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

NLP Search based Content Retrieval System

- Improved CIPIO's content retrieval library to efficiently fetch relevant content stored on AWS S3 buckets.
- Designed a retrieval system to accurately match user queries with content embeddings.
- Scripted the conversion of user content into OpenAI CLIP embeddings, stored them on Qdrant clusters, and integrated additional metadata such as object embeddings using Faster R-CNN, audio embeddings using Whisper & text embeddings for captions.
- Enhanced retrieval efficiency by over 80%, significantly improving the system's ability to fetch accurate and relevant content.

Von Roll USA Inc. May. 2023 – Aug. 2023
Data Science Intern | Analytics 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%

LSTM Model-based Time Series Forecasting of Sales and Consumption

- Devised demand forecasting system in supply chain, enhancing accuracy in predicting market demand and consumption
- Conducted meticulous data preprocessing, encompassing normalization, sequence generation and temporal data
- Adopted advanced deep learning to unravel complex demand patterns, leading to a 40% improvement in forecast accuracy

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

Secure B2B tax calculation based micro-service system

- Crafted Java-Spring Boot micro-services with Apache Spark Structured Streaming, ensuring seamless data extraction
- Developed tax calculation algorithms, shifting to GST-based pricing for heightened accuracy and compliance up-to 90%
- Orchestrated fault-tolerant micro-services deployment on Kubernetes, elevating scalability, and resource efficiency

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

- Led the creation of an NL2SQL model using LangChain, enabling users to execute SQL commands through natural language, eliminating the need for SQL proficiency
- Utilized dynamic few-shot examples and context-aware table selection for precise, context-specific query handling, ensuring accurate model responses
- 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