

Yash Raj

+91 9555866452 | @ yashraj25118110@gmail.com | LinkedIn | GitHub

PROFESSIONAL SUMMARY

AI Engineer with 2.8 years of experience building and shipping end to end Generative AI, Agentic AI, and Machine Learning systems. Comfortable owning the full stack from data pipelines and the model layer to APIs, deployment, and frontend integration. Skilled at building scalable AI agents, RAG systems, and chatbots using Python, FastAPI, LangChain, LangGraph, and LLMs, with a track record of boosting operational efficiency by up to 90%. Hands on with cloud and MLOps tooling across GCP, AWS, Azure, Docker, Kubernetes, and GitLab CI/CD. Driven by shipping production AI products that solve real business problems.

WORK EXPERIENCE

Optimal Virtual Employee (OVE)

Noida, India

AI Engineer

Nov 2025 – Present

- Built **AIRA**, a production grade real time voice AI interview platform using **LiveKit WebRTC**, **Deepgram STT**, **ElevenLabs TTS**, and **Groq LLaMA 3.3 70B**, replacing manual HR screening for clients including **City Home** (Florida's top furniture retailer) and **GUS** (global group of schools).
- Designed and implemented **company type personalization** with dynamic opening, wrap up, and closing scripts per client (HR, recruitment, admissions) and correct TTS pronunciation, deployed via **FastAPI** backend and **GitLab CI/CD** to production.
- Engineered **candidate proctoring** on the React/Vite frontend using **BlazeFace** (TensorFlow.js) for real time face detection, tab switch monitoring, and violation based interview termination.
- Developed automated AI driven **interview reports** with 5 dimension scoring (skills, communication, problem solving, culture fit, role match), persisted to MySQL and exposed via REST APIs.
- Developed a **Knowledge Graph PoC** for multimodal archival data (scanned PDFs, images, newspapers) using **OCR**, **Computer Vision**, and **Multimodal LLMs** (Gemini 2.0 Flash), extracting structured entities and relationships with **NetworkX** and **PyVis**, generating outputs for **Neo4j** ingestion with 85 to 94% extraction accuracy.

AGL-Hakuhodo

Gurugram, India

Data Science Analyst

Oct 2023 – Nov 2025

- Developed a **Generative AI chatbot** using **OpenAI** and **Groq**, improving SEO team efficiency by **50%** through customizable content generation.
- Implemented **RAG pipelines** over 50+ client files using **OpenAI embeddings** and **All-MiniLM-L6-v2**, boosting information extraction efficiency by **90%**.
- Enabled natural language data insights and chart generation, reducing dependency on **Power BI** using **Pandas AI** and **LIDA**.
- Built an **AI Agent** using **LangChain** and **LangGraph** for domain specific retrieval and intelligent tool execution.
- Developed predictive models (regression, churn) using **Scikit-learn** and **XGBoost**, achieving **85 to 88% accuracy and precision**.
- Deployed a **Llama 2** powered chatbot on **GCP Cloud Run**, reducing client query response time by **60%**.
- Built an Agentic HR Virtual Assistant with RAG and JWT based multi tenant authentication, cutting HR workload by 50% and achieving 96% query resolution accuracy.

PROJECTS

AIRA, AI Real Time Interview Agent

Nov 2025 – Present

- Built a production grade **real time voice interview platform** using **LiveKit WebRTC**, **Deepgram STT**, **ElevenLabs TTS**, and **Groq LLaMA 3.3 70B**, replacing manual HR screening for clients including **City Home** and **GUS** (global group of schools).
- Engineered **multi client personalization** with dynamic opening and closing scripts per **company_type**, **candidate proctoring** via **BlazeFace** (TensorFlow.js) for real time face detection and tab switch violation tracking, and **auto generated interview reports** with 5 dimension scoring saved to MySQL.

- Deployed end to end on **Azure VM** with **FastAPI** backend, **React/Vite** frontend, and **GitLab CI/CD** pipeline. Integrated with a **Laravel** admissions portal for session creation and candidate management.

