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) *Expected: Graduation May 2025*
Master's in Computer Science with **Machine Learning** Specialization; Grade: **4.2/4.3**
Relevant Coursework: ***Deep Learning, Natural Language Processing, Generative Models, Data Science***

NATIONAL INSTITUTE OF TECHNOLOGY - WARANGAL *Graduated: May 2021*
Bachelor of Technology, **Computer Science & Engineering**;
Relevant Coursework: ***Algorithms, Object Oriented Programming, DBMS, Software Engineering***

EXPERIENCES

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

- Designed and deployed a scalable end-to-end **Multilingual MultiModal Vector Search** pipeline using state-of-the-art embedding models (e.g., **PyTorch**/**Hugging Face**) and **Milvus**, enhancing document retrieval for diverse data modalities.
- Built a **semi-supervised dataset** leveraging minimal labeled data, embeddings, and **HDBSCAN clustering**, improving model performance with unsupervised insights. And developed **Gradio**-based UI scripts for model evaluation and easy experimentation, enabling quick iteration on vector search and embedding benchmarks.

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

- Contributed to JusticeBot AI for legal queries by training on Canadian legal data to generate contextually accurate responses.
- Implemented **Retrieval Augmented Generation (RAG)** pipelines, fine-tuned **LLaMA2** with **QLoRA**, and leveraged **LLaMA indexing** to boost retrieval accuracy by **40%**
- Applied legal text vectorization and improved retrieval functions, enabling prosecutor-style response generation.

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

- Designed **DocChat** for streamlined document retrieval, using **BERT models**, **Pinecone** vector database, and **LLM prompt engineering** to enable efficient, natural language query processing
- Developed an unsupervised recommendation system using **word2vec** and **GloVe** embeddings, reducing ticket resolution time by **23%** via similar ticket recommendations
- Built a Recommended Actions Framework, integrating **ML** and **NLP** models across the ServiceNow ecosystem using **React**, **Java**, and **GraphQL**, resulting in streamlined organizational processes and real-time suggestions

MathWorks EDG Intern – Hyderabad, India *April 2020 – June 2020*

- Implemented a scalable clipboard management solution for MATLAB's browser-based version, enabling robust multi-browser support for diverse data types in clipboard actions.

PUBLICATIONS

Efficient Detection of Disguised Faces using Photos/Sketches from Low-Quality Surveillance Footage Link
In: 18th **IEEE** International Conference on Automatic Face and Gesture Recognition 2024 (Accepted)

Leveraging AI for Natural Disaster Management : Takeaways From The Moroccan Earthquake Link
In: Artificial Intelligence for Humanitarian Assistance and Disaster Response Workshop, **NeurIPS** 2023 (Accepted)

SELECTED PROJECTS

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

- Developed an augmentation architecture combining a general-purpose **LLM** (**Zephyr 1.6B**) with task-specific LLMs via **cross-attention layers**, improving efficiency and multilingual performance for low-resource languages. Conducted experiments with partial LLM freezing (training only cross-attention layers/projection matrices) to enable efficient fine-tuning
- Also integrated a smaller **CLIP** component for initial vision capabilities, training it on a captioning dataset to demonstrate cross-modal understanding

Pairwise-MMMF (Recommender Systems), Bachelor's Thesis under **Prof. Venkateswara Rao Kagita** Link

- Derived an algorithm with a new loss function and gradients, optimized on hinge loss of pairwise distance between points which reduced the bias problem of Hierarchical Matrix Factorization algorithm.
- Implemented the algorithm with metrics like **ZOE**, **MAE** and **RMSE** in Python, and worked with MovieLens 100K Dataset.

MetaMenu – Hyderabad, India Link
Co-Founder *Jan 2022 – August 2023*

- Developed an **augmented reality** app to showcase items in an interactive way from our digital menu app
- Built Full stack web application using **React** for front-end, **NodeJS** for backend and **Firebase** as database, and hosted it in **Google Cloud.** Application served more than **100,000** users

SKILLS

Technical: Python, C++, Java, JavaScript, React, TensorFlow, PyTorch, Transformers, LangChain, QLoRA, RAG, Vector Databases (Milvus/Pinecone), SQL, Hadoop, Docker, Kubernetes, Jenkins, MLflow, AWS, GCP, Azure, Linux, Gradio