Rakshit Shah

+1 (408) 460-3840 | Chicago, IL, 60090
rakshitrajeshshah1@gmail.com| www.rakshitai.info | linkedin.com/in/rakshitshah28/ | github.com/rakshitshah280701

Summary

Agentic AI Engineer with experience in building production-ready solutions using MCP across computer vision, NLP, and multimodal AI. Skilled in LLM fine-tuning, vector databases, MLOps, Computer Vision and cloud platforms (AWS, Azure). Published researcher in IEEE Xplore with expertise in Responsible AI and Generative AI. Currently focused on LLM-based systems, end-to-end ML pipelines, and human-centered AI applications.

Skills

- **Programming Languages:** Python, JavaScript, Java, C++
- **Machine Learning & AI:** PyTorch, TensorFlow, LangChain, LLMs, RAG, FAISS
- **Agentic Systems & Graph:** CrewAI, Model Context Protocol (MCP), Neo4j (Cypher)
- **Frontend,Backend & APIs:** React, Tailwind CSS, HTML/CSS, FastAPI, Flask, REST, WebSockets, Socket.IO
- **Cloud, DevOps & Analytics:** AWS, Docker, GitHub Actions, Linux, Pandas, NumPy, Power BI, Git

Education

California State University East Bay — May 2025
MS in Computer Science — GPA: 3.74 / 4.00

AP Shah Institute of Technology, University of Mumbai, India — June 2023
Bachelor of Engineering in Computer Engineering — GPA: 8.72 / 10

Experience

AI Agentic Engineer | Avesta Computer Service Ltd., Somerset, NJ (Remote from Chicago, IL) **June 2025 - Present**

- Built 0 to 1 MVP - **full-stack AI-powered technical support chatbot** using **React, FastAPI, and Socket.IO**, integrated with **CrewAI** and **Neo4j Graph Database**, and building both the client side functionality and Agent side functionality for support by enabling network troubleshooting, server diagnostics, and customer support with real-time interactive avatar capabilities
- Designed an advanced **Retrieval Augmented Generation (RAG) system with FAISS** vector database and **LongContextReorder** optimization, achieving 40% improved document retrieval accuracy through intelligent document categorization (Defects, Product Docs, Case Notes) and streaming responses.
- Built a **multi-agent AI system using CrewAI** with six specialized agents (e.g., Technical Support, Defect Analysis), leveraging a Neo4j knowledge graph with 218 troubleshooting steps for dynamic workflow routing and real-time sentiment analysis.
- Deployed production-ready application on **AWS (EC2, Amplify) with automated CI/CD** via GitHub Actions, achieving 99.9% uptime, **SSL/TLS encryption, and CORS configuration**, supported by Nginx reverse proxy and systemd for persistent backend operation.
- Refactored the core orchestrator to build CopilotSession to support memory retention, context switching, and event-driven step routing across graph search, command validation, and natural language inputs.

Student Assistant (ML & AI Researcher and Moderator) | CSU East Bay, Computer Science Dept, Hayward, CA **Feb 2024 - Nov 2024**

- Researched and implemented a neural network **from first principles (NumPy)**: derived closed-form updates for bias (θ) and threshold (β), built a tunable-slope sigmoid with a stable derivative, added **layer-width normalization (Zo)** and **ReLU-gated backprop**; vectorized ops to prevent overflow/NaNs and improve convergence.
- Built an experimentation framework with an end-to-end training loop, loss/accuracy tracking, parameter-init utilities, and activation/gradient instrumentation; validated behavior on **XOR-style toy tasks** and compared against a baseline MLP.
- Trained **YOLOv5/YOLOv8** for underwater object detection: curated datasets, set up augmentation/splits, and evaluated with **mAP/precision-recall** to benchmark improvements across test sets.

Data Collection Intern & Moderator | Sciffer Analytics Pvt. Ltd , Pune, Maharashtra, India **May 2021 - Nov 2021**

- Developed web scraping python script (BeautifulSoup, Selenium) to extract 50,000+ images, ensuring high-quality dataset generation for machine learning models.
- Maintained an annotation pipeline using LabelImg, resulting in a 15% improvement in YOLO model accuracy.
- Monitored work of Junior Interns and worked closely with Data Scientist to improve the accuracy of the Model.

Research Paper

1. Linear Regression vs LSTM for Time Series Data — Paper Link- https://ieeexplore.ieee.org/document/9848887
   Role: First Author | **Focus:** Time-series forecasting, deep learning vs. classical baselines
   - Framed a head-to-head evaluation of Linear Regression vs. LSTM for univariate/multivariate series; implemented reproducible training/evaluation with holdout and cross-validation.
   - Engineered features and temporal windows; compared error profiles (MAE/MSE/RMSE) and sensitivity to data length, showing when simpler models are competitive vs. when LSTM's temporal memory helps.
   - Documented performance/complexity trade-offs and recommended model selection heuristics for small vs. larger datasets.
2. Heartbeat prediction using Mel Spectrogram and MFCC value — Paper Link- https://ieeexplore.ieee.org/document/10150129
   Role: First Author | **Focus:** Biomedical audio, feature engineering, supervised classification
   - Built an end-to-end phonocardiogram pipeline: pre-processing/denoising → Mel-spectrogram + MFCC extraction → model training for normal vs. abnormal heartbeat detection.
   - Benchmarked multiple supervised classifiers; analyzed robustness across recordings and emphasized clinically interpretable features.
   - Reported practical guidance on audio feature choices (Mel/MFCC) for resource-constrained systems.

Selected Projects

InstructAware – Generative Instructional Narration for Situational Awareness ***CSU East Bay, iLab | Dec 2024 – March 2025***
Built a multimodal real-time navigation system using YOLOv8 + PaddleOCR with fine-tuned LLMs (DeepSeek-R1, GPT-3.5, T5, LLaMA) integrated via InstructAware into an Android app (SceneSense) for signboard detection and text extraction; applied RLHF and BLEURT/COMET/SBERT evaluation to boost contextual accuracy and reduce hallucinations.

DevMind - AI-Powered Developer Productivity Agent
Built a local-first dev assistant on Claude Desktop using Model Context Protocol (MCP), delivering custom MCP servers for file create/edit/search/move and intelligent project-structure analysis with codebase indexing.

Personal Portfolio Website | www.rakshitai.info
Built a responsive React + Tailwind portfolio with animated UI and dynamic routing—integrating a gemma3 LLM chatbot over personal data (FAISS semantic search) for real-time Q&A—and a serverless Next.js backend handling keepalive, Slack alerts, IP geolocation, and Hetzner wake-ups.

StockSage - AI-Powered NSE Stock Forecasting App
Built a full-stack stock prediction platform: GRU models trained on 5 years of NSE data with TA indicators (RSI, MACD, Bollinger Bands, ADX), an automated per-symbol pipeline for scaling/feature engineering/windowing, and REST APIs for live tickers.