NIKHIL REDDY POTTANIGARI

Mobile: +14387227132 Mail: nikhilreddy3888@gmail.com | Github: nikhilreddy3888 | Portfolio: nikhilreddy | Linkedin: Nikhil Reddy

EDUCATION

MILA - QUEBEC AI INSTITUTE, MONTREAL (Affiliated with UdeM/McGill) *Graduated: Dec 2024*
Master's in Computer Science with **Machine Learning** Specialization; Grade: **4.2/4.3**

NATIONAL INSTITUTE OF TECHNOLOGY - WARANGAL *Graduated: May 2021*
Bachelor of Technology, **Computer Science & Engineering**;

EXPERIENCES

ServiceNow – Montreal, Canada
Applied Research Scientist *May 2024 – Present*

- Built a Document Intelligence agent within the ServiceNow Agentic Framework enabling automated summarization, Q&A, information extraction, classification, and cross-document reasoning across enterprise documents.
- Implemented and benchmarked multimodal vector search pipeline using text, vision & multimodal embedding models, supporting 10+ languages and 20+ document types, and deployed scalable Kubernetes microservices for document parsing, image captioning and content embedding.
- Collaborated with NVIDIA on dataset curation, fine-tuning, and benchmarking for their document based models spanning OCR, retrieval, and Q&A tasks, contributed to evaluation frameworks across these model variants.
- Transitioned to the Multimodal Services team, extended the platform to audio and video use cases, built efficient pipelines for embedding generation, Q&A, and content analysis of audio/video content.
- Engineered semi-supervised labeling pipeline using HDBSCAN clustering, reducing manual annotation by 60%.
- Led fraud document detection research: five-agent modular AI system (layout, semantic, anomaly, tampering, metadata) achieving 92% accuracy on generated synthetic fraud datasets using LaTeX-to-PDF and Diffusion models.

AI Safety & Red Teaming

- Contributed to LLM red-teaming and adversarial testing on multimodal document agents by simulating prompt-injection, jailbreak, and data-exfiltration attacks, including hidden instruction attacks embedded in PDFs and images.
- Added AI safety guardrails using prompt hardening, input sanitization (prompt-injection detection) and adversarial evaluation datasets to improve robustness of production document AI pipelines.

Cyberjustice Laboratory of Udem – Montreal, Canada
Student Researcher *Nov 2023 – May 2024*

- Built JusticeBot AI for legal query processing by fine-tuning LLaMA 2 on Canadian legal corpus via QLoRA; implemented RAG pipeline with custom legal embeddings improving retrieval accuracy by 40%.
- Co-authored two research papers on LLM evaluation for mediation and AI-based dispute resolution, published npm package 'text-moderate' for content moderation.

ServiceNow – Hyderabad, India
Software Engineer - II *June 2021 – Aug 2023*

- Designed DocChat, an enterprise document retrieval system using BERT embeddings and Pinecone vector database for natural language query processing.
- Built an unsupervised recommendation engine using text embeddings, reducing average ticket resolution time by 23%, developed Recommended Actions Framework with React, Java, and GraphQL.

SELECTED PROJECTS

LLM Augmented Mixture of Experts (LLM-Aug-MoE)

- Hybrid architecture combining Zephyr-1.6B with task-specific expert models via cross-attention; improved multilingual performance on low-resource languages by 28%, cutting compute cost 60% via selective layer freezing.

Vision Augmented Large Language Models (VA-LLM)

- Multimodal system integrating CLIP ViT with StableLM via cross-attention for joint text-image reasoning; Dockerized ML workflow with Hugging Face Hub integration for reproducible research.

NHL Expected Goals (xG) Prediction & Live Deployment | Université de Montréal

- Real-time ML pipeline with Dockerized Flask APIs, Streamlit dashboard, live NHL API integration, and CometML model registry for dynamic model hot-swapping.

PUBLICATIONS

- **Bridging the Modality Gap: Enhancing Document Retrieval with Multimodal Embeddings** — Master's Thesis, MILA / Université de Montréal, 2024
- **Let's Evaluate Step-by-Step: A Robust Evaluation Method of LLM Applications** — *2nd Workshop on Generative AI and Law, GenLaw '24 (Submitted)*
- **Robots in the Middle: Evaluating LLMs in Dispute Resolution** — *JURIX 2024 (Accepted)*
- **Leveraging AI for Natural Disaster Management: Takeaways from the Moroccan Earthquake** — *AI for Humanitarian Assistance Workshop, NeurIPS 2023 (Accepted)*
- **Efficient Detection of Disguised Faces from Low-Quality Surveillance Footage** — *IEEE FG 2024 (Accepted)*