ANSHUMAN KUMAR

Bengaluru · anshumankumar.mail@gmail.com · 9818164505 ·
https://www.anshumankumar.dev/

Work Experience

Deloitte (Bangalore)

AI Consultant December 2023 - Present

- Architected a Generative AI-based Conversation Agent with RAG System and Text-to-SQL Capabilities for an OEM manufacturer, leveraging open-source such as Large Language Models such as Mistral 8x7B, Codestral 22B, LLavaNext 34B.
- Implemented a Retrieval-Augmented Generation (RAG) system using PGVector as the vector database. This system addressed questions based on images, text from PDFs, and other unstructured documents containing solution architectures and manuals for the PLM (Product Lifecycle Management) Application.
- The solution significantly streamlined operations, boosting productivity by enabling quick and accurate access to critical data insights. It potentially reduced document retrieval time by 30% and lowered employee cognitive load while improving query response times by 20-25%.
- Took ownership of testing and deploying AI solutions on Kubernetes on BareMetal using Nginx, ensuring robust and scalable infrastructure. Set up CI/CD pipelines using Gitlab CI.
- Business Rule Change Simulation on Supply Chain: Developed a solution to simulate the impact of changing business rules on outbound shipments for the OEM manufacturer, where I worked on utilizing Feast for feature management during ML model inferencing, ensuring streamlined feature retrieval processes.
- Trained ML models to predict shipment cycle times and classify shipment status (late or on-time). Achieved a prediction accuracy of 95%.
- Created a chat bot to provide step by step assistance to fibre mechanics working on the field using Azure AI Search to answer questions using KB articles and manuals Azure Open AI, performing RAG (Retrieval Augmentation Generation).

Quantiphi (Bengaluru)

Senior Conversation Bot Engineer November 2020 - December 2023

- Designed fault-tolerant architectures with GCP for real-time client dashboards, enhancing customer satisfaction insights and agent efficiency.
- Developed ML-driven chat/voice bots for government and private sector, streamlining user queries and improving service accessibility.
- Created COVID-19 unemployment insurance chatbot, enabling seamless data retrieval and support for customers during a critical period.
- Integrated on-premises APIs to fetch user details, enhancing data accuracy and service reliability for clients.

Certifications

GCP Professional Cloud Developer **December 2022-2024**

AWS Certified Developer – Associate **November 2024-2027**

SKILLS

Languages: Python, JavaScript

ML/AI: Prompt Engineering, Scikit Learn, Agent Developement, Keras, Vertex AI, Bedrock

Cloud: AWS, GCP

CI/CD: Gitlab, Github Actions, Kubernetes, PCF

Bot Development: Dialogflow, RASA

Databases: Postgres, MySQL, Firestore, Document DB

EDUCATION

Manipal Institute of Technology Manipal B. Tech Computer Science and Engineering *GPA: 7.35* 2016 - 2020

The Mother's International School New Delhi CBSE *GPA: 93.6%*

ACHIEVEMENTS & EXTRA CURRICULARS

1) Contribute to open-source applications such as writing a solution to automate various coding tasks like unit tests, code fixes, searching for solutions on Stack Overflow while also offering Code Completion. This won the third prize at the Jovian Hackathon 2023, which was held over the course of 24 hours in Bangalore.

2) Fitness: Love cycling and taking part in marathons (took part during Vedanta Delhi Half Marathon in 2022, Bangalore Half Marathon in 2023, a 200K cycling event amongst other things.