VISHNU SAI VARDHAN REDDY BASI REDDY GARI

vishnubr@umd.edu | +1 2272751335 | College Park, MD | LinkedIn | Github | Website | Medium

EDUCATION

Univeristy of Maryland, *United States* *Aug 2025 - Jun 2027*

Master's of Science in Applied Machine Learning

Courses : NLP, Statitics & Probability

Vellore Institute of Technology, *India* ***CGPA : 8.36/10*** *Jul 2019 - Jun 2023*

Bachelor's of Science in Computer Science Engineering with specialization in Data Science

Courses : Data Structures & Algorithms, Deep Learning, Python, Database Management, Cloud Computing, Artificial Intelligence

EXPERIENCE

Machine Learning Engineer | Maruti Suzuki India Limited, *Bengaluru, India* *Jun 2023 - Jul 2025*

Agentic Data Analyst

- Automated Data analysis by Architecting and Builing a multi-agentic system using Agency Swarm and **Azure Open AI API**
- Engineered memory using Memgraph for short-term context graphs and ChromaDB+Neo4j for long-term **RAG** retrieval.
- Enhanced **contextual reasoning** with LangChain pipelines, achieving **99%** retrieval accuracy and **30%** lower token usage.
- Built a Streamlit interface to enhance user experience, reducing analysis time from **10 days to 10 minutes** and human effort by 90%.

Telematics Chatbot - Natural Language to SQL

- Designed a GenAI chatbot to democratize data access by translating natural language into optimized SQL for self-service analytics.
- **Fine-tuned Flan-T5** using LoRA and PEFT on domain SQL/NL datasets, achieving **99%** accurate query-to-database mapping.
- Optimized inference using ONNX Runtime and mixed-precision quantization, reducing GPU memory by **25%** and latency by 40%.
- Drove business impact by boosting user engagement by **80%** and improving cross-team collaboration by 58%

Machine Learning Intern | Maruti Suzuki India Limited, *Bengaluru, India* *Jan 2023 - Jun 2023*

Accident Detection using LSTM AutoEncoders

- Developed an LSTM autoencoder to predict accidents by modeling sequential time-series data and detecting anomalies.
- Engineered features using rolling statistics and window-based methods to capture temporal patterns and improve prediction accuracy.
- Validated insights through EDA and hypothesis testing, achieving an **F1-score of 82%** for reliable deployment.
- Created a benchmark dataset from processed accident data to enable hotspot dashboards and safety design improvements.

PROJECTS

AI Event Reminder | *Tesseract, MLX, Swift, Core ML* Github

- Launched an iOS app that extracts text from images via **OCR** and converts it into structured calendar events.
- Fine-tuned SmolLM-360M using **MLX** + Unified Memory for accurate event entity extraction.
- Optimized on-device inference for **<2s latency**, ensuring smooth, real-time iOS performance.
- Boosted user engagement with **50%** adoption in the first month, simplifying event management for students.

Visual Search Engine | *Pytorch, FAISS, Cuda, Streamlit, Hugging face, tansformers* Github

- Developed a CLIP-powered multimodal search system to recommend visually similar apparel from uploaded images.
- Engineered object-level detection extensions to identify **multiple entities (person,shirt)** for context-aware visual similarity retrieval.
- Optimized embedding pipelines using FAISS, PyTorch, and **Transformers**, achieving <200 ms latency across 100K+ fashion images.
- Deployed a scalable, containerized pipeline on AWS SageMaker with Streamlit UI, boosting search speed by 40%.

SKILLS

Programming	:	Python, SQL, C++, C, CUDA, R
Frameworks	:	Pytorch, MLX, Tensorflow, Langchain, Tesseract, Core ML, Agency Swarm, ADK, Crew AI, Unsloth AI
Generative AI	:	Agentic AI, RAG, LLM Finetuning, Model Profiling, Ollama, OpenAI, Gemini, Distributed Training,ONNX
Databases	:	Vector DB - Chroma DB, Pinecone | RDBMS - MYSQL, DB2
Cloud & Web	:	Aws, Azure, GCP, Docker, Kubernets, Git, CI/CD, ML flow, LLMops, Stream-lit, FastAPI

CERTIFICATES

AWS - Certified Machine Learning Engineer Associate

Azure - Developing AI Applications on Azure