Vijay Murari Tiyyala

Contact Email | Contact Phone | LinkedIn Profile | Personal Website | Baltimore

Technical Skills

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

Education

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

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

Work Experience

ML Researcher - ***Johns Hopkins University*** Aug 2023 – Present
- Developed an empathic medical chatbot using LlaMA2, enhancing patient interactions. Utilized Pytorch/SLURM for distributed model fine-tuning in a multi-GPU environment.
- Utilized Apache Solr Cloud for large-scale data indexing and retrieval thereby reducing compute and retrieval time.
- Aligned the model by using Direct Preference Optimization (DPO)/RLHF training.
- Facilitated seamless model deployment to AWS using Docker, ensuring scalable and reliable access.

ML Research Intern - ***Center for Language and Speech Processing*** Jun 2023 – Aug 2023
- Led RAG chatbot development, integrating LLMs with Apache Solr Cloud to achieve rapid data indexing and retrieval, significantly reducing query response times and lowering compute resource demands.
- Achieved a 50% reduction in compute costs by using PEFT, LoRA, and QLoRA for efficient LlaMA2 training and quantization.
- Enhanced response accuracy by employing advanced re-ranking techniques to ensure the retrieval of the most relevant documents.
- Managed end-to-end chatbot deployment using Docker and FastAPI.

Graduate Research Assistant - ***Johns Hopkins University*** Jan 2023 – Jun 2023
- Enhanced machine translation for medical terminologies in low-resource languages, improving accessibility.
- Analyzed compound word translation statistics across languages from over 15,000 processed terms.
- Developed a translation pipeline for 300+ languages, utilizing compound-splitting algorithms to reconstruct lost or unknown terms.

Business Technology Analyst - ***Deloitte USI*** Jul 2021 – Jun 2022
- Developed stored procedures and scripts for API data integration, optimized queries, and visualized insights in PowerBI.
- Achieved 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.

Projects

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.
- Implemented a server-client architecture using Flask for efficient data transfer between the Mistral 7B and the front end.
- Developed a dynamic user interface to visualize simulation outcomes, enhancing the tool's usability for policymakers.

ResearchNavigator - ***Python, PyTorch, HuggingFace, Git, HTML/CSS, JavaScript***
- Developed an AI 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***
- Created Codeforces data for problem submissions, preprocessed them into input-output pairs and Instruction-tuned CodeLlama.
- Enabled CodeLlama to accurately interpret natural language instructions for code editing, enhancing its user interaction.

Grammar Autocorrected ASR: Enhancing Transcription Accuracy - ***Python, PyTorch, NLTK, Git***
- Achieved over 93% grammatical accuracy in transcriptions from noisy English audio by training NLP auto-correction model.

Publications

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