Balu Harshavardan Koduru

+1 (716) 907-9338 | balu.koduru99@gmail.com | github.com/BaluHarshavardan99 | linkedin.com/in/balu-koduru/

Experience

AI Engineer, TeammateMe (Remote) ***Aug 2024 - Present***

- Designed and developed a scalable **multi-agent chatbot** using **LangChain** framework, integrating with a remote SQL database and **Neo4j** knowledge graph to enable dynamic, context-aware responses to user queries.
- Incorporated **Chroma Database** and fine-tuned AI models to enhance query precision and reduce latency by **30%**, optimizing performance for high-volume data interactions.
- Deployed chatbot using **AWS Lambda**, achieving **99.5%** uptime and seamless scalability for API-based operations.

Machine Learning Researcher, SUNY RF (Buffalo, NY) ***Sep 2022 - May 2024***

- Led the development of EndoAssistant, a **Visual Question Answering (VQA)** model for endoscopic surgery analysis, and created the first image-caption dataset in endoscopy, enhancing surgical data interpretation. ***(ICLR 2025)***
- Fine-tuned LLaVA and OpenAI's CLIP, multi-modal **LLMs**, to create a specialized **Endoscopy AI assistant** that answered surgical procedure-related questionnaires with precision, improving response accuracy.
- Integrated Whisper API for **Automatic Speech Recognition (ASR)** to generate accurate captions from medical speech, reducing transcription errors by **33%** through text correction leveraging **SpaCy** and **GPT-4 API**.
- Developed a robust **NLP pipeline**, utilizing **BERT** and fine-tuned **Meta's Llama 2**, to extract structured data from patient health records, streamlining medical documentation and analysis.

Machine Learning Engineer, ACPS Group (Norway) ***Jan 2021 - Jun 2022***

- Implemented a **Multi-Agent Reinforcement Learning** algorithm to control autonomous drone fleets, leveraging Unreal Engine to optimize complex environments for effective UAV navigation and coordination.
- Developed CNN-based deep learning models in PyTorch and TensorFlow for classifying UAVs via radio signals with 99.09% accuracy and analyzing human speech signals to classify environments with 95.2% accuracy. **(Publication: IEEE)**

Data Analyst, JSW Energy (India) ***Jun 2019 - Dec 2019***

- Engineered a **deep learning model** in **TensorFlow** to manage and predict Carbon levels in 240 MW thermal power plant and provided data-driven recommendations enhancing efficiency by **15%**.
- Leveraged **Exploratory Data Analysis (EDA)** to identify plant parameters, and developed interactive dashboards in **Tableau** to visualize key performance indicators and track plant efficiency metrics.
- Executed **ETL pipeline**, merging sensor data with plant parameters, facilitating seamless data integration, & standardization.

Publications (Google Scholar)

- Recent Advances in Thermal Imaging and It's Applications using Machine Learning: A Review: ***IEEE Sensors***
- Classification of UAVs Using Time-Frequency Analysis of Remote Control Signals and CNN - ***IEEE iSES 2022***
- Light Weight Deep CNN for Background Sound Classification in Speech Signals - ***JASA***

Projects

Prompt Engineering: Hallucination mitigation in Chatbots

- Enhanced LLM accuracy by implementing prompt structuring and iterative refinement; decreased hallucinations by **40%**, leading to a **25%** increase in factual, verifiable information output.
- Formulated transformer model built on **RoBERTa** for hallucination mitigation in chat-bots achieving **87.4%** accuracy.

Real-Time Face Mask Detection Using Advanced Object Detection Models

- Designed and implemented a machine learning model for real-time face mask detection, leveraging state-of-the-art object detection frameworks such as **Faster-RCNN** and **YOLO v8**.
- Conducted extensive experimentation and fine-tuning, achieving **97.6%** classification accuracy while optimizing model performance for deployment in scalable applications.

Education

University at Buffalo, The State University of New York ***May 2024***
Master of Science (MS) in Artificial Intelligence ***GPA:*** *3.75/4.00*

Birla Institute of Technology and Sciences Pilani, India (BITS Pilani)
Bachelor of Engineering (B.E.) in Electronics and Instrumentation

Relevant Skills and Competencies

- **Programming Languages:** Python, C++, MySQL, PostgreSQL, MATLAB.
- **Libraries/Modules:** PyTorch, TensorFlow, Keras, OpenCV, Scikit-learn, Pandas, Hugging Face, NLTK, Spacy, LangChain.
- **Machine Learning Skills:** MLOps, Generative AI, Natural Language Processing (NLP), Large Language Models (LLM), RAGs, Prompt Engineering, Computer Vision, Graph Neural Networks, Reinforcement Learning, Exploratory Data Analysis (EDA), Data Visualization, ETL.
- **Tools/Cloud:** Git, Tableau, Llama, GPT API, PySpark, AWS, MS Office Suite, Anaconda, ROS, AirSim, Linux.