# TARUN REDDY THANDU

☎ +1 (857) 693-4743 ✉ tarutornado@gmail.com LinkedIn 🌐 Portfolio

## Professional Summary

Machine Learning Engineer with expertise in GenAI, NLP, LLMs, and computer vision, delivering scalable AI solutions in healthcare and enterprise domains. Proven ability to fine-tune and deploy advanced models using Python, PyTorch, Hugging Face, and AWS to drive measurable business impact.

## Technical Skills

| | |
|---|---|
| Languages and Frameworks | : Python, PyTorch, TensorFlow, Keras, Scikit-learn, FastAPI, Git |
| NLP and GenAI | : Hugging Face Transformers, spaCy, RAG (LangChain, Haystack), Entity Extraction |
| MLOps and Cloud | : AWS (Lambda, Fargate, S3, RDS, DynamoDB, OpenSearch), DVC, MLflow, RESTful APIs |
| Databases and Search | : Milvus, FAISS, Qdrant, Pinecone, Weaviate, Elasticsearch, PostgreSQL, MongoDB |

## Experience

**HerHeard** | Full-Stack AI Engineer Intern **Sep 2024 - Dec 2024**
FastAPI, LangChain, Milvus, FAISS, AWS, DVC, MLflow, Haystack Cambridge, MA

- Developed and deployed a HIPAA-compliant GenAI conversational assistant for patient-provider support, integrating **LangChain** with healthcare-adapted **Mistral-7B** and **GPT-4 Turbo** models; leveraged **Milvus v2.4** for high-throughput **RAG** over clinical notes and guidelines, achieving a 35% boost in patient query accuracy and 28% faster provider response.
- Engineered a multi-stage **RAG** pipeline with **FAISS** for prototyping and **Milvus** for production, orchestrated via **LangChain Expression Language (LCEL)**; evaluated retrieval consistency using **RAGAS** and **DeepEval** to minimize hallucinations.
- Fine-tuned **LLMs** (**Mistral-7B**, **Llama-3-8B-Instruct**) on de-identified clinical data using **Axolotl** with **LoRA** and **QLoRA** adapters, validating outputs with **OpenAI Evals** and custom medical safety metrics for domain compliance.
- Designed event-driven backend microservices with **FastAPI**, deploying on **AWS Lambda** and **Fargate**; managed model/data versioning with **DVC** and **MLflow**, and implemented **Haystack 2.0** and **Whisper-v3** for real-time ingestion and multimodal retrieval from PDF, HL7, FHIR, and audio sources.

**Infosys** | Specialist Programmer **Sep 2021 - Jan 2023**
Python, Transformers, spaCy, Elasticsearch, FastAPI, AWS, SQL, AWS Glue, Terraform, Docker, Kubernetes Bangalore, India

- Developed and productionized NLP pipelines using **Python**, **Hugging Face Transformers** (**BERT-base**, **RoBERTa-base**), and **spaCy v3** for large-scale document classification and NER, automating compliance checks and data extraction from contracts, invoices, and support tickets—including those in the finance and insurance sectors—reducing manual processing effort by 40%.
- Orchestrated and automated data workflows using **SQL** and **AWS Glue** to process high-volume documents and drive downstream analytics, supporting regulatory compliance and reporting needs.
- Containerized NLP microservices with **Docker** and deployed scalable applications on **Kubernetes** clusters; configured cloud infrastructure using **Terraform** for automated, robust CI/CD and seamless integration with enterprise systems.
- Built and deployed semantic search and automated ticket routing services leveraging **SentenceTransformers**, **Elasticsearch**, and **DistilBERT**; improved retrieval relevance by 30% and cut helpdesk response times by 33%.

**VIGIL Lab @ IIT Hyderabad** | Computer Vision Research Intern **Mar 2021 - Aug 2021**
PyTorch, TensorFlow, Keras, OpenCV, Git, Weights and Biases Hyderabad, India

- Engineered a memory-efficient object detection pipeline for aerial images using a custom **YOLOv4-OBB** model in **PyTorch**, reducing model size by 30% and improving detection accuracy by 12% on **DOTA** and **VisDrone** benchmark datasets.
- Researched and benchmarked deep learning architectures—including dilation networks, attention modules (**SE**, **CBAM**), and **EfficientNet** variants—across **PyTorch**, **TensorFlow**, and **Keras**, leveraging transfer learning and mixed-precision training to enhance feature extraction efficiency by 18%.
- Developed advanced data preprocessing and augmentation pipelines with **OpenCV** and **torchvision.transforms**, automated evaluation with **Scikit-learn** and **TensorBoard**, and managed experiment reproducibility using **Git** and **Weights & Biases**.

## Projects

**Domain-Specific LLM Alignment via RLHF** **Feb 2024 - Apr 2024**
Python, Hugging Face Transformers, PyTorch, RLHF, Weights and Biases

- Fine-tuned and aligned a language model for **biomedical knowledge extraction** using **Reinforcement Learning from Human Feedback (RLHF)**, leveraging expert-annotated datasets to calibrate LLM responses for target discovery and literature triage.
- Designed and implemented a prompt engineering and feedback loop for iterative model refinement, integrating **conformal prediction** and uncertainty quantification techniques to ensure factual accuracy, reduce hallucinations, and improve scientific retrieval precision.

## Education

**Khoury College of Computer Sciences, Northeastern University** | Boston, MA **Jan 2023 - May 2025**
Master of Science in Artificial Intelligence - Head TA for Natural Language Processing **GPA : 3.96/4**
Courses: ML, NLP, LLMs, AI for Human Computer Interaction, Deep Learning, Reinforcement Learning