Sanjay Makwana
AI Developer
+916351336586 | makwanasanjaylm@gmail.com | BHAVNAGAR, Gujarat, 364002, India | LinkedIn

SUMMARY

Innovative Generative AI Developer with 2+ years of experience in designing and deploying scalable AI solutions using LLMs, RAG pipelines, and multimodal models. Proven expertise in OpenAI, Gemini, Hugging Face, and LangChain, with a strong focus on NLP, custom prompt engineering, and end-to-end AI architecture. Adept at building full-stack AI applications (Streamlit, FastAPI) and integrating vector databases (Qdrant, Pinecone) to deliver measurable business impact. Passionate about leveraging state-of-the-art GenAI to solve complex problems and drive automation.

Key Strengths:

- LLM Deployment: Fine-tuning, multi-agent workflows (CrewAI), and hybrid-search RAG systems.
- Multimodal AI: CV parsing, document intelligence, and conversational agents.
- Full-Stack Integration: API development (FastAPI/Flask), cloud deployment (AWS), and UI frameworks (Streamlit/React).
- Metrics-Driven: Consistently delivered solutions with >80% accuracy, 40% efficiency gains, and 30% user engagement boosts.

WORK EXPERIENCE

Software Engineer, Tatvasoft, Ahmedabad, India **July 2023 – Present**

- RAG-Based Financial Assistant: Built a hybrid-search chatbot (Gemini + AWS) with 85% accuracy for financial document Q&A, leveraging Qdrant vector DB and semantic chunking.
- Multi-LLM Support Bot: Developed a LangChain + CrewAI customer support agent with dynamic routing across GPT-4/Gemini, reducing response time by 40%.
- PDF Chatbot: Engineered a hybrid-search PDF analyzer (Gemini Pro + Qdrant) improving retrieval accuracy by 80% via dense/sparse embeddings.
- CV Parsing Pipeline: Automated extraction (OpenCV + PyTorch) for 10K+ CVs/month, cutting manual effort by 65%.
- Data Scraping: Accelerated dataset collection by 60% using Scrapy and LLM-based cleaning.
- Deployment Leadership: Spearheaded end-to-end deployment of AI apps boosting engagement by 30%

Trainee Software Engineer, Tatvasoft, Ahmedabad, India **Jan 2023 - Jun 2023**

- Specialized in LLMs (GPT-4, Gemini, Claude), prompt engineering, and RAG pipelines.
- Designed multi-agent workflows (LangChain, CrewAI) and task-specific GPTs for automation.
- Deployed lightweight customer support bots (OpenAI API) and internal tools (FastAPI/Streamlit) with JWT authentication.

SKILLS

Languages: Python / JavaScript
AI & ML Frameworks: TensorFlow / PyTorch / Hugging Face / Ollama, LLaVA / AWS Bedrock
LLM Tooling: LangChain / LlamaIndex / CrewAI / OpenAI API / Gemini API
Web & Backend: Streamlit / FastAPI / Flask / ReactJS / Flask-SocketIO
Vector DBs & Search: FAISS / Qdrant / Pinecone / ChromaDB / Weaviate / ElasticSearch
Cloud & DevOps: AWS / Git / GitHub
NLP Libraries: NLTK / spaCy / Scikit-learn / Pandas / NumPy / Matplotlib / OpenCV
APIs & Integrations: RESTful APIs / Webhooks / WebSockets / Third-party APIs

EDUCATION

Bachelor of Engineering (B.E.) in Information Technology **July 2019 – May 2023**
Government Engineering College, Bhavnagar, India | **CGPA: 8.24**

High School Diploma (HSC), State Board **June 2018 – May 2019**
Shree Gyan Guru Vidhyapith, Bhavnagar, India | **Percentage: 73.5%**

PROJECTS

AI Chatbot for Customer Support

- **Tech Stack:** LangChain / OpenAI / Gemini / FAISS / MongoDB
- **Description:** Created a chatbot using Retrieval-Augmented Generation (RAG) that provides accurate, real-time responses from company documents. Integrated document ingestion, vector search, and support for uploading new files for ongoing learning.

AI Art Generator (Generative Art | Stable Diffusion)

- **Tech Stack:** Stable Diffusion v1.5 / Diffusers / PyTorch / Transformers / FastAPI
- **Description:** Developed a web-based generative art tool using Stable Diffusion v1.5. Integrated prompt engineering with live previews and style selectors. Used FastAPI for image generation endpoints, MongoDB to store prompt/image metadata, and local/AWS S3 for image storage. Enabled users to create surreal, anime, and photorealistic art directly from textual prompts.

Facial Colour Detection System

- **Tech Stack:** OpenCV / MTCNN / NumPy / Pandas / KMeans / FastApi / OpenAI
- **Description:** Developed a Seasonal Color Analysis system using facial recognition and computer vision techniques. Leveraged OpenCV and MTCNN for high-accuracy detection of skin, hair, and eye colors. The system maps detected tones to seasonal color palettes (Winter, Spring, Summer, Autumn) and suggests suitable fashion choices.

Smart Voice Assistant

- **Tech Stack:** Flask / LangChain / Gemini API / FAISS
- **Description:** Built a voice-controlled AI assistant using LLM (GPT), speech recognition (Whisper/STT), and TTS for natural interactions. Enabled voice commands for reminders, weather, emails, and Q&A via API integrations (Open Weather, Gmail). Achieved 90%+ accuracy in command parsing.

AI Writing Assistant

- **Tech Stack:** Flask / Gemini API / OpenAI / FAISS
- **Description:** Built a real-time writing assistant for grammar and spelling correction, writing style suggestions, readability scoring, and optional explanations to help users understand their mistakes. Custom prompts allowed tone and style adjustments, enhancing user experience with NLP and GenAI techniques.

AI-Powered Course Creator

- **Tech Stack:** React js / Gemini API / PDF parsing
- **Description:** Developed an AI system that extracts structure and content from academic PDFs to auto-generate course outlines, lessons, and quizzes. Used Multi-Chain Prompting for higher accuracy and allowed users to manually refine generated content.