Multimodal Archive to Knowledge Graph Builder

Nov 2025 – Present

- Developed an AI driven pipeline that processes multimodal archive data (images, scanned PDFs, newspapers, documents) using **OCR, Computer Vision, and Multimodal LLMs** (Gemini 2.0 Flash) to extract structured entities and relationships.
- Constructed a full **Knowledge Graph** using **NetworkX** and **PyVis**, generating **interactive visualizations** and structured outputs (JSON, CSV, GraphML) for graph database ingestion into **Neo4j**, achieving 85 to 94% entity extraction accuracy.

Agentic HR Assistant

Nov 2024 – May 2025

- Built a **Flask API** based chatbot for a major Japanese client, leveraging **GenAI RAG architecture** for multi company data integration. Access secured by **JWT**, restricting data to the corresponding organization.
- Enabled real time HR functionalities (leave applications, attendance reports, policy queries) using regex driven intent matching in Python, increasing HR operational efficiency by **70%** and providing instant responses to common questions.

Insight Generator

Aug 2024 – May 2025

- Developed an end to end **Streamlit ChatBot** for a client's 50+ data files. Integrated files using **OpenAI embeddings** and the **All-MiniLM-L6-v2 model**, enabling the chatbot to generate statistical insights and replace Excel functionalities.
- The interactive UI allows the client to query insights, select content length, and choose languages, drastically improving efficiency and ease of use for data analysis and reporting.

SKILLS

Programming Languages: Python, React, NodeJS, SQL, C/C++

Machine Learning: Regression, Classification, Clustering, Ensemble Methods, Feature Engineering, MLOps

Deep Learning: CNN, RNN, LSTM, Transformers, BERT, Computer Vision (YOLO, OpenCV)

Generative AI: LLMs, LangChain, LangGraph, RAG, AI Agents, Agentic AI, Vector Embeddings, Prompt Engineering, Fine Tuning, Knowledge Graphs

Frameworks & Libraries: FastAPI, Flask, Streamlit, TensorFlow, PyTorch, Scikit-learn, XGBoost, Pandas, NumPy, Matplotlib, Seaborn, LIDA, Pandas AI, Hugging Face, FAISS, NetworkX

Tools & Platforms: Git, GitLab CI/CD, Docker, Kubernetes, MySQL, MongoDB, Neo4j, PySpark, Hive, Power BI, n8n

Cloud Technologies: AWS (S3, EC2, SageMaker, Lambda), GCP (Cloud Run, Vertex AI, Dialogflow CX), Azure (VM, App Service)

LLMs & APIs: OpenAI, Groq, LLaMA, Gemini, Claude, Hugging Face

Voice AI & Real Time: LiveKit WebRTC, Deepgram, ElevenLabs, Cartesia, TTS, STT

EDUCATION

KIET Group of Institutions (AKTU)

Delhi, NCR

B-Tech in Information Technology; CGPA: 7.6/10

Oct 2020 – Jun 2024

Modern National Public School (CBSE)

Delhi, NCR

12th Board; CGPA: 7.39/10

Mar 2019 – Apr 2020

Modern National Public School (CBSE)

Delhi, NCR

10th Board; CGPA: 8.6/10

Mar 2017 – Apr 2018

CO-CURRICULAR AND EXTRACURRICULAR ACTIVITIES

Core Member (TEDxKIET)

Oct 2022 – Dec 2022

- Volunteered for the inaugural TEDxKIET Women event
- Served as Sponsorship Head for the main TEDxKIET event, "Living the Wonka World"

Campus Ambassador (Techkriti IITK)

Sep 2022 – May 2023

- Participated as a Campus Ambassador Intern, leveraging event management and marketing strategies to promote the event.

DSA and Coding

- LeetCode | HackerRank

Certifications

- | | |
|------------------------------------|---------------------|
| • C++ (Ducat) | Apr 2021 – Aug 2021 |
| • AI/ML (AWS Academy) | May 2023 – Jul 2023 |
| • Data Analysis using Python (IBM) | Feb 2024 – Apr 2024 |