Sashidhar Reddy Duddukunta

dsasidharreddy867@gmail.com | +91 9390302684
LinkedIn | GitHub | Medium

Summary

Data Scientist with 2+ years of experience specializing in production grade Generative AI and Machine Learning. Proficient in leveraging **OpenAI, Azure, LangChain, and AutoGen** tools for innovative solutions. Deep understanding of **Retrieval-Augmented Generation (RAG), fine-tuning** large language models, and Transformers Architecture. Proven track record in applying LLM models to solve complex business challenges.

Education

Indian Institute of Information Technology, Tiruchirappalli May 2023
Bachelor of Technology CGPA: 7.7/10

Skills

Languages: Python, C, SQL, LaTeX.
Database: MySQL, Postgresql, CosmosDB, BlobStorage.
VectorDB: Azure Cognitive Search, Chroma, Pinecone, Faiss.
Libraries: Langchain, Autogen, Semantic Kernel, PyTorch, TensorFlow, OpenAI, FastAPI, etc.
Technologies & Tools: Docker, MySQL, Azure, AWS, Flask, Git, MLFlow, Redis, CI/CD Pipelines.
Data Science: Machine Learning, Probability & Statistics, XG-Boost, SVM, Regression, Classification, Numpy, Pandas, Decision Tree, Random Forest, Scikit-Learn.
NLP: Deep Learning, BERT, T-5, Fine-Tuning, Embeddings, Cross Encoders, Transformers.
GenAI: LLMs, RAG, Graph RAG, Re-Ranking, LLMops, Fine-Tuning, DPO, Quantization, LORA, Prompt Engineering, Neo-4j, GPT-4, LLAMA-3, Mistral, Gemini, Chatbots.

Work Experience

Senior Data Scientist 02-2022-present
Celebal Technologies

Advanced Knowledge Mining System

→Led the development and management of the life cycle of GenAI chatbots, scaling their deployment to end users.

- Developed AI tools such as knowledge mining (QnA), document summarization, BRD generation, and graph generation using **multi-agent** systems and large language models (LLMs), supporting **Real-time upload** and querying of various file types like PDF, Excel, PPT, doc, Image, audio, video.
- Designed and implemented a **RAG System** with hierarchical levels (organization, division, department, users), dynamically determining whether to search answers in structured or unstructured databases.
- Achieved a **7x Reduction** in preprocessing time for Excel files and a **3x Reduction** for other files using multi-threading and async processing techniques.
- Innovatively used Azure AI search to scale the uploading capacity of structured Excel files from 5 to unlimited, and productionized the entire system within the Azure ecosystem.
- Developed a multi-agent system to generate interactive graphs from Excel files.
- Implemented a caching mechanism and developed a custom streaming response algorithm to enhance system performance.
- Adhered to best coding practices like S.O.L.I.D principles and advanced exception handling techniques, while ensuring robust system security.

Travel AI Assistant

- Led the development of an Interactive AI Travel Assistant that effectively collects user preferences, generates comprehensive itinerary plans, recommends destinations, and facilitates flight bookings.
- Integrated the Bing Search API to provide video and Image previews of recommended destinations and itinerary locations.

- Integrated the AI Assistant with Flights API for ticket booking, Hotels API for accommodation recommendations, Weather API for real-time updates, and Restaurant API for dining suggestions.
- Orchestrated and applied prompt engineering techniques, including Chain of Thought (CoT), to significantly reduce hallucinations in LLM models.
- Streaming the responses from Assistant and integrated them with an avatar for enhanced user interaction.

Routing Agent

- Conducted comprehensive data analysis, including preprocessing and exploratory data analysis (EDA), on the provided dataset.
- Developed and fine-tuned a custom BERT classification model to categorize messages into academic and non-academic categories, with further subcategories within non-academic messages, and routed messages to specialized trainers based on these classifications.
- Improved model accuracy by 2% by incorporating custom dense layers on top of the BERT model.
- Successfully trained the model across multiple GPUs and deployed it using a GPU-optimized Docker image.
- Achieved a significant reduction in model size and **inference latency** by 5 seconds through optimization using ONNX format and GPU acceleration.

Machine Learning Intern 10/2021-12/2021
QuantsCase

- Developed algorithms to automate stock trading and manage portfolio data, with comprehensive details saved in Excel files.
- Build NLP models using Bert Bard that predicts the Sentiment of a given call transcript and summarize intent of the call. Prepared a detail PowerBI dashboards showing sentiment, intent, duration of call etc.
- Used API's like Yahoo Finance to fetch fundamental data, and applied various ML algorithms to predict changes in share value for the next quarter.
- Used API's like Yahoo Finance to fetch fundamental data, and applied various ML algorithms to predict changes in share value for the next quarter.

Projects

End-to-End Summarization *Feb. 2022 - Apr. 2022*

- Developed and trained a T-5 model for text summarization on a news articles dataset, achieving high-quality summaries.
- Utilized ROUGE metrics for evaluation, demonstrating significant improvements in summarization accuracy and relevance.

Text Generation *May. 2022 - May. 2022*

- Implemented and fine-tuned GPT-2 for user input-based text generation, trained on an extensive English dataset.
- Achieved high-quality and coherent text outputs, demonstrating the model's effectiveness in generating contextually relevant content.

Publication

Leveraging Pretrained Language Models for Multilingual Protest Detection CASE 2021
Paper Link - Click Here June 2021 - Aug 2021

Certifications

Microsoft Azure Fundamentals Azure Associate Data Scientist Microsoft Azure AI Engineer