Dwija Parikh

+1 (713) 715-9407
dwija@uw.edu | linkedin.com/in/dwija-parikh | dwijap.github.io

EDUCATION

University of Washington
M.S. Computational Linguistics — Seattle, WA
Coursework: Shallow & Deep Processing for NLP, Machine Translation, Syntax Engineering, Phonetics

University of Houston
B.S. Computer Science, B.S. Mathematics (Data Science Option) — Houston, TX
Coursework: Linguistics, Advanced NLP, Data Structures & Algorithms

TECHNICAL SKILLS

Programming Languages: Python, R, C/C++
Tools & Frameworks: PyTorch, Tensorflow, Huggingface, Scikit-learn, Pandas, NLTK, SpaCy, Praat
Data Operations: SQL, MongoDB, NoSQL, Tableau
Development: AWS, CUDA, Docker, CI/CD, Git
Languages: English, Hindi, Gujarati

EXPERIENCE

NVIDIA — Sep 2023 - Present
NLP Engineer — Remote

- Built and optimized human-in-the-loop **annotation data pipelines** for diverse models like **ASR/TTS (Riva), LLMs (NeMo), and vision language models (VILA, Cosmos)** supporting **1k+ members**, ensuring scalability and high-quality data preparation
- **Collaborated cross-functionally with research and product teams** to evaluate, refine, and enhance multimodal models and products aligning technical deliverables with customer needs across **50+ projects**
- Designed and implemented **synthetic data generation techniques** for Retrieval-Augmented Generation (RAG) in the NeMo LLM service, leveraging both rule-based systems and LLMs to enhance downstream performance
- Provided **linguistic expertise for dataset evaluation,** leveraging customized **agreement metrics** and **quality scores** tailored to specific use cases, enabling data-driven decision-making for stakeholders

Hewlett Packard Data Science Institute — Jun 2021 – Aug 2021
Data Science Intern — Houston, TX

- Developed **data pipelines and workflows** to process over **1 million data points** in healthcare insurance claims, optimizing ovarian and prostrate cancer treatment pathways for over **65k patients**
- Applied unsupervised learning to uncover trends in patient pathways using **Directed Acyclic Graphs (DAGs),** community detection, and predictive link analysis for treatment visualization

PROJECTS

Product Recommender System Using Graph Neural Networks

- Developed an **ontology based** product recommendation system leveraging **Graph Neural Networks (GNNs**) to enhance recommendation accuracy, achieving a 72% accuracy in predictive performance
- Integrated **knowledge graphs** to enrich product data representation, enabling better semantic understanding and enhancing downstream machine learning task
- Constructed a comprehensive ontology that captured **domain-specific knowledge** and relationships between products, enabling an understanding of user preferences and item characteristics thereby improving customer engagement

PUBLICATIONS

Targeted Multilingual Adaptation for Low-resource Language Families, *EMNLP Findings 2024*
C.M. Downey, Terra Blevins, Dhwani Serai, **Dwija Parikh**, Shane Steinert-Threlkeld

- Adapted XLM-R for **low-resource language families**, improving performance and accuracy on POS tagging & dependency parsing through targeted multilingual training strategies and evaluated hyperparameters to enhance performance across **15+ languages**
- Identified key hyperparameters through regression analysis, establishing best practices for up-sampling low-resource languages without compromising high-resource language performance

Normalization and Back-transliteration for Code-Switched Text, *CALCS @ NAACL 2021*
Dwija Parikh and Thamar Solorio

- Developed a preprocessing module specifically designed for **code-switched** data, utilizing a hybrid approach that combined rule-based phonemic transcription methods with machine learning techniques, including a **seq2seq** model employing **LSTM** networks, resulting in an accuracy rate of **78.6%** and **0.8 BLEU score**
- Engineered a novel **grapheme-to-phoneme (G2P)** conversion technique specifically tailored for Romanized Hindi data, enhancing the processing and analysis of code-switched text in social media contexts