SARABJOT SINGH

Chennai, Tamil Nadu • +91 6200342799 • sarabjot.singh799@gmail.com
• in/sarabjot-singh-93256b22b • github/ricky-aufvaa

SUMMARY

Generative AI Engineer with over 1 year of experience in the Data and Analytics domain, specializing in building, fine-tuning, and deploying Large Language Models (LLMs) for domain-specific use cases. UI frameworks including Streamlit and Chainlit, alongside integration with Flask. Recognized for problem-solving, algorithm development, and code optimization in complex projects.

WORK EXPERIENCE

Ashok Leyland

June 2023 - Present
Chennai, Tamil Nadu

- **Developed and Deployed In-house Generative AI Platform:** Created a comprehensive platform with three applications tailored for Engineers, Designers, and Functional Heads using tools like Python, Langchain, LLamaindex and AWS.
- **Fine-tuned Llama-3 Model:** Achieved significant improvements in retrieval-augmented generation (RAG) performance by fine-tuning the Llama-3 8B model with automotive- specific data.
- **Experience with Multiple Large Language Models:** Worked with various LLMs including T5, FlanT5, Cohere, GPT-3.5, GPT-4, and LLM Sherpa for diverse use cases in Generative AI.
- **Built AI-Powered Sales Platform:** Developed an LLM-driven application to assist sales personnel, enhancing efficiency with Generative AI, leveraging OCR, vector databases, embedding models, Neo4j, OpenAI API, and LangChain.
- **Developed Custom OCR Pipelines:** Built and optimized OCR workflows to extract, refine, and structure text from scanned documents, enabling seamless AI-driven insights.
- **Tools used:** Python, SQL, Langchain, Llamaindex, Streamlit, AWS, ChromaDB, Flask, OpenAI, LLMSherpa, Bedrock LLMs, Marker, Unsloth.

PROJECTS

Document Q & A - RAG

March 2024 - March 2025

- **Dynamic Document Assistant Development:** Created a dynamic document assistant using the Retrieval-Augmented Generation (RAG) approach, enabling users to query uploaded documents effectively.
- **Technologies Utilized:** Python, Langchain, LlamaIndex, Streamlit, Flask, OpenAI GPT-4, Anthropic LLIVIs, LLMSherpa, ChromaDB, AWS.
- **Advanced Query Capabilities:** Implemented LlamaIndex Agents to facilitate responses to various question types, including direct inquiries, comparative analyses, reasoning questions, and cross-references.
- **User Interface Design:** Designed an interactive user interface with Streamlit that enables document uploads and query submissions.

Failure Analysis Assistant

April 2024 -March 2025

- **Overview:** Developed a RAG based Component Failure Service Assistant to support service engineers in analyzing and resolving vehicle component failures, utilizing advanced generative AI technologies.
- **Technologies Utilized:** Python, Langchain, LlamaIndex, Streamlit, Flask, OpenAI GPT-4, Anthropic LLMs, LLMSherpa, ChromaDB, AWS, Knowledge graphs, Marker.
- **Text extraction:** Extracted and indexed data from technical PDFs (including text, tables, and images) using Marker to create structured metadata for efficient querying and retrieval.
- **Interactive UI:** Designed an interactive user interface using Streamlit for seamless interaction with the model, enhancing user experience and engagement with the system.

EDUCATION

August 2020 - November 2024

B.Tech. - Electronics and Communications Engineering
Chandigarh Engineering College, Landran, Punjab

- Lead Google Developer Student's club.
- Was the head of the Electronics and Communication Engineering Club.

Guru Nanak Higher Secondary School, Doranda, Ranchi *2019*

SKILLS

- Pandas, Numpy, Langchain, Llamaindex, Mysql, Chroma, AWS : Ec2, S3, Athena, Sagemaker, Azure: Data factory, Data bricks, Pyspark, Hadoop, Hive, Keras, Tensorflow, Streamlit, Flask, Generative AI, NLP
- Linux administration, Linux shell scripting, Web Scraping & Data Extraction, Fast API, Flask