Navya Battula

Portfolio: navyabattula.github.io
Github: github.com/navyabattula
Email: navyabio12@gmail.com
Mobile: +1 (820)-758-8379

About

Highly motivated and skilled **AI Engineer** with over an **year of experience in developing and deploying Generative AI and Deep learning models**. Demonstrated expertise in using cutting-edge AI tools and frameworks to build robust, scalable solutions. Proficient in a wide range of technologies including **PySpark, Hadoop, PostgreSQL, TensorFlow, PyTorch, Open CV and more**. Proven track record of success through hands-on projects and internships, with significant contributions in Computer Vision, Machine Learning, ChatBots and AI-driven document management systems. **Published researcher** with a strong foundation in both theoretical and applied AI.

Education

- **University of California, Santa Barbara** — Santa Barbara, USA
  Masters - Computer Science; GPA: 3.87 — *September 2021 - June 2023*
  Courses: *Special topics Deep Learning, Information Retrieval systems, Run time systems, Machine Learning for Graphs, Machine learning for networking systems, Software fuzzing, Scalable internet services, Advanced topics Computer Vision*
- **PVP Siddhartha Institute of Technology** — Vijayawada, India
  Bachelor of Technology - Computer Science and Engineering; GPA: 9.53 — *July 2017 - July 2021*
  Courses: *Operating Systems, Data Structures, Analysis Of Algorithms, Automata Theory, Computer Architecture, Android Development, Networking, Databases*

Skills Summary

- **Languages**: Python, C, C++, SQL, R, Bash, JAVA
- **Frameworks**: Scikit,TensorFlow, Keras, Torch, Huggingface, Pyspark, Big Query, Open CV, Hadoop
- **Tools**: Docker, GIT, PostgreSQL, Oracle, MySQL, SQLite, PowerBI,Matplotlib, Streamlit
- **AI/ML Technologies**: LangChain, LlamaIndex, Pinecone VectorDB, Llama LLM, Gemma, Mistral, OpenAI GPT, Anthropic Claude
- **Cloud Platforms**: AWS (SageMaker, Lambda, EC2), Google Cloud Platform
- **MLOPs**: Docker, Kubernetes, MLflow, Weights & Biases

Experience

- **OQ Point LLC** — Onsite - Redmond, WA
  AI/ML Engineer — *June 2023 - Present*
  As a key member of the AI development team I worked on and deployed a sophisticated **AI-powered content recommendation system** using **LangChain and Pinecone VectorDB**, resulting in a **40% increase in user engagement**. Developed a **multi-model AI assistant** integrating **OpenAI GPT and Anthropic Claude**, enhancing customer support efficiency by 60%. Implemented a **real-time document analysis pipeline** using **LlamaIndex and PostgreSQL**, processing over **1M documents daily with 99.9% accuracy**. Optimized large language model inference using **AWS SageMaker**, **reducing latency by 35% and costs by 25%**. Development of a Generative AI platform for creating synthetic training data, improving model performance by 30% across various use cases.
- **Productive Robotics** — Onsite - Santa Barbara, CA
  Computer Vision Intern (Part-time, Research Collaboration) — *March 2022 - June 2022*
  Developed an **Open CV module** that detects surroundings in a metal shop based on **depth information of door and color information of the light**. Made use of bounding boxes to identify door surroundings and used color channel operations to distinguish between lights. Had experience with **C++ and low latency programming** while coding depth detection modules for robot camera aparatus.
- **Indian Academy of Sciences** — Remote - Bangalore, KA
  Summer Research Fellowship (Internship) — *July 2020 - Nov 2020*
  Designed a **neural network model to predict emission and transmission probabilities based on packet drop statistics** in the Gilbert Elliot channel. Employed the **Sim2net** Python package to collect packet drop statistics as input for the model. Achieved a **remarkable accuracy** of approximately **75%**, **surpassing the existing state-of-the-art Baum Welch algorithm**, which yielded approximately 55% accuracy. **Published research paper** accepted by the **Indian Academy of Sciences Summer Research Fellow Reports journal**.
- **Indian Severs** — Hybrid - Vijayawada, AP
  Machine Learning Engineer Intern (Part-time) — *June 2019 - June 2020*
  Developed an **Open CV framework along with de-noising** that is capable of **diagnosing Diabetic Retinopathy** from retinal fundus images. Reports accuracy of **96% on normal images and 91% on noisy images**.

Projects

- **Chat with SQL - A Agent based CSM model to chat with databases.**: Developed an AI-powered **customer support chatbot using Streamlit** that seamlessly connects to an **SQL database** to retrieve and provide real-time sales data to users. The chatbot leverages **LangChain and open source LLMs** like **Llama and Gemma** for natural language processing, enabling customers to **query sales records with ease**. The solution integrates **Pandas** to format and display query results in a structured, tabular form, ensuring clear and accurate information delivery. Implemented caching mechanisms to optimize database interactions and improve performance. The system also features robust error handling and fallback methods, ensuring a smooth user experience. This chatbot significantly **enhances customer support operations** by allowing agents and users to quickly access critical data without manual intervention. .
- **Q&A ChatBot using Langchain and Open Source LLMs**: Created a **Question and Answer Chatbot** utilizing the **Langchain Framework** that utilizes the choice of **Llama 2, Llama 3 and Gemma 2 models**. Designed an intuitive user interface using **Streamlit**. Maintained chat history using Langchain tools to retain short term memory. .
- **Search Engine Application based on AI Agents using Open Source LLMS**: Created a **Search Engine Tool** utilizing the **Langchain Framework and AI Agents** along with LLMs **Llama 2, Llama 3 and Gemma 2 models**. The search engine uses Tools like **Wikipedia, Arxiv and Duck Duck Go Browser** to scrape through online content and answer specific queries. Designed an intuitive user interface using **Streamlit**. .
- **Network Data Processing Pipeline Framework PERRY (MS Project)**:
  Designed and implemented **PERRY**, an **adaptable data processing framework for networking data**. This versatile solution efficiently **cleans, processes, and converts raw packet capture data into refined feature sets** (packet, burst, flow) with speed and fault tolerance, accommodating even large datasets.
- **Physical Training app using TensorflowJS**: Created a **Physical Training web application** utilizing the **TensorFlowJS-based MoveNet model** for **real-time exercise tracking through pose estimation** from a webcam. Designed an intuitive user interface facilitating effortless dataset curation and training. The **lightweight nature** of TensorFlowJS ensures smooth performance, even on **low-end mobile devices**. .

Publications

- **PERRY: Flexible and Scalable Data Preprocessing System for "ML for Networks" Pipelines (MS Thesis)**: PERRY is my masters thesis dissertation in which I discuss the prevailing issue of tight coupling between data processing and model training parts in Machine learning for networking pipelines and then try to address this problem with a flexible and scalable data processing framework called PERRY. Leveraging state of the art tools and being scalable with limited resources makes our framework an easy to use solution for networking researchers.
- **netFound: Foundation Model for Network Security (Research paper)**: In ML for network security, traditional workflows rely on high-quality labeled data and manual feature engineering, but limited datasets and human expertise hinder feature selection, leading to models struggling to capture crucial relationships and generalize effectively. Inspired by recent advancements in ML application domains like GPT-4 and Vision Transformers, we have developed netFound, a foundational model for network security. This model undergoes pre-training using self-supervised algorithms applied to readily available unlabeled network packet traces. netFound's design incorporates hierarchical and multi-modal attributes of network traffic, effectively capturing hidden networking contexts, including application logic, communication protocols, and network conditions. With this pre-trained foundation in place, we can fine-tune netFound for a wide array of downstream tasks, even when dealing with low-quality, limited, and noisy labeled data. Our experiments demonstrate netFound's superiority over existing state-of-the-art ML-based solutions across three distinct network downstream tasks: traffic classification, network intrusion detection, and APT detection. Furthermore, we emphasize netFound's robustness against noisy and missing labels, as well as its ability to generalize across temporal variations and diverse network environments. Finally, through a series of ablation studies, we provide comprehensive insights into how our design choices enable netFound to more effectively capture hidden networking contexts, further solidifying its performance and utility in network security applications. .
- **Optimal Parameter Estimation for Low Latency Communication using Deep Neural Networks (Reasearch article)**: Developed a neural network model to predict the emission and transmission probabilities in the Gilbert Elliot channel. Collected packet drop statistics using Sim2net package python as input for the model. Reported ˜75% accuracy which was better than previously explored Markov chain models. Paper accepted at Indian Academy of sciences Summer Research Fellow reports journal.