(571) 547-9928
nikhilreddy.kodumuru@gwu.edu

Nikhil Reddy Kodumuru

in/Nikhil-Reddy-Kodumuru
Arlington, Virginia
github.com/Nikhilreddy024

EDUCATION

George Washington University	MS in Computer Science - GPA: 3.90	08/2023 – present
Vasavi College of Engineering	Bachelor's in Information Technology - GPA: 8.19	05/2019 – 04/2023

EXPERIENCE

Integrative Nucleic Acid Frameworks -GWU — Research Assistant, Washington DC 08/2024 – present

- Developed models for conditional generation of various RNA types for medical research using graph transformer and diffusion models, including a specialized RNA-focused version of AlphaFold that improved structure prediction accuracy by 30%.
- Designed sequence and structure-aware RNA generative models with an optimized encoding strategy for a GPT-like architecture, reducing processing time by 70% compared to conventional methods, leveraging PyTorch, clustering analysis, and AWS SageMaker.

Brane Enterprises — ML Engineer 03/2023 – 6/2023

- Engineered and fine-tuned a facial recognition system for authorization, utilizing MTCNN for face detection and ArcFace for embedding extraction. Enhanced detection accuracy by 15% through model architecture and training pipeline optimizations.
- Implemented a computer vision system using PyTesseract and OpenCV to detect specific features at warehouse turns, enabling autonomous right turns. Successfully executed the project from data collection to deployment on an NVIDIA Jetson Nano in just 10 days.
- Conducted Research in reinforcement learning for autonomous vehicles and explored SLAM (Simultaneous Localization and Mapping) for mobile robot navigation, advancing self-driving capabilities in dynamic environments.

Edmound Software Solutions — Jr. ML Engineer, India 02/2022 – 10/2022

- Created an end-to-end NLP pipeline using Hugging Face Transformers and MLflow for sentiment analysis, processing daily customer reviews with 85% accuracy and automated model retraining workflows.
- Built and maintained data preprocessing pipelines using Apache Airflow and pandas, handling 5GB+ of daily user interaction data, resulting in a 50% reduction in feature engineering time and improved model training efficiency.
- Containerized legacy ML pipelines using Docker multi-stage builds and implementing health checks for improved reliability in Kubernetes clusters.

PROJECTS

Project EGRET: Emotion-Aware Conversational AI — Gen AI Project 07/2024 – 12/2024

- Designed EGRET, an innovative emotion-aware AI chatbot leveraging LLaMA-3B and a custom GCN architecture, achieving a 57% improvement in perplexity (14.04 vs. 32-35 baseline) through optimized fine-tuning and emotion mapping.
- Developed a multi-task learning framework incorporating real-time emotion classification, enhancing BLEU-4 scores by 57.1% (0.1257 vs. 0.0800) compared to Facebook AI's baseline using the EmpatheticDialogues dataset; the state-of-the-art results led to a paper submission at IEEE CAI 2025 (currently under review).
- Built a GCN-based emotional context tracking system with LoRA optimization, effectively processing 2,000 conversations and achieving a 0.4101 ROUGE-L score while maintaining minimal computational overhead.

Project InterACT — Deep learning Project (Multimodal AI) 01/2024 – 04/2024

- Designed an advanced Multimodal AI system integrating object detection with an LLM-powered conversational agent to enable real-time task automation.
- Managed the complete ML pipeline, from data preprocessing to model deployment, achieving 94% accuracy in object detection using YOLO-World and fine-tuning Llama-2-7B for improved conversational performance.
- Enhanced efficiency by reducing resource consumption by 15% and increasing conversational accuracy by 12% through fine-tuning with QLoRA and PEFT techniques, successfully deploying the model on AWS EC2.

SKILLS

Programming:	Python, Java, C, SQL, Shell Scripting, Matlab, Algorithm Design, Data Structures.
Specializations:	Generative AI, Natural Language Processing, Computer Vision, Reinforcement Learning.
Frameworks & lib's:	PyTorch, TensorFlow, Keras, Jax, Langchain, Transformers, crewAI .
Tools & Cloud:	Docker,Kubernetes, Git, WandB, Grafana, AWS, Google Cloud Platform(Vertex AI, Compute Engine, Vision API)

CERTIFICATIONS

1. Deep learning Certificate — Issued by Deeplearning.ai 04/2022
2. Google Cloud Certifications — Issued by Coursera 12/2022
3. Cloud Practitioner — Issued by AWS 11/2023

ACHIEVEMENTS/EXTRACURRICULARS

- Authored publication on Mechanistic Interpretability in Large Language models, presented at ICSADL 2025, contributing to advancements in AI Governance.
- SEAS Merit Based Scholarship – George Washington University, 50% Scholarship (awarded to top 10% students)
- Secured 2nd place in a university hackathon by developing an LLM-powered Chain-of-Thought RAG model on graph-based physics academic data using Claude-Sonnet..