Sanket Sonkusare

sanketsonkusare.me | sanketsonkusare | sanketsonkusare01@gmail.com | 7559448705

Work Experience

AI Engineer, ScroBits Technologies Aug 2025 - Present

– • Built an Agentic Enterprise RAG system using FastAPI, LangGraph, and Pinecone, enabling multi-tenant semantic search, human-in-loop escalation, and real-time dashboards, reducing query time by 60% and token usage by 40% used FastMCP for tools and prompts.
• Built a multi-agent AI support chatbot using FastAPI, LangGraph, Pinecone, and Google Gemini, achieving 95% query routing accuracy, 40% latency reduction, real-time human escalation, and zero-config embeddable deployment.
• Engineered and deployed a multi-agent AI system using LangChain, Gemini, and RAG with pinecone DB, integrating Retriever and Query agents via FastAPI and WhatsApp webhook for a Media Monitoring & Knowledge Assistant.

Machine Learning Intern, Manastik Nov 2023 - Apr 2024

– • Leveraging TFLite, built a high-performance MediaPipe pose detection model with 93% accuracy for real-time applications.
• Built a pose estimation model with a repetition counter, trained on a dataset of 10,000 exercise photos, achieving 95% accuracy.
• Designed a product flow of posture detection in the app with 3 different scoring metrics based on yoga performed.

Projects

AutoVoyce – AI-Powered YouTube Knowledge Base GitHub Repo — Live Demo

– AI video research assistant leveraging a Python-based RAG pipeline (LangChain + Pinecone) to concurrently index 50+ YouTube transcripts for sub-second semantic search and context-aware Q n A, delivered via a scalable Next.js + FastAPI full-stack with real-time speech-to-text, text-to-speech (ElevenLabs), and a custom transcript ingestion microservice for reliable, API-independent video analysis.

Convo - AI Powerd Chat Application GitHub Repo — Live Demo

– A modern real-time chat application with integrated AI assistant, with JWT auth, using MERN stack. Deployed on Render with responsive UI (Tailwind) and MongoDB Atlas integration.

Education

2023 - 2025 M.tech (Data Science and Analytics) at **MIT WPU, Pune** (GPA: 8.3)
2019 - 2023 B.tech (Computer Engineering) at **DPCOE, Pune** (GPA: 8.0)

Skills

Languages	Python, Java, C++, R
Development	MERN Stack
ML/AI	LLMs, RAG, Multi-Agent AI, LangChain, LangGraph, Prompt Engineering, NLP, Semantic Search, FastMCP, Supervised/Unsupervised ML, MediaPipe, Data Engineering.
Others	Docker, AWS, Redis.