HARSHA VARDHAN REDDY EMANI

Tsundur, Guntur, Andhra Pradesh, India | 6301791287 | ehvrreddy456@gmail.com | LinkedIn | Github

Enthusiastic AI, HPC, and Computer Vision developer with strong expertise in Python and modern machine learning frameworks, focused on deploying deep learning solutions for real-world impact.AI Research and Systems Engineering student specializing in High-Performance Computing and Large Language Model optimization. Hands-on experience fine-tuning and building models from scratch across AMD MI300X, Intel Gaudi3, and NVIDIA L40S GPU platforms.

SKILLS AND ABIILITIES

- **Languages:** Python, SQL, Java, C, HTML, CSS, JavaScript, Dart
- **Frameworks:** Pytorch, Langchain, LangGraph, Scrapy, Pandas, Numpy, Scikit-Learn, Fastapi, Flutter , Nodejs , Expressjs
- **Databases:** MySQL, Oracle SQL,Berkely DB, MongoDB
- **Tools:** Git, GitHub, Docker
- **Platforms:** Jupyter Notebook, Visual studio Code, Linux (Ubuntu)
- **GPU:** CUDA, OpenMP, AMD ROCm, Intel
- **Soft Skills:** Technical Public Speaking (SCI 2025 Tutorial), Peer Mentorship (AIHUB), People Management, Communication Skills

EXPERIENCE

Speaker | SCI 2025 Conference ***Dec 2025***

- Selected to deliver a 3-hour technical tutorial titled **"Mathematical Foundation for Large and Small Language Models"**.
- Demonstrated live LLM deployment and performance benchmarking on **Intel Gaudi3 and AMD ,NVIDIA hardware**.

Intern - High Performance Computing (HPC)
Centre for Development of Advanced Computing (CDAC), Pune
April 2025 - June 2025

- Contributed to the sub-project **"High Performance Data Management Using Berkeley DB for Telecom Traffic Data."**
- Designed and optimized **graph-based data structures in C/C++** for modeling complex telecom networks.
- Leveraged **Berkeley DB and OpenMP** to parallelize and accelerate large-scale data access.
- Gained exposure to **HPC cluster environments**, performance profiling, and scalable data visualization..

PROJECTS

Medical LLM & SLM Development – Cross-Hardware Training
Tech Stack: Python, PyTorch, Hugging Face, ROCm/CUDA, DeepSpeed, LoRA, DDP

- Built a **GPT-2–style Small Language Model from scratch**, implementing custom Causal Self-Attention, LayerNorm, and Transformer blocks.
- Enabled **cross-hardware, multi-GPU training** on an 8-GPU AMD MI300X system using Distributed Data Parallel (DDP) and gradient accumulation.
- Performed **Supervised Fine-Tuning (SFT)** of a 7B medical LLM using **LoRA** for parameter-efficient adaptation.
- Optimized training stability and memory efficiency for **verifiable medical reasoning tasks** using SFTTrainer and mixed-precision strategies.

SocialGPT – Multi-Agent RAG System
Tech Stack: Python, LangGraph, FastAPI, Vector Databases, LLMs

- Developed a LangGraph-based multi-agent Retrieval-Augmented Generation (RAG) system with an intelligent router node to switch between vector search and real-time API retrieval.
- Implemented automated roll-number extraction for accurate intent parsing and dynamic query routing.
- Integrated real-time data fetching from university result endpoints for live student information access.
- Deployed the end-to-end RAG workflow using FastAPI, enabling scalable and low-latency API-based inference.

TrackCoders – Campus-Scale Student Analytics Platform | Website
Tech Stack: FastAPI, MongoDB, React, REST APIs, Linux Server Deployment,Docker

- Architected and deployed a **production-grade student analytics platform** used campus-wide by **5,000+ students and faculty**.
- Built a **scalable FastAPI + MongoDB backend** with a React frontend for real-time student status tracking.
- Deployed on **local campus servers**, optimized for high internal traffic and low-latency access.
- Streamlined **administrative and academic workflows** through real-time dashboards and progress visualization.

Prompt Classification ML - Hierarchical Prompt Categorization | LINK
Tech Stack: Python, Scikit-learn, TF-IDF, Pandas, NLTK, Pickle

- Developed a full ML pipeline for classifying prompts using TF-IDF, clustering, and hierarchical classification techniques.
- Preprocessed and augmented datasets (expanded_prompts.csv) to improve model accuracy and generalization.
- Persisted trained models and vectorizers using pickle for consistent offline and web-based inference.
- Documented the entire workflow using Jupyter notebooks and generated professional reports in DOCX/PDF formats.

ACHIEVEMENTS | EXTRA CIRCILLARS

- **Core Member & Platform Manager -** AIHUB VVIT **Organization**
  Spearheading AI/ML project development, mentoring peers in applied AI, and organizing community tech events. Deployed and currently maintain the official AIHUB VVIT website and backend infrastructure
- Exploring and implementing **GPU-accelerated** AI pipelines, including model parallelism, distributed training, and performance benchmarking.
- **Organized 3 - 4 student hackathons and technical challenges** under AIHUB VVIT, encouraging innovation and team-based problem-solving.
- **Member of ACM (Association for Computing Machinery)** student chapter, participating in seminars, coding events, and tech talks.

CERTIFICATIONS

Programming in Java NPTEL online Certification - 2024

- Gained strong proficiency in core Java concepts including **OOP, multithreading, exception handling, and file I/O**.
- Worked on assignments and proctored exam based on real-world programming problems using **Java SE**.

Generative AI by Google Cloud Google Cloud Skill Boost (L4G) - 2024

- Learned **foundation models, large language models (LLMs), and prompt engineering**.

EDUCATION

Vasireddy Venkatadri University(VVITU) **Andhra Pradesh, India**
Bachelor of Technology (B.Tech); GPA: 8.7 Aug 2023 – Ongoing