Aadit Deshpande

linkedin.com/in/aadit-deshpande/
Email : aaditd@andrew.cmu.edu
Mobile : (717)-917-9082

Education

Carnegie Mellon University — Pittsburgh, PA
Master's Degree, LTI, School of Computer Science — ***GPA: 4.17/4.3*** — *Expected Dec 2024*
- **Courses** : Machine Learning, Advanced NLP, Speech Processing, Search Engines, Generative AI, Linguistics

Birla Institute of Technology and Science, Pilani — Pilani, India
Bachelor of Engineering Computer Science — ***GPA: 9.64/10*** — *May 2023*

Experience

Siemens Digital Industries Software — Cincinnati, OH
Software Development Intern, Product Engineering Software — *May 2024 - Aug 2024*
- Developed a Video Retrieval Augmented Generation (RAG) chatbot for Siemens NX using Agile methodology.
- Improved Vector DB ranking (ChromaDB) MAP by **5** points by switching to the text-ada-002 embedding model.
- Augmented retrieval index with QA pairs and video scene descriptions using **Claude-3-Sonnet** (AWS Bedrock).
- Identified 3 most common RAG failure cases using **RAGAS** evaluation suite and **GPT-4** (Azure OpenAI).
- Optimized Azure GPT-4 inference by **40%** using streaming, dynamic batching, and generation size control.

American Express, AI Labs — Bangalore, India (Remote)
Data Science Analyst Intern, Credit and Fraud Risk (CFR) — *Jul 2022 - Dec 2022*
- Augmented customer complaints platform database with **6 months** of **external data** scraped using Reddit API.
- Improved Reddit thread analysis efficiency by **30%** by optimizing the **unsupervised intent detection** codebase.
- Enhanced existing Reddit pipeline by implementing an unsupervised **aspect-based sentiment analysis** module (sentenceBERT, nltk.vader) and a novel retrieval metric for 'engagement'.
- Consolidated the Reddit insights pipeline into single internal tool for CFR, through cross-functional collaboration.

Projects

Evaluating the rhyming capabilities of Large language models (Prof. D Mortensen) — Mar 2024 – May 2024
- Constructed multilingual rhyme evaluation datasets (**5k** word pairs) using CMUDict and Celex2 data.
- Evaluated open-source LLMs (Llama2-7b, Llama3-8b, CrystalChat-7b) using vLLM with prompt engineering.
- Identified **Llama3-8b** as the best model by F1 score on the English (**0.687**) and Dutch rhyme (**0.647**) datasets.
- Proposed an error typology based on LLM hallucinations due to implicit and explicit reasoning mismatch.

Cascaded vs. End-to-End Speech-to-Speech Translation (S2ST) systems (es-en) — Feb 2024 – May 2024
- Compared real-time S2ST systems: Cascaded (pretrained **OWSM v3.1** (Speech-to-text translation) + VIT TTS (text-to-speech)), and End-