Vibhav Jaiswal

mail.vibhav@gmail.com (91) 8130892883, (91) 9873517170 in/vibhavjaiswal github.com/VibhavJaiswal

SUMMARY

GenAI Engineer with hands-on experience in RAG pipelines, LLM chatbots, Azure AI, and multimodal systems, delivering production-ready solutions using LangChain, OpenAI, FastAPI, and YOLO across real-world projects with scalable deployment and API integration.

SKILLS

- **Programming**: Python, LangChain, OpenAI, Hugging Face, FastAPI, Flask, SQL, REST APIs
- **Cloud**: Azure ML, Azure SDK, Cognitive Services, Bot Framework, Custom Vision
- **Analytics**: ML, NLP, RAG, forecasting, segmentation, multimodal data
- **Development**: LLM chatbots, vector search, RAG pipelines, microservices, YOLO automation
- **Collaboration**: API integration, LLMOps, cross-functional AI projects (Deakin, Purdue, UT Austin)
- **Maintenance**: Deployment, monitoring, scalability, production-ready AI systems

EXPERIENCE

AI projects at Deakin, Purdue, UT Austin + self-initiated real-world

Academic and self-initiated real-world projects via global university programs **December 2021 - Present**

- Executed AI projects in RAG, LLM chatbots, and analytics through academic collaborations with Deakin, Purdue, and UT Austin, alongside self-initiated real-world projects focused on deployment, scalability, and impact.
- **RAG Chatbot Engineering**: Built a production-ready assistant using FastAPI, OpenAI, Sentence-BERT, ChromaDB, and Redis for real-time, vector-based contextual Q&A.
- **LangChain Pipelines**: Developed modular RAG workflows with LangChain and OpenAI, integrating PromptTemplates and retrievers for scalable knowledge retrieval.
- **Azure Cognitive AI**: Deployed Q&A bots using Azure Cognitive Services and REST APIs, aligned with Azure AI Search for enterprise-ready chatbot solutions.
- **Multimodal AI Agents**: Created an intelligent agent using Azure Bot Framework, LUIS, and Vision/Speech APIs for multimodal interaction in real-time support.
- **LLM Microservices**: Designed lightweight Flask APIs with LLaMA and custom retrievers for hybrid RAG integration into backend systems.
- **Full-Stack AI Automation**: Delivered an end-to-end consultation platform with YOLO-based image analysis, OpenAI integration, and calendar-based scheduling.

Deputy Manager

TOSHIBA, Toshiba JSW Power Systems Pvt. Ltd. **November 2012 - February 2021, Gurgaon**

- **Data-Driven Engineering Optimization**: Applied Python, SQL, and predictive analytics to enhance power plant equipment performance, leveraging machine learning for efficiency improvements and operational insights.
- **AI-Driven Predictive Maintenance**: Developed ML-based predictive maintenance models using classification algorithms, feature engineering, and time series analysis to forecast equipment failures and reduce downtime.
- **Industrial Data Analytics with SQL**: Utilized SQL (CTEs, Window Functions, Subqueries) to extract and analyze large-scale operational datasets, optimizing engineering workflows and data-driven decision-making.
- **Time Series Forecasting for Energy Analytics**: Applied ARMA models to forecast energy consumption, integrating Pandas, NumPy, and visualization techniques for trend analysis and process optimization.

Assistant Manager

L&T-Sargent & Lundy Limited **August 2006 - November 2012, Faridabad**

- **Pipe Rack Planning & Visualization:** Implemented data analytics for pipe rack planning and visualization, optimizing engineering workflows.
- **Data-Driven BOQ & Procurement Optimization:** Utilized data-driven techniques to enhance BOQ extraction, vendor drawing evaluation, and procurement processes, improving project coordination and cost efficiency.

EDUCATION

MASTER OF DATA SCIENCE (GLOBAL)
DEAKIN UNIVERSITY • 2024
- Completed in June 2024

Post Graduate Program in Data Science & Business Analytics
The University of Texas at Austin and Great Lakes Executive Learning • 2023
- Completed in January 2023

B.TECH.
Uttar Pradesh Technical University, Lucknow

CERTIFICATIONS

APPLIED GENERATIVE AI SPECIALIZATION
PURDUE UNIVERSITY • 2024
- Completed in September 2024

PROJECTS

AI-Powered Tattoo Booking and Consultation Platform
Language and Frameworks: Python, FastAPI, PyTorch, TensorFlow, YOLO, OpenCV, HTML, CSS, JavaScript, Google Calendar API, SQLite, Rasa, Jinja2, REST API, scikit-learn, NumPy, pandas
- Designed and implemented an AI system to analyze client-submitted tattoo reference images, predicting complexity, size, price, and session duration using deep learning models.
- Developed an end-to-end FastAPI backend integrated with Google Calendar to automate appointment scheduling and manage availability in real-time.
- Built a client-facing interface for uploading images, receiving instant quotes, and booking sessions, ensuring a smooth and user-friendly experience.
- Integrated an AI-powered chatbot to handle client inquiries, send reminders, and assist with bookings, while ensuring data security and compliance throughout the system.

AI-Powered HR Assistant: A Conversational Chatbot for Employee Support
Language and Frameworks: Python, FastAPI, OpenAI API, Sentence-BERT (SBERT), Redis, ChromaDB, RapidFuzz
- Developed an AI-powered HR chatbot using Stack AI to handle employee queries and automate HR responses.
- Used SBERT and fuzzy matching for accurate FAQ retrieval and semantic search.
- Built a FastAPI-based API with Redis caching and ChromaDB for efficient query handling.
- Integrated GPT-3.5 for handling complex queries beyond predefined responses.

Crafting an AI-Powered HR Assistant with LangChain and OpenAI
Language and Frameworks: Python, LangChain, Chroma, OpenAI, PyPDFLoader, OpenAIEmbeddings, RecursiveCharacterTextSplitter, VectorDBQA, RetrievalQA
- Integrated OpenAI's language model with LangChain to develop an AI-powered HR assistant.
- Loaded HR policy documents from PDF files and converted them into vector representations using Chroma and OpenAI embeddings.
- Built a question-answering system using the VectorDBQA and RetrievalQA modules for document-based queries.
- Split long texts efficiently with RecursiveCharacterTextSplitter to optimize the retrieval and processing of information.

Implementing Retrieval-Augmented Generation (RAG) with LangChain and OpenAI
Language and Frameworks: Python, LangChain, OpenAI, PromptTemplate, LLMChain, Retrieval-Augmented Generation (RAG)
- Integrated the OpenAI API with LangChain to build a Retrieval-Augmented Generation (RAG) system.
- Used the OpenAI language model to generate creative and context-specific responses.
- Built custom prompt templates with LangChain to structure queries and responses.
- Implemented LLMChain to manage the text generation flow, enhancing the retrieval and generation process.