Tarun Reddy

Boston, MA | (857)-693-4743 | thandu.t@northeastern.edu | www.linkedin.com/in/tarun-reddy

EDUCATION

Khoury College of Computer Sciences, Northeastern University, Boston, MA May 2025

Master of Science in Artificial Intelligence (GPA 3.96/4)

- Courses: Large Language Models, AI for Human Computer Interaction, Deep Learning, Machine Learning, Natural Language Processing, Reinforcement Learning
- Roles: Head Teaching Assistant for Graduate Level Natural Language Processing

National Institute of Technology, India Jun 2020

Bachelor of Technology in Electronics and Communication Engineering (Gold Medalist)

- Roles: Class Representative, Training and Placement Coordinator, Head of Electronics Club

SKILLS

Machine Learning: *TensorFlow, PyTorch, scikit-learn, Keras, XGBoost* **Deep Learning:** *Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM)* **Programming Languages:** *Python, C++, JavaScript* **Automotive Technology:** *OpenCV, ROS (Robot Operating System), CARLA Simulator* **Data Analysis:** *Pandas, NumPy, MATLAB* **Computer Vision:** *OpenCV, TensorFlow Object Detection, YOLO (You Only Look Once)* **Software Development:** *Git, Docker, Agile methodologies* **Sensor Technologies:** *LiDAR, RADAR, GPS* **Cloud Computing:** *AWS, Google Cloud Platform (GCP), Microsoft Azure*

WORK EXPERIENCE

HerHeard, Full-Stack AI Engineer Intern, Cambridge, MA Sep 2024 - Dec 2024

- Developed conversational chatbot using LLMs (RAG, knowledge graphs, agentic workflows), evaluated performance with advanced metrics, and integrated AI into production React apps.
- Contributed to design and implementation of a personalized daily journal and dynamic news feed dashboard, leveraging cloud APIs and scalable backend services.
- Enhanced user interaction through prompt engineering and LLM evaluation, deploying new features via CI/CD pipelines for a secure and global user base.

Infosys, Specialist Programmer, Bangalore, India Sep 2021 - Jan 2023

- Built REST APIs and microservices using Spring Boot, Flask, and Jenkins on AWS, automating deployments and enhancing system reliability for Apple DevSecOps Portal.
- Developed PostgreSQL-backed dashboards with Angular for deployment tracking; performed predictive reliability monitoring using statistical analysis.
- Collaborated with Apple teams in Agile sprints; integrated Jenkins, AWS Secrets Manager, and Microsoft 365 tools across cross-platform environments.

Srisys Inc., Software Engineer Intern, Hyderabad, India Mar 2021 - Aug 2021

- Developed ERP extensions and cloud-based apps using Java, Spring Boot, Angular, and PostgreSQL; built CI/CD pipelines with Jenkins and Puppet.
- Integrated AWS services for scalable backend solutions; delivered custom features for clients in manufacturing and healthcare sectors.
- Collaborated with cross-functional teams to ensure timely delivery of robust and maintainable solutions meeting client requirements.

IIT Hyderabad (VIGIL Lab), Research Intern, Hyderabad, India Aug 2021 - Feb 2022

- Conducted research on object detection, dilation, and attention networks using PyTorch, TensorFlow, Keras, and OpenCV; contributed to computer vision advancements.
- Built and evaluated ML pipelines for video data, enhancing model performance with advanced preprocessing and deep learning architectures.
- Validated results through rigorous testing and applied state-of-the-art techniques on academic datasets in collaboration with the VIGIL Lab.

IIT Indore, Research Associate, Indore, India Jul 2020 - Nov 2020

- Developed method for DSP hardware IP piracy prevention by generating palmprint biometrics-based digital signatures and embedding them in RTL design.
- Altered register allocation for enhanced security, resulting in a peer-reviewed IEEE Transactions on Consumer Electronics publication.
- Collaborated with interdisciplinary teams to validate novel hardware security solutions and document reproducible design methodologies.

IIT Indore, Research Intern, Indore, India May 2019 - Jul 2019

- Worked on transient fault tolerance in DSP cores using compiler-driven transformations and simulated annealing for floorplanning optimization.
- Applied AI search methods such as hill climbing and particle swarm optimization to improve delay, area, and power metrics in VLSI systems.
- Documented improvements in DSP core reliability and efficiency, contributing to ongoing research in digital hardware design optimization.

PROJECTS

Generative AI, Domain-Specialized RAG Chatbot with Multimodal Agentic AI Feb 2024 - Jun 2024

- Built RAG chatbot tailored for healthcare, law, and finance using LangChain, Pinecone, OpenAI GPT, and AWS; integrated knowledge graphs for domain expertise.
- Developed agentic workflows with multimodal LLMs and Crew AI; improved user experience via RLHF fine-tuning, enhancing both response accuracy and human alignment.
- Deployed scalable cloud solution for real-time interactions; managed secure API integration and robust data handling with CI/CD pipelines and observability tools.

Machine Learning, Safe and Aligned LLMs with LoRA, PEFT, and Quantization Jan 2024 - Apr 2024

- Fine-tuned open-source LLMs (LLaMA, Mistral) from Hugging Face using LoRA, PEFT, and quantization; enhanced safety and relevance for practical deployments.
- Evaluated models on harmfulness, relevance, and hallucination using RAGAS and FMeval; ensured alignment with ethical AI standards and reliability requirements.
- Optimized inference and continuous monitoring; integrated with regulatory-compliant GenAI pipelines and performed robust multi-metric reporting.

Reinforcement Learning, Automated Stock Trading & Portfolio Optimization Bot Oct 2023 - Jan 2024

- Developed automated stock trading bot using PPO, SAC, and A2C algorithms in PyTorch; integrated financial data via yFinance API and custom feature engineering.
- Executed extensive backtesting, optimizing risk-adjusted returns; used Sharpe ratio and max drawdown to quantify performance gains and portfolio stability.
- Implemented real-time analytics dashboard; evaluated deployment scenarios, enhancing both system interpretability and trading strategy transparency.

Computer Vision, ASL Sign Language Detection with Deep Learning Aug 2023 - Nov 2023

- Engineered ASL sign detection system using CNNs, LSTMs, Transformers in TensorFlow and PyTorch; leveraged OpenCV for video preprocessing and augmentation.
- Achieved high classification accuracy with cross-validation and rigorous evaluation; improved robustness via model ensembling and feature fusion.
- Validated in real-world scenarios; deployed RESTful Flask API for accessible inference in assistive technology and communication tools.

Vision-Language Models, Lightweight Multimodal Visual Question Answering with Gemma Jun 2023 - Aug 2023

- Developed visual question-answering model inspired by LLaVA using Google's Gemma architecture; applied vision-language alignment for efficient edge deployment.
- Integrated performance optimizations and resource-constrained design; demonstrated significant gains on standard VQA benchmarks with minimal latency.
- Enabled multimodal reasoning on mobile and IoT platforms; validated deployment with cross-device benchmarking and low-power constraints.

Software Engineering, Modular Java Image Processing Suite Feb 2023 - May 2023

- Engineered robust image-processing application in Java using MVC and SOLID principles; implemented filters, transformations, and editing features with JavaFX.
- Applied Factory and Observer design patterns; ensured modularity and maintainability for production-scale deployment and ongoing extensibility.
- Conducted unit/integration testing; delivered scalable solution with clear documentation, facilitating team collaboration and future feature growth.

Computer Vision / Deep Learning, Rotation-Invariant Object Detection in Aerial Drone Imagery Nov 2022 - Feb 2023

- Built object detection system for aerial drone imagery using YOLOv8, PyTorch, Detectron2; focused on identifying randomly oriented objects at scale.
- Implemented data augmentation and rotation-invariant training; achieved high-speed, precise detection suitable for real-world aerial applications.
- Optimized model for real-time processing; validated through rigorous field tests, demonstrating operational readiness for deployment.