Vijay Murari Tiyyala

mleng.nlp@gmail.com | +12405711678 | LinkedIn Profile | Personal Website | Baltimore

Education

Master's in Computer Science - ***Johns Hopkins University,*** *Baltimore* Dec 2023
Focus: Machine Learning, Data Science, NLP, Databases, Information Retrieval, Statistics

Bachelor of Technology in Computer Science - ***VR Siddhartha Engineering College,*** *India* Jun 2021

Technical Skills

- Programming Languages: Python, Java, R, C++, C
- Frameworks and Libraries: PyTorch, TensorFlow, Keras, Langchain, HuggingFace, Deepspeed, Scikit-learn, MLflow
- Tools and Platforms: Docker, AWS, Azure, GCP, Git, Kubernetes, PowerBI, Airflow, Weights & Biases
- Data Management: SQL, NoSQL, PostgreSQL, Apache Solr, Apache Spark, Hadoop, Elasticsearch
- Misc: HTML/CSS, PHP, Linux, Shell Scripting, Distributed Computing, CI/CD

Work Experience

NLP Researcher - ***Johns Hopkins University,*** *Full-Time* Aug 2023 – Present
- Engineered an empathetic medical chatbot using **LlaMA3**, boosting response accuracy to **88.7%** on a human-annotated test dataset, enhancing patient interaction quality.
- Reduced training time by **50%** by leveraging **PyTorch/SLURM** in a **multi-GPU** environment for efficient **distributed** training.
- Leveraged **Apache Solr Cloud** for indexing and retrieval of **2.5TB** of textual data, optimizing compute and access times.
- Enhanced model empathy and factuality through **Direct Preference Optimization (DPO)/RLHF** training.
- Facilitated seamless model deployment to **AWS** using **Docker**, ensuring scalable and reliable access.

NLP Research Intern - ***Center for Language and Speech Processing,*** *Full-Time* Jun 2023 – Sep 2023
- Led **RAG** chatbot development & integration with **Apache Solr Cloud** to achieve rapid data indexing and retrieval, significantly reducing user search time by **70%**, and boosting web traffic by **40%**.
- Achieved a **50%** reduction in compute costs by using **PEFT**, **LoRA**, and **QLoRA** for efficient **LlaMA2** training and quantization.
- Optimized document retrieval recall to **90%** by integrating re-ranking and chunk summarization, optimizing search result relevance.
- Managed end-to-end chatbot deployment using **Docker** and **FastAPI**.

Graduate Research Assistant - ***Johns Hopkins University,*** *Part-Time* Jan 2023 – Jun 2023
- Improved machine translation accuracy to **86%** for medical terminologies in low-resource languages, improving accessibility.
- Analyzed **15,000+** compound words, creating a model to improve English translations.
- Designed a **300+** language translation pipeline, enhancing term reconstruction with compound splitting algorithms.

Business Technology Analyst - ***Deloitte USI,*** *Full-Time* Jul 2021 – Jun 2022
- Developed stored procedures and scripts for integrating clients' tax data via APIs, and visualized analytical insights in **PowerBI**.
- Accomplished a **20%** reduction in tax data processing time by refining **SQL** procedures for optimization.
- Boosted client retention by **30%** through improved analytics and reporting, by collaborating with various teams in analyzing and deploying data solutions.

Publications

Published

1. **Kreyòl-MT: Building MT for Latin American, Caribbean, and Colonial African Creole Languages**, *NAACL 2024*.

Under Review

1. **ANALOBENCH: Benchmarking the Identification of Abstract and Long-context Analogies**, submitted to *ACL 2024*.

Projects

Cannabis Use Detection in Clinical EMR - ***Python, PyTorch, Git***
- Trained NLP models such as BERT, RoBERTa, and ClinicalBERT to increase detection accuracy of cannabis use in EHRs by 97%
- Achieved 92% accuracy in distinguishing medicinal and recreational cannabis use from unstructured text, enhancing data quality.
- Collaborated with clinical researchers to validate model outputs, ensuring compliance with HIPAA and high data fidelity.

Adverse AI: Automated Discovery of Adverse Event Reports from Unstructured Text - ***Python, PyTorch, Git***
- Led the development of 'Adverse AI', achieving 97.5% accuracy in identifying adverse events from diverse text sources including medical reports, social media.

- Automated extraction and analysis of adverse event data by training models like BERT and RoBERTa reducing manual review time by 90%.
- Open-sourced the tool to enable widespread adoption and continuous improvement by the healthcare community.

SAMOYEDS - ***Python, PyTorch, HuggingFace, Git, Flask, HTML/CSS, JavaScript***
- Led the design and development of the SAMOYEDS application, a policy simulation tool using LLMs focusing on public health.
- Enabled SAMOYEDS to simulate diverse human personas, predicting public health policy responses with **76% accuracy**, enhancing policymaker decision-making.

Benoit: Better English Noisy Audio Transcripts - ***Python, PyTorch, TorchAudio, TorchText, Colab***
- Developed a grammar-correcting ASR model for non-native English speaker audio.
- Created synthetic dataset by back-translating English sentences from a low-resource language and passing them to Microsoft SAPI5 TTS to create a proxy for non-native English audio.
- Used a GRU-based seq2seq denoising autoencoder on top of a pre-trained Wav2Vec 2.0 (frozen) for grammatically correct ASR.

ResearchNavigator - ***Python, PyTorch, HuggingFace, Git, HTML/CSS, JavaScript***
- Created an AI information retrieval system/search engine with an interface for research papers, utilized **LDA** for clustering, and LLMs to generate summaries.

Code Editing via Natural Language Instructions - ***Python, PyTorch, HuggingFace, BeautifulSoup, Git, SLURM***
- Improved code editing by Instruction-tuning **CodeLlama2**, achieving a **37% pass@1 accuracy** in interpreting natural language instructions, significantly streamlining the coding workflow.