Navya Battula

Portfolio: navyabattula.github.io
Github: github.com/navyabattula
LinkedIn:navya-battula

Email: navyabio12@gmail.com
Mobile: +91 83282 77463

About

With 3 years of experience as an **AI/ML Engineer**, I specialize in **architecting production-grade LLM systems**— spanning **hybrid recommendation engines, RAG-based entity extraction, and agentic AI chatbots**. I design and deploy scalable platform-agnostic solutions using LangChain, LangGraph, and AWS Bedrock, with deep expertise in **fine-tuning, prompt optimization, intermediate representation translation, and enterprise guardrails**. My work consistently delivers measurable business impact: 40%+ improvements in engagement and efficiency, 25% cost reductions, and 99.9% system reliability. I **bridge the gap between research and operations**, turning **complex AI capabilities into resilient, high-performance enterprise products**. **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 - September 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, Selenium
- **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

- **Acuitilabs** — Pune, India
  ML Engineer — *October 2025 - Present*
  1. Architected a **platform-agnostic, SaaS-based AI assistant** capable of seamless integration across **3 distinct operational portals** within the organization, utilizing **LangGraph** to orchestrate complex agentic workflows that maintain state and context across multi-turn conversations.
  2. Developed a novel **Intermediate Representation (IR)** layer that decouples natural language intent from execution logic; this abstraction enables the system to dynamically translate a single user query into **dialect-specific SQL** (PostgreSQL, Snowflake, Redshift, MySQL) depending on the underlying connected platform.
  3. Implemented **hierarchical guardrails** and security filters, performing **real-time PII redaction, prompt injection detection, and topic restriction**, ensuring compliance with enterprise security policies without sacrificing conversational fluidity.
  4. Engineered an **adaptive intent recognition module** that combines LLM-based classification with few-shot prompting to handle ambiguous user requests, routing queries to specialized sub-agents based on context.
  5. Established a **multi-tenant configuration** framework allowing distinct client portals to maintain unique knowledge bases, SQL schemas, and brand voice parameters while sharing the same underlying agentic engine.
  6. Engineered a **dynamic RAG-based feedback loop** for entity extraction within the Smart Match payment reconciliation system, leveraging AWS Bedrock (Llama/Claude) to parse unstructured payment notes and remittance data with high precision.
  7. Designed and iterated on **context-aware dynamic prompts** utilizing few-shot learning and self-refining prompt generation, enabling the system to adapt to varied financial document formats and improving entity extraction accuracy while significantly reducing hallucination rates.

- **Alphanome.AI** — Remote
  AI Researcher Freelance — *April 2025 - May 2025*
  Conceptualized and implemented Antetic AI, a project utilizing **Multi-Agent Reinforcement Learning (MARL)** to model emergent **collective intelligence based on ant colony principles**. Focused on designing and optimizing objective functions to drive effective collaboration and communication strategies among agents operating in diverse simulation environments.

- **OQ Point LLC** Onsite - Redmond, WA
  ML Engineer *April 2023 - March 2025*
  1. As a key member of the AI development team I worked on and deployed a sophisticated **AI-powered content recommendation system** using **GPT, BERT Foundation Models and Pinecone VectorDB**, resulting in a **40% increase in user engagement**.
  2. Performed **fine-tuning** on these models for generating personalized recommendations based on user needs that led to the spike in engangement.
  3. Developed a **multi-modal AI assistant** integrating **OpenAI GPT and Anthropic Claude**, enhancing customer support efficiency by 60%.
  4. Implemented a **real-time document analysis pipeline** using **LlamaIndex and PostgreSQL**, processing over **1M documents daily with 99.9% accuracy**.
  5. Optimized large language model inference using **AWS SageMaker**, **reducing latency by 35% and costs by 25%**.
  6. Developed the components following a **microservices architecture** and integrated them using **RESTful APIs**. Utilized AWS technologies like **AWS Sagemaker, AWS Glue, API Gateway, EC2, Elastic Load Balancer, AWS ECS, AWS S3, AWS CloudWatch, AWS ECS** for developing the application on cloud.

- **Productive Robotics** Onsite - Santa Barbara, CA
  Computer Vision Intern (Part-time, Research Collaboration) *March 2022 - June 2022*
  Developed an **OpenCV-based module** for **3D environmental analysis** using depth information and color channel operations to detect and classify objects (e.g., doors, lights) in a metal shop. Implemented bounding box algorithms for precise object localization and **depth-based 3D reconstruction**, enabling scalable environmental mapping. Designed robust color differentiation techniques to adapt to varying lighting conditions. Applied **PyTorch for ML modeling and evaluation**, with potential to extend the pipeline for **3D Gaussian Splatting (3DGS) or Neural Radiance Fields (NeRF)** for novel view synthesis and synthetic data generation, reducing manual labeling efforts for simulation and training.
- **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: Developed netFound, a foundational ML model for network security that leverages self-supervised pre-training on unlabeled network packet traces to eliminate reliance on manual feature engineering and labeled datasets. By incorporating a hierarchical, multi-modal architecture, the model captures hidden network contexts (e.g., application logic, protocols) and generalizes effectively across dynamic environments. Experiments demonstrated state-of-the-art performance in traffic classification, intrusion detection, and APT detection, outperforming existing solutions under low-quality/noisy labels, while ablation studies validated robustness to temporal shifts, missing data, and design choices critical for real-world deployment. The framework reduces dependency on curated data and enhances adaptability for diverse security tasks.**: .
- **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.