Sriharsha Annamaneni

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

With over eight years of experience in Machine Learning and Deep Learning, I specialize in building Generative AI products using Large Language Models, Retrieval-Augmented Generation, and vector search. My work focuses on integrating diverse data sources to enhance the relevance and efficiency of RAG-based solutions, driving innovation and impactful results in complex AI projects.

Experience

- **Senior GENAI Data Scientist**, Fractal, Bengaluru May 2024 - Present
  - Directed a cross-functional team to successfully launch an MVP chatbot for text2SQL, using Langchain and Langraph. Boosted SQL generation accuracy by **27%**, reducing query errors and increasing system reliability by changing table schemas.
  - Innovated the development of four specialized agents (router, clarity, text2SQL, output) to enhance query interpretation, reducing average response time to **50** seconds.
  - Scaled a robust solution across business units, supporting product owners with a 99.9% uptime, deployed as a microservice in Azure Kubernetes Service (AKS) with a Streamlit-based frontend.
- **Senior AI Engineer**, Bosch, Bengaluru May 2021 - April 2024
  - Spearheaded the 'Ditto' project, which optimized the detection of similar software bugs through a semantic text similarity system, leading to a 30% reduction in bug resolution time by leveraging RAG and Qdrant technologies.
  - Pioneered the development of an advanced Interior Monitoring System, enhancing road safety and passenger comfort by implementing Seat Belt Detection and Drowsiness Detection. Reduced errors by 40% using GradCam++ and improved uncertainty handling with evidential deep learning.
  - Transformed the NLP process by engineering an MPNet-based system to automate bug-to-test-case linkage, achieving a **75**x increase in efficiency and **95%** test coverage, which significantly reduced manual effort.
- **Research Engineer**, Sirena Technologies, Bengaluru Oct 2019 - Jun 2020
  - Developed an offline wake-up word detection system using GRU Networks, reducing latency by **13%** and enhancing real-time responsiveness for embedded systems in consumer electronics.
  - Engineered a Siamese Deep Neural Network for facial recognition, achieving **99.8%** accuracy, which improved security protocols across multiple applications.
- **Computer Vision Engineer**, Aimlytics, Hyderabad Oct 2020 - May 2021
  - Developed a scalable, automated speech dubbing solution combining ASR, Speaker Diarization, and TTS, achieving high-quality, natural-sounding output suitable for diverse media applications.
  - Customized and optimized a Text-to-Speech (TTS) model using the Indic TTS dataset, effectively capturing and reproducing regional accents to improve user engagement and accessibility across multiple languages.
- **Research Fellow**, IIIT Hyderabad Jan 2017 - Dec 2019
  - Conducted applied research under Prof. C.V. Jawahar, focusing on real-world applications of Deep Learning for Road Audit Systems. Specialized in Model Compression techniques and Semantic Segmentation for Autonomous Navigation on Indian roads.
  - Optimized and implemented models such as PSPNet, ERFNet, MobileNet, and DeeplabV3 in PyTorch, contributing to advancements in autonomous driving technology, with models deployed in pilot projects.
  - Co-authored a research paper on "Efficient Semantic Segmentation using Gradual Grouping," presented at CVPR Workshop 2018, awarded Best Runner-up for its significant industry impact.

Education

- **M.S in Electrical Engineering**, Florida Institute of Technology **2016**
- **B.E. in Electronics and Communication Engineering**, Manipal Institute of Technology **2014**

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**