RESTful APIs, HTML5, CSS3, Tailwind CSS, WebRTC
- **AI/ML:** TensorFlow, Keras, Scikit-Learn, NLP, Computer Vision, OpenCV Vector DBs, LLMs, CNN, LSTM, Transformer, Deep Learning, Generative AI, MLOps, N8N, MCP
- **Databases & Cloud:** MySQL, PostgreSQL, MongoDB, Oracle, SQLAlchemy, Azure, AWS, Docker, Git, Linux, Serverless
- **Methodologies:** Agile, Test-Driven Development, Code Reviews, Backend, End-To-End

PROJECTS

- **Image Feature Detection & Captioning | Neural Networks, Streamlit** – Github
  Developed an AI app using CNN/VGG-16 for image feature extraction and an LSTM/Transformer-based captioning model achieving a BLEU score of 0.80, with a Streamlit interface for user-friendly interaction.
- **Knowledge Hub (Personal Semantic Search & Study Assistant) | Flask, Next.js, PostgreSQL + pgvector, PyMuPDF, Tesseract/TrOCR, OpenCV, Ollama** – Github
  Built a local-first portal that ingests PDFs/images/handwritten notes; runs OCR + chunking; and delivers hybrid search (FTS + semantic) with citations and an LLM-backed answer API via Ollama.
- **Project-TechUpdates | Python, TypeScript, JavaScript** – Github
  Built a headline aggregation and classification tool using Python for scraping, TypeScript for data handling, and a minimal frontend to deliver ad-free, categorized tech news updates.

EXTRACURRICULAR ACTIVITIES

- Event Leadership: Organized "Blind Code," a sub-event of ASCENT 2022 (ISTE).
- Competition Success:
  - Won Tech-A-Thon at IIFL.
  - Runner-up at Crack The Code (Jai Hind, 2018).
  - Participated in Feynwick (KJSIEIT, 2021) and Trident (TSEC, 2019).
- Courses: 0-100 Cohort 2.0 (Harikart Singh), TensorFlow for deeplearning.