VIPUL RAJIV SARODE

Syracuse,NY | (315) 278-5565 | vsarode@syr.edu | https://www.linkedin.com/in/vipulsarode | https://medium.com/@vipul.sarode007

EDUCATION

Syracuse University, New York — August 2022 - May 2024
M.S. Applied Data Science
Relevant Coursework: Applied Machine Learning | Big Data Analytics | Natural Language Processing | Text Mining

TECHNICAL SKILLS

- **Programming Languages:** Python(Pandas, Numpy, Scipy, OOP), R, SQL
- **Parallel Computing:** CUDA, PyTorch Distributed Training, Dask
- **Programming Frameworks:** PyTorch, TensorFlow, Keras, HuggingFace, LangChain, LangGraph, Scikit-Learn, PyReft
- **Databases:** Qdrant, ChromaDB, Pinecone, Weaviate, MongoDB, MySQL, NoSQL, Hive, PostgreSQL, BigQuery
- **Technologies:** Git, DVC, Docker, Kubernetes, Weights and Biases, Airflow, Kafka
- **AWS:** Bedrock, Sagemaker, Lambda, EC2, S3

EXPERIENCE

Machine Learning Engineer at Omdena — November 2023 - Present
Developing Personalized AI Travel Advisors for Paris Olympics

- Leading the development of AI agents capable of personalized interaction, assisting users with itinerary planning, recommending authentic local attractions, providing metro guidance, and offering cultural advice to mitigate Paris syndrome and ensure health and safety awareness during travel.
- Integrating language translation features and extended travel advice beyond Olympics-related activities, encompassing comprehensive support for tourists, including language translation and cultural immersion.
- Mentoring 10 students in fine-tuning Large Language Models (LLMs), constructing scalable retrieval-augmented generators (RAGs), developing resilient AI agents, and implementing strategies to mitigate hallucinations in AI systems.

Gaming Technology: Creating Lifelike Non-Player Characters and Procedural Tasks

- Fine-tuning Mistral-7B on Wizard of Oz documents from Project Gutenberg, utilizing Hugging Face for dataset upload, recompiled Faiss embeddings with Langchain library for optimization, and establishing LM studio server.
- Executing RAG querying to impersonate characters in fiction books, accomplishing mining tasks from conversations engaging displayed UI.

Data Scientist at Jain Irrigation Systems — May 2020 - June 2022

- Forecasted sales of 150 products for next two quarters, employing deep learning time- series analysis techniques.
- Collected and analyzed sales data from three distribution centers in Maharashtra to identify patterns, trends, and seasonality, ensuring accurate sales forecasting.
- Crafted robust time-series forecasting model adopting LSTM networks in Tensorflow obtaining accuracy of 88% and improved decision-making capabilities for critical business insights.
- Collaborated and communicated with two cross-functional teams comprising supply chain and production to provide recommendations and align sales forecasts with production planning and inventory management.

PROJECTS

FinAdvisor AI — April 2024

- Implemented real-time financial news feature pipeline utilizing Alpaca API, WebSocket for data ingestion, Bytewax connector for extraction and cleaning, and MiniLM-L6-v2 encoder model for text embedding. Deployed on AWS EC2 with potential scalability to AWS EKS for multi-node setup.
- Performed distillation of Falcon-7B model from GPT-4 generated financial questions and answers and QLoRA, logged loss and model weights to Comet's experiment tracker and deployed training pipeline to Beam for training on A100 Nvidia GPUs.
- Designed an inference pipeline employing encoder, RAG, Qdrant Vector DB, and Falcon-7B model for RESTful API deployment via LangChain and Beam.

RAG API with Amazon Bedrock and Azure OpenAI — March 2024

- Built an end-to-end Retrieval-Augmented Generation (RAG) tool harnessing AWS Bedrock Service for embedding models and knowledge base, Amazon OpenSearch Service for vector database, and Azure OpenAI models for language processing, demonstrating expertise in integrated system development.
- Developed a Python API using FastAPI and configured Lambda functions to seamlessly tie together various services, showcasing proficiency in creating efficient and scalable retrieval and generation systems.

dsGPT: LLM for Data Science Code — January 2024

- Fine-tuned Gemma Instruct 7B with DScoder dataset from HuggingFace to generate data science code using LoRA leveraging custom prompt templates for instruction.