Annu Ahlawat

8448800862 | aannu1302@gmail.com | linkedin.com/in/ja

Education

Maharshi Dayanand University — Haryana
Bachelors of Technology in Computer Science — *October 2020 - May 2024*

Experience

Junior Executive – Generative AI — June 2025 – Oct 2025
Creative Thinks Media Pvt. Ltd. — *Noida, Uttar Pradesh*

- Delivered end-to-end GenAI solutions using FastAPI, Python, and OpenAI LLMs for production-ready deployments.
- Built an LLM-powered RAG chatbot (FastAPI + Pinecone) and a contextual property search engine (FastAPI + Pinecone), enhancing retrieval accuracy and user experience.
- Deployed and optimized both systems on AWS with secure APIs and efficient inference pipelines.

Computer Vision Engineer — Nov 2024 – May 2025
ORBO.AI — *Noida, UP*

- Built a contextual makeup search engine using LLMs and semantic retrieval, improving search accuracy by 35%.
- Integrated Google SERP API for real-time product discovery, reducing data latency by 40%.
- Implemented AI ranking with Milvus/Weaviate vector DBs, enhancing personalized recommendations and engagement.

Artificial Intelligence Developer — Jan 2024 – Oct 2024
Illuminate Technologies — *Gurugram, Haryana*

- Trained Stable Diffusion and developed RAG systems using OpenAI/Hugging Face, improving retrieval accuracy by 30%.
- Executed ML/DL research including K-means clustering on protein datasets with 80% enzyme group accuracy.
- Built and deployed LLM-powered real estate search and RAG applications on AWS, ensuring scalable performance.

Projects

Application Tracking System | *Gemini Vision Model, Full-stack Development*

- Developed a full-stack ATS with Gemini Vision for automated resume parsing and evaluation, reducing manual review time by 60%.
- Designed an end-to-end workflow with API gateway, asynchronous processing, and secure data pipelines, reducing manual review time by 60%.

LLM-RAG Assistant | *Python, Openai, Web Scrapping, Pinecone, AWS*

- Created an AI assistant for a customer complaints and process firm using RAG architecture and vector search, improving search precision by 45%
- Built scalable FastAPI microservices deployed on AWS with orchestrated ingestion, vector indexing, and query pipelines, improving search precision by 45%.

Technical Skills

Languages: Python
LLMs: OpenAI, OLLAMA, Claude, Gemini
Libraries: Hugging Face, TensorFlow, PyTorch, Keras, Scikit-learn, NLTK
Database: Firebase, NoSQL MongoDB, MySQL
Vector Database: Milvus, FAISS, Weaviate, Chroma, Pinecone
Grapgh Database: Neo4j Aura DB
Frameworks: React, FastAPI, LangChain, Re-rankers Cohere, LlamaIndex,
Models: Stable Diffusion, GAN, VAE, Transformers, BERT, CNN Models
Automation Tools: n8n
Developer Tools: Git, Docker, VS Code, AWS
Agentic AI : CrewAI, LangGraph, MCP