Hitesh Sudam Patil

+1(716)936-4858 | hiteshsu@buffalo.edu | linkedin.com/in/hitesh-sudam-patil/ | Github | Website

EXPERIENCE

CIPIO Inc. Jan. 2024 – Present
Machine Learning Engineer | Generative AI Team McLean, VA

Multi-Modal Content Retrieval System

- Spearheaded the development of scalable multi-modal content retrieval system, ensuring highly accurate search results from configurable knowledge bases such as AWS S3 buckets
- Addressed the issue of keyframe information retention during conversion of videos to embeddings, present in previous system
- Incorporated OpenAI CLIP embeddings along with Faster R-CNN for object detection within keyframes and OpenAI Whisper for audio transcription, leveraging this metadata during retrieval for better similarity score
- Achieved 85% retrieval efficiency, reducing average search time by 80 seconds and boosting user satisfaction by 30%

Von Roll USA Inc. May. 2023 – Aug. 2023
Data Science Intern | Machine Learning Team Schenectady, NY

Internal In-House Chatbot for Knowledge Sharing

- Engineered advanced Retrieval Augmented Generation (RAG) chatbot for sharing knowledge across 6 internal teams
- Integrated LlamaIndex ReAct Agents for automating the process of gaining insights from SAP generated tabular reports
- Streamlined report generation by using Llama Hub Tools for data visualization, decreasing time consumed by 60%
- Worked on integration of search and re-ranking pipelines, so that response times consistently stayed under 5 seconds

Ajio, Jio Platforms Limited Nov. 2020 – Jun. 2022
Software Development Engineer | Back-end Data Processing Team Mumbai, India

Comprehensive B2B Tax Calculation and Data Processing System

- Crafted Java-Spring Boot micro-services with Apache Kafka Streaming, ensuring seamless invoice data extraction
- Automatized feature extraction from invoices by using Azure Cognitive Services, from over 10,000 invoices per day and deployed a PySpark-based data pipeline to store the extracted features in Azure Data Lake Storage (ADLS)
- Orchestrated fault-tolerant micro-services deployment on Kubernetes, elevating scalability, and resource efficiency
- Integrated the micro-service system into Ajio's Service Mesh for catering to 9 internal teams ensuring widespread adoption

PROJECTS

Document Data Analyzer [demo] [code] Nov. 2023

- Spearheaded building RAG (Retrieval Augmented Generation) chatbot powered by vector databases and LLMs
- Leveraged HuggingFace Sentence Transformer "all-MiniLM-L6-V2" for embeddings and FAISS Similarity Search
- Assimilated OpenAI's API, LangChain framework and FAISS vector store for orchestrating the pipeline for querying model
- Optimized model output through advanced prompt engineering techniques such as Self-Refine and Chain-of-Thought

LLM-Powered SQL DB agent [demo] [code] Apr. 2024

- Built NL2SQL model with LangChain, for users to run SQL commands via natural language, eliminating prior SQL knowledge
- Used dynamic few-shot examples and context-aware table selection to ensure precise, context-specific query handling
- Acquired improved latency by hosting MySQL server on Amazon RDS and executing the query using SQLAlchemy
- Integrated memory capabilities for maintaining conversational context, reducing query resolution time by 40%

EDUCATION

University at Buffalo, Buffalo Aug. 2022 – Dec. 2023
Master of Science in Artificial Intelligence New York, US

University of Mumbai, Mumbai Jul. 2016 – Oct. 2020
Bachelor of Engineering, Computer Science Mumbai, India

SKILLS

Languages : Python, Java, C++, C, PLSQL, SQL, MongoDB

Technologies : Retrieval Augmented Generation(RAG), Large Language Models(LLMs), Natural Language Processing(NLP), Machine Learning, Deep Learning, Web Services, Data Structures, Algorithms, Prompt Engineering, Indexing, Quantization

Cloud : Amazon Web Services (AWS), Azure, Google Cloud Platform (GCP), Oracle, Atlas, OpenAI

Frameworks and Libraries : Pytorch, Tensorflow, MLFlow, Kafka, SentenceTransformers(embeddings, re-rankers), Databricks, Snowflake, PySpark, Langchain, LlamaIndex, VectorDBs(ChromaDB, Faiss, elasticsearch), MLFlow, Git, Jenkins, Big Data, Hadoop, Flask, Docker, Kubernetes, Pandas, Streamlit, Flask-RESTful, FastAPI, YOLO, Django, XGBoost, GAN, ActiveMQ, Springboot