

SUMMARY

AI Engineer with hands-on expertise in designing and deploying, production-grade Generative AI solutions across diverse real-world applications. Specialized in building advanced LLM-driven systems, including Retrieval-Augmented Generation (RAG) pipelines, custom prompt engineering strategies, and intelligent automation workflows using tools like n8n and Apache NiFi. Proven track record in orchestrating end-to-end AI workflows from experimentation to deployment with a focus on performance, reliability, and business impact.

TECHNICAL SKILLS

Languages: Python
Frameworks: Flask, FastAPI, REST APIs
AI/ML: RAG, Generative AI, Hugging Face, OpenAI APIs, RAG Pipelines
Vector Databases: Pinecone, ChromaDB, ElasticSearch
Automation Tools: Apache NiFi, n8n
Databases: SQL, MongoDB, ElasticDB

KEY PROJECTS

- Real-Time Earnings Call Intelligence Platform**

OpenAI Whisper, Python, FastAPI, WebSockets, React, AI/ML.

– Built an intelligent system that processes live financial calls in real-time, automatically transcribing audio, segmenting content into logical sections, and extracting key financial insights as calls unfold. The platform enables fund managers to make data-driven investment decisions during earnings calls by providing instant access to critical financial information through a real-time dashboard with live audio visualization, transcriptions, and section timeline navigation.
- AI-Powered Resume Optimization Pipeline**

Apache NiFi, n8n, OpenAI

– -> Identified the inefficiency in manual resume screening, leading to a solution that cut **effort by over 95%**. Designed an automated pipeline using **OpenAI** to parse, score, and personalize resumes for specific **job descriptions**. Orchestrated workflows with **Apache NiFi** and integrated **Microsoft Teams** to enable on-demand resume delivery. Accelerated recruiter decision-making with real-time access to tailored, ranked candidate profiles
- Smart Learning Engine with RAG and Vector Search**

Python, Pinecone, Langchain, FastAPI

– Tackled poor content discoverability by designing a **smart retrieval** system for educational videos and articles. Mapped user queries to relevant syllabus topics using **semantic similarity** and structured topic buckets. Used Pinecone to perform high-speed **vector search** across multimodal embeddings from both video transcripts and article summaries. Delivered the single most relevant video and article per topic, enhancing user satisfaction and platform engagement.
- LLM Powered Sales Assistant with RAG and Voice Integration**

OpenAI Whisper

– Built a real-time, voice-enabled AI assistant for sales teams using RAG architecture and OpenAI Whisper. When a customer asks product-specific queries, the assistant instantly fetches answers from internal product manuals via ChromaDB vector search. This empowers sales with accurate, context-aware responses—reducing customer query handling time and boosting confidence during live interactions.

WORK EXPERIENCE

- ShorthillsAI**

Software Engineer Intern

Feb 2025 – Present

Gurugram, Haryana

– Automated organizational workflows using Apache NiFi & n8n, reducing manual intervention by 90% and boosting operational efficiency.
- ProdigalAI**

AI Project Intern

Sep 2024 – Nov 2024

Remote

– Developed a domain-specific chatbot using LlamaIndex and Gemini embeddings to solve unstructured document query retrieval through RAG pipelines
- Deutsche Telekom Digital Labs**

Data Scientist Intern

Jun 2024 – Aug 2024

Gurugram, Haryana

– Engineered a non-LLM document parser achieving >95% extraction coverage from HR documents and built real-time PDF analytics dashboard using Streamlit.

– Contributed to the development of a comprehensive evaluation dataset for AWS Bedrock model assessment, creating structured test cases and performance metrics to evaluate model accuracy, response quality, and domain-specific understanding for internal HR chatbot applications.

EDUCATION

–Bachelor of Technology in Artificial Intelligence and Machine Learning

Guru Gobind Singh Indraprastha University, New Delhi

2021-25

CGPA: 8.4

CERTIFICATIONS & ACHIEVEMENTS

GATE DA 2025 (AIR 663) | Diploma in Programming & Data Science - IIT Madras (2024)

Postman API Fundamental Student Expert (2023)