Sai Sravan Yarlagadda

saisravy@andrew.cmu.edu | (585) 910-3383 | linkedin.com/in/saisravany | github.com/Sai-Yarlagadda

EDUCATION

Carnegie Mellon University **Pittsburgh, PA**
Master of Science, Artificial Intelligence Engineering- Mechanical Engineering GPA: 4.0/4 | Dec 2024
Relevant Courses: Advanced Natural Language Processing, Deep Learning, Machine Learning, Data Engineering.

Vellore Institute of Technology **Vellore, India**
Bachelor of Technology, Mechanical Engineering GPA: 3.89/4 | July 2023
Relevant Courses: Programming, Object Oriented Programming, Optimization, Applied Linear Algebra, Statistics.

EXPERIENCE

Mechanical And AI Laboratory, Research Assistant - NLP, Mar 2024 – Present **Pittsburgh, PA**
- Modeled an NLP system employing a self-refinement approach to generate CAD models based on input prompts.
- Analyzed the performance of various LLMs like Gpt-4, LLaMA-70b for generating macros used to create 3D models.
- Finetuned various image-to-text models like BLIP2 and ViT-GPT2 for evaluating the output generated by the LLM.
- Utilized Nvidia A6000 Lambda with a GPU capacity of 150GB and AWS EC2 instances to run these large models.

Carnegie Mellon University, Course Assistant – AIML (24787), Jan 2024 – Present **Pittsburgh, PA**
- Teaching students various concepts of Machine Learning and Artificial Intelligence and python and holding regular office hours every week.

Packaging Dynamics Laboratory, Research Assistant, Jan 2023 – May 2023 **Rochester, NY**
- Investigated vehicle vibration techniques and the effects in packaged products due to the scuffing caused by vibrations.
- Developed a setup for quantifying the wear and tear of the corrugated plastic containers during transportation.
- Captured power spectral density profiles of a truck that travelled on interstate and city roads of Rochester and simulated these profiles on the test setup and compared the package with respect to main scuffed package.
- Ran various data analysis to identify the events that occurred with a range of two vibration intensities.

ACADEMIC PROJECTS

Retrieval Augmented Generation System capable for answering questions regarding CMU's LTI department
- Engineered an end-to-end NLP system tailored for answering domain specific queries related to Carnegie Mellon University.
- Implemented Advanced RAG methods like reranking and multi-query retrieval for enhancing accuracy in question answering.
- Enhanced models capability to answer complex questions by integrating a dense retrieval model with LLaMA2 model.
- Obtained an F1 score of 41% that is a notable threefold improvement over using LLAMA2 model independently.

Implementing LLAMA2 from Scratch: A Comprehensive Approach to Machine Learning Architecture Design
- Designed and implemented the LLAMA2 model which is a leading open-source language model by Meta AI.
- Demonstrated a strong understanding of complex topics like RoPE and included this in the architecture designed.
- Executed zero-shot sentiment analysis on CFIMDB datasets and finetuned the model on the Tiny Stories dataset.
- Achieved results stating the finetuned model is 32% more coherent in generating answers than the model without finetuning.

End to End Data Engineering Process on the NSL- KDD dataset
- Predicted and classified network intrusions by understanding the patterns in previous occurred attacks.
- Performed various big data analytics and data cleaning on the data using PySpark and used postgres to handle the SQL databases of a dataset consisted of 125k entries and 41 features...
- Worked on various machine learning models, including logistic regression, support vector mechanisms and gradient boosted decision trees, to classify network intrusions accurately.
- Orchestrated the whole data engineering process including deployment in the Google Cloud Platform (GCP)

Implementation and optimization of a stereo Visual Odometry SLAM System from scratch
- Designed a custom SLAM system to gain comprehensive insight into every stage of the Localization and Mapping pipeline.
- Established loop closure detection using Bag of Words algorithm and implemented Georgia Tech Smoothing and Mapping framework for backend development.

SKILLS

Programming: Advanced: Python; C++, Intermediate: MATLAB
Libraries: Advanced: OpenAI, Lang Chain, Hugging Face, Spark MLlib OpenCV, Pytorch, Keras, Pytorch
Tools: Docker, SQL, Google Cloud Platform(GCP), AWS, ChromaDB