Amit Kumar

iamitkumar2007@gmail.com | +919889460738 | linkedIn | Github

Professional Summary

Full-Stack AI Developer specializing in production-grade LLM applications, including scalable RAG pipelines and multi-step AI workflows built with LangChain and LangGraph. Skilled in developing end-to-end AI systems using FastAPI, vector search, pre-trained models, and cloud-deployed, scalable architectures.

Education

M.tech in Artificial Intelligence Aug 2023 – July 2025
Maulana Azad National Institute of Technology , Bhopal , Madhya Pradesh

B.tech in Computer Science and Engineering Aug 2018 – July 2022
Institute of Engineering and Technology, Bundelkhand University, Jhansi , Uttar Pradesh

Internship Experience

Full Stack Developer Intern, *Wabtec Corporation, Bengaluru* Sep 2024 – Aug 2025

- Built asset management web-appllication using ReactJS, FastAPI, and PostgreSQL for 200+ users.
- Automated workflows with Dockerized microservices, improving uptime by 15%.
- Integrated SSO login, SSL certificate , email alerts, and AWS cloud deployment for scalable operations.

Projects

AI-Powered Customer Support Agent July 2025 **-** August 2025

- Built LangGraph and Langchain based multi-agent system with groq LLM for FAQs, Order related queries to DB, and Email automation.
- Automated 65% of customer queries, reducing escalations from 70% to 25% with <2s response.
- Deployed via FastAPI + Docker microservice integrated with cloud databases.

NL2SQL Chatbot April 2025 - June 2025

- Natural language to SQL system using Streamlit, Python, and SQLite
- Converted 100+ text prompts into optimized SQL queries with 80% accuracy improvement.

Semantic Search Engine (RAG-based) Jan 2025 - Feb 2025

- Vector-based document search using FAISS, Sentence Transformers, and Groq LLM.
- Achieved 95% retrieval accuracy across 200+ documents, 3× faster results.

YouTube video chatbot Dec 2025 - Jan 2026

- Built a YouTube-based RAG chatbot using LangChain and Streamlit, enabling real-time Q&A over video transcripts with cached vector retrieval for low-latency responses.
- Designed a UI to improve user experiance with session-aware chat management, automatic context reset on video change, and optimized LLM pipeline reuse.

Research Publication

Optimising Fault Tolerance and Latency of Federated Learning using Edge Servers and Pre-trained Model,
Published under IEEE Xplore Sep - 2025

Technical Skills

- **Languages :** Python
- **AI / ML :** CNN, RNN, NLP, Transfer Learning, LLMs, Federated Learning, GenAI
- **Frameworks & Libraries:** LangChain, LangGraph, FastAPI, Streamlit, ReactJS, TensorFlow, Scikit-learn, Pandas, NumPy, Matplotlib
- **Databases :** PostgreSQL, Chroma (Vector DB), FAISS (Vector DB).
- **Cloud & Devops :** AWS (S3, Lambda, EC2, Bedrock, CloudWatch, Cost Explorer, Vector DB), Docker , Gitlab CI/CD, Github.