Sriharsha Annamaneni

sriharsha0806@gmail.com (+91) 798 178 7689 sriharshavenugopal.com

Objective

Focused on building GENAI products leveraging Large Language Models (LLMs), Retrieval Augmented Generation (RAG), and vector search, with a strong ability to integrate heterogeneous data sources to enhance relevance in RAG applications.

Industrial Experience

- **Senior GENAI Data Scientist**, Fractal, Bengaluru May 2024 - Present
  - Led a cross-functional team to build an MVP chatbot for text2SQL applications using Langchain and Langraph. Improved system accuracy by conducting root cause analysis and redesigning schema by **27%** through schema redesign, significantly enhancing SQL query generation and overall performance.
  - Developed four specialized agents (router, clarity, text2SQL, output), enhancing the system's ability to accurately interpret and process complex business queries with an average response time of **50** seconds.
  - Scaled the solution across multiple business units, ensuring robustness and adaptability; deployed the chatbot as a microservice in Azure Kubernetes Service (AKS) with a user-friendly Streamlit-based frontend UI.
- **Senior AI Engineer**, Bosch, Bengaluru May 2021 - April 2024
  - Launched 'Ditto,' a project aimed at detecting similar software bug defects through semantic text similarity techniques. Enhanced the pipeline with a RAG (Retrieval-Augmented Generation) based system to optimize efficiency using **Qdrant**.
  - Spearheaded the creation of an Interior Monitoring System for enhancing road safety, passenger comfort, and driver focus. Implemented advanced features such as Seat Belt Detection and Drowsiness Detection using Evidential Deep Learning. Leveraged Explainable AI algorithm GradCam++ to minimize False Positives and True Negatives, leading to a significant reduction in errors.
  - Developed a smart system leveraging NLP and the MPNet sentence encoder model to associate bugs with test cases, achieving a **75**-fold increase in efficiency and **95**% test coverage.
- **Computer Vision Engineer**, Aimlytics, Hyderabad Oct 2020 - May 2021
  - Engineered an automated speech dubbing solution incorporating ASR, Speaker Diarization, and TTS capabilities, leading to a more accurate and natural-sounding output.
  - Customized a TTS synthesis model using the Indic TTS dataset, effectively capturing regional accents and dialects.
- **Research Engineer**, Sirena Technologies, Bengaluru Oct 2019 - Jun 2020
  - Crafted an offline wake-up word detection system utilizing GRU Networks for real-time responsiveness, significantly reducing latency by **13**%.
  - Formulated a Siamese Deep Neural Network for facial recognition, achieving a **99.8**% accuracy rate, contributing to enhanced security measures.

Education

- **M.S in Electrical Engineering**, Florida Institute of Technology, Melbourne, FL 2016 *GPA: 3.7/4.0*
- **B.E. in Electronics and Communication Engineering**, Manipal Institute of Technology, Manipal, India 2014 *GPA: 7.0/10*

Publications

1. N. Vallurapalli, **A. Sriharsha**, G. Varma, C.V. Jawahar, M. Mathew, S. Nagori, *Efficient Semantic Segmentation using Gradual Grouping*, CVPR Workshop 2018 (oral), Best Runner-up Award.
2. P. Keny, A. Menon, M. Rao, U. Gaitonde, A. Gupta, **A. Sriharsha**, *Development of antenna deployment circuit for nano-satellites*, ECCTD 2013.

Tools and Technologies

Python, PyTorch, OpenCV, scikit-learn, Rust, Optuna, D-Tale, **HuggingFace**, MongoDB, PostgreSQL, PySpark, **Langchain**, Semantic Kernel, **Vertex AI**, Heroku, Pillow, SpaCy, Pomegranate, nltk, **dspy**, Streamlit, FastAPI, Gradio, Docker, AWS, Rubrix, Git, Data Version Control, Luigi, **MLFlow**