Om Bhaltilak

+91-9767910617 | omnbhaltilak@gmail.com | Pune, India
LinkedIn | GitHub | LeetCode | HackerRank

EDUCATION

AISSMS Institute of Information Technology — Pune, India
B.Tech. in Artificial Intelligence & Data Science — 2022 – 2026

- **CGPA**: 7.82

EXPERIENCE

Infosys Springboard — Feb. 2025 – Mar. 2025
Intern – AI-Powered Fraud Management System for UID Aadhaar — Pune, India

- Developed an AI-based fraud detection system for UID Aadhaar document verification using OCR pipelines, classification models, and YOLO-based object detection.
- Implemented advanced image preprocessing techniques to enhance model accuracy on low-quality and noisy inputs.
- Implemented fraud scoring and generated structured Excel/JSON reports to streamline verification workflows and reduce manual errors.
- Utilized Python, TensorFlow/PyTorch, YOLO, OpenCV, Tesseract OCR, Pandas, NumPy, and Flask; proposed enhancements including multilingual OCR and real-time fraud alert integration.

PROJECTS

YouTube RAG Assistant Multilingual Video AI Chatbot | *Python, LangChain, FAISS* — GitHub

- Architected a full-stack AI browser extension and scalable backend allowing users to chat with long-form YouTube videos in real-time using an advanced Retrieval-Augmented Generation (RAG) pipeline.
- Engineered a robust data ingestion pipeline featuring language detection, on-the-fly translation, and sliding-window chunking with semantic overlap to maintain context boundaries across large transcripts.
- Built a two-stage retrieval pipeline using FAISS with all-MiniLM embeddings for high-speed vector search, followed by an MS-MARCO Cross-Encoder for reranking retrieved documents to maximize context precision.
- Implemented Agentic RAG features including LLM-driven query rewriting, guardrailing, and dynamic intent routing.
- Designed a strided document sampling strategy (map-reduce style) to accurately summarize multi-hour videos without exceeding LLM context windows or losing global context.
- Utilized strict prompt engineering with Qwen2.5-7B-Instruct to generate hallucination-free, dynamically sized answers, strictly enforced with multiple inline clickable video timestamps [MM:SS].
- **Tech Stack:** Python, LangChain, FAISS, Hugging Face Inference Endpoints (Qwen 2.5), Sentence-Transformers (Cross-Encoders), Flask, Docker, JS.

Offline Document Finder (ODF) AI Semantic Search | *Python, ChromaDB, ONNX* — GitHub

- Architected and developed a fully offline, privacy-first AI-powered desktop search engine enabling semantic retrieval across local documents (PDF, DOCX, XLSX, TXT) using natural language queries instead of exact keyword matching.
- Designed a high-performance semantic indexing pipeline using FastEmbed with ONNX Runtime, dynamically leveraging NVIDIA CUDA for GPU acceleration while maintaining optimized CPU-only fallback execution.
- Implemented persistent local vector storage using ChromaDB, enabling efficient similarity search with disk-based embedding storage for scalable document indexing.
- Built a lightweight desktop interface using CustomTkinter with global system-wide hotkey support (Ctrl + K) for instant search access.
- Packaged the application into a standalone Windows executable using PyInstaller, ensuring zero external dependencies and 100% local execution.
- **Tech Stack:** Python, ChromaDB, FastEmbed, ONNX Runtime (CPU/GPU), CustomTkinter, PyInstaller, PyMuPDF (Fitz), OpenPyXL

CERTIFICATIONS

OCI 2025 Generative AI Professional | Oracle
The Joy of Computing using Python | NPTEL (IIT Madras)

SKILLS

Languages : Python, C++, JavaScript, HTML/CSS
Generative AI & Frameworks : Agentic RAG, LangChain, Prompt Engineering, FAISS, ChromaDB
AI, ML & Data Science : NLP, Deep Learning, Machine Learning, Pandas, NumPy, Matplotlib, Seaborn
Databases & BI Tools : SQL (MySQL), MongoDB, Power BI, Tableau
Web Development & Core : React.js, Tailwind CSS, Bootstrap, DSA, OOP