

Rohit Kumar

✉ rohitcode005@gmail.com ☎ +91-9798494187 📁 portfolio in linkedin 🐙 github 🏆 kaggle

Technologies

- Programming and Query Languages: Python, SQL, JavaScript
- Machine Learning and Deep Learning: TensorFlow, Pytorch, OpenCV, CNN, Transformer, LoRA, PEFT, Llmgraph
- Big Data and Frameworks: PySpark, Databricks, Flask, Django, FastAPI
- Cloud and Tools: AWS lambda, AWS S3, Azure ML, Docker, Git, MLflow

Education

Sarala Birla University

B.Tech in AI (Artificial Intelligence)

Nov 2021 – May 2025

Experience

AI Engineer

peptidewebmd (qubecare.ai)

Bhopal, India

may 2025 – present 2025

- Designed and deployed AI-driven automation agents for social and workflow platforms using Python, LangChain, and n8n, enhancing real-time decision-making and adaptive user interaction.
- Developed a plan-execute multi-agent architecture leveraging LangGraph and LLM-based reasoning, integrating voice-capable agents through LiveKit for natural conversational interfaces.

Data Extraction Engineer

Relu Consultancy

Bhopal, India

Feb 2025 – may 2025

- Built end-to-end data ingestion and retrieval systems using Spring Boot APIs, MySQL, and ReactJS extensions to handle over 300K+ unstructured and financial records efficiently.
- Automated large-scale data scraping and integration pipelines, ensuring structured access to web and external data sources for AI model training and evaluation.
- Contributed to backend automation, scheduling, and performance monitoring for data-driven applications.

Data Scientist Intern

Enshrine Global Systems (Gradstem)

Oakland County, USA

June 2024 – April 2025

- Engineered retrieval-augmented generation (RAG) systems using LangChain, FastAPI, and OpenAI function calling for contextual reasoning and document analysis.
- Built scalable ML pipelines on AWS Lambda, automating resume parsing and document summarization workflows for data-driven recruitment.
- Integrated structured and unstructured data into knowledge-based AI systems to enable dynamic contextual responses and domain-specific insights.
- Collaborated with cross-functional teams to deploy, evaluate, and document AI solutions for production use.

Prompt Engineer outlier.ai

Remote Sept 2023 – Feb 2024

- Review LLM model code generation Data science and ML training expert improved prompt generative AI model for better performance worked on google (bard i.e gemini) to review and train the ML model prompts

Data Engineer Intern Taiyo.ai

USA (Remote) Mar – Sept 2021

- Designed a Big Data pipeline using PySpark on Databricks and Python to process World Bank datasets, integrating automated web scraping with Selenium to standardize multilingual data for ML models.

Projects

LangGraph: Multi-Agent Collaboration and Tool Calling

Technologies: LangChain, Llmgraph

- Developed a multi-agent collaboration framework using LangGraph, enabling tool calling and seamless interaction between a Research Agent and a Python Code Executor to dynamically generate matplotlib charts from queries.
- Integrated Instagram API and email workflows into the system, enabling data retrieval, file writing, document review, email automation, and report generation through specialized agents.
- Designed and implemented a Supervisor Agent from scratch, orchestrating multiple agents, tools, and layered super-

visors to ensure scalable collaboration and coordinated execution.

Smart Procurement & Sentiment Analysis (CI/CD pipeline)

[Live](#) · [GitHub](#) · [Video](#)

Technologies: LangChain, GPT-4o, LLaMA, Azure, Vector Embeddings, Supabase, Azure, Firebase

- Built and deployed an end-to-end platform that scrapes MENA-region news to extract tenders, company mentions, and sentiment for procurement teams and analysts.
- Used LangChain pipelines and 1536-dimension vector embeddings for semantic search and standardized entity recognition (e.g., linking “Facebook” to “Meta”).
- Integrated real-time trend dashboards and filtering for companies, countries, and commodities based on public sentiment and procurement signals.

Built LLM from Scratch: GPT-2 Transformer Model

[GitHub](#)

Technologies: Python

- Built a GPT-2 based LLM in PyTorch with transformer blocks, multi-head self-attention, token/positional embeddings, and GELU-activated feed-forward layers.
- Used byte-pair encoding (5104 tokens) and a custom TikToken-based dataloader for efficient training.
- Pretrained for 10 epochs with loss-perplexity monitoring and performed instruction/domain-specific fine-tuning.

CCTV Video Violence Detection (CNN, Transformer, OpenCV, Pytorch)

[GitHub](#)

- Developed a deep learning system for real-time violence detection in CCTV feeds using CNN, ViT, and RNN/Transformer models, achieving 97% accuracy.
- Built a full pipeline with video preprocessing (OpenCV, NumPy), Flask-based deployment, and AWS hosting; generated timestamps and violence GIFs for easy review.
- Applied MLflow for tracking metrics and DVC for dataset/model versioning, showcasing strong computer vision

Fine-Tuning FLAN-T5 for Less-Toxic Text Summarization (Flan-T5, python)

[GitHub](#)

- Fine-tuned FLAN-T5 with PPO and PEFT to create less-toxic summaries, using the DialogSum dataset and RoBERTa for toxicity evaluation. Achieved an 11.14% reduction in mean toxicity scores, leveraging Hugging Face Transformers and TRL libraries.

MLflow - Data Science and ML MLOps CI/CD

[Live](#) · [GitHub](#)

- Developed a full-stack web platform using ReactJS, Flask, and TailwindCSS that allows data scientists to perform data preprocessing, feature selection, and ML model training without writing code
- Enabled end-to-end workflows with automated code generation and support for 17 regression and 20 classification algorithms (e.g., XGBoost, SVM, Random Forest)
- **Tools:** ReactJS, Flask API, Pandas, NumPy, Scikit-learn, JavaScript, MLflow

TalkKAI – RAG-based Conversational AI Platform

[live](#)

Technologies: Python, LangChain, FastAPI, FAISS, reactjs, Azure

- Built a conversational AI platform using LangChain and FastAPI that allows users to interact with PDF documents

AgenticHR - AI Agent (Python, Solidity, AI Agent, Ethereum)

- Built an AI-powered HR system with automated candidate extraction, summarization, and formatting. Created a custom vector DB and fine-tuned “all-MiniLM-L6-v2” on resume-job datasets. Used a two-step selection: vector similarity search followed by an AI agent for top candidate ranking.

Achievements

- [🏆 Samsung Solve for Tomorrow Top 30 Winner](#) - Ranked in the Top 30 of 70,000+ participants nationwide developed **Laper App**, an innovative mobile application that provides instant solutions for developers.
- [🏆 Winner, Anveshana Synopsys, Delhi \(02/2024\)](#) - Engineered an IoT project using ESP Cam, Arduino and ML to prevent train accidents with obstacle and drowsiness detection, winning cash prizes and goodies.
- **Amazon ML Challenge (09/2024)** - Secured All India Rank (AIR) 87 out of 75000+ Participants in Amazon ML challenge