Vijay Murari Tiyyala

✉ mleng.nlp@gmail.com | ☎ +12405711678 | vijaymuraritiyyala | iMvijay23 | Baltimore

Technical Skills

- Programming Languages: Python, Java, R, C++
- Model Training and Deployment: PyTorch, TensorFlow, HuggingFace, SLURM, Deepspeed, Spark, Docker, Kubernetes
- Data Management and OS: SQL, Apache Solr, Airflow, Linux, Shell Scripting, Git, PowerBI
- Web and Cloud Technologies: HTML/CSS, PHP, GCP (including AI/ML services), Azure, AWS

Education

Master's in Computer Science - ***Johns Hopkins University,*** *Baltimore* Graduated – Dec 2023

Focus: Machine Learning, Data Science, NLP, Databases, Information Retrieval, Statistics

Work Experience

NLP Researcher - ***Center for Language and Speech Processing*** Aug 2023 – Dec 2023

- Developed an empathic medical chatbot using Llama2, enhancing patient interactions. Utilized Pytorch/SLURM for model training and fine-tuning in a multi-GPU environment.
- Implemented Apache Solr for data indexing and retrieval. Implemented Direct Preference Optimization (DPO)/RLHF for human-preference alignment.
- Contributed to systems design for AI application, focusing on efficient data and ML pipelines.

NLP Research Intern - ***Center for Language and Speech Processing*** Jun 2023 – Aug 2023

- Led Retrieval Augmented Generation (RAG) chatbot project, employing LLMs for real-time querying. Enhanced Apache Solr Cloud integration for faster data indexing.
- Implemented Parameter-Efficient Fine-Tuning (PEFT), LORA, and QLORA for Llama2, significantly reducing compute costs.
- Deployed chatbot using Docker and FastAPI, showcasing skills in end-to-end application development and deployment.

Multilingual Machine Translation Researcher - ***Center for Language and Speech Processing*** Jan 2023 – Jun 2023

- Initiated a major project to improve machine translation of medical terminologies, aiding access to healthcare information in low-resource languages
- Automated web scraping tools to compile a database of 15,000+ medical terms.
- Developed a robust translation pipeline for 300+ languages, utilizing advanced compound-splitting algorithms, boosting translation accuracy by 25%.

Business Technology Analyst - ***Deloitte USI*** Jul 2021 – Jun 2022

- Managed and optimized SQL databases, developing visualization tools and API data scripts, enhancing operational efficiency.
- Improved SQL-based stored procedures, achieving a 20% reduction in tax data processing times.
- Facilitated cross-departmental collaboration, implementing enterprise-level data solutions, contributing to a 30% increase in client retention.

Technical Projects

SAMOYEDS - ***Simulating Agents for Modeling Outcomes and Estimations to Direct Social-policy***

- Spearheaded the design and development of the SAMOYEDS application, a policy simulation tool focusing on public health.
- Implemented a server-client architecture using Flask for efficient data transfer between the Mistral 7B and the frontend.
- Developed a dynamic user interface to visualize simulation outcomes, enhancing the tool's usability for policy makers.

CLSP-ResearchNavigator - ***An AI-Enhanced Academic Repository***

- Engineered a full-stack web application for efficient querying of a database encompassing over 3,000+ CLSP research papers.

CodeTalk - ***Code Editing via Natural Language Instructions***

- Compiled a dataset for language model training, aimed at assisting in code editing tasks. Utilized Dijkstra's algorithm for identifying related codes, enhancing data quality for CodeLlama fine-tuning.

ImageClassify-VT - ***Unsupervised Image Classifier with Vision Transformers***

- Crafted an innovative image classification model using Meta's DINOv2, attaining a remarkable 86% ImageNet accuracy.

Publications

NAACL'24 - ***Multilingual Machine Translation Paper Under review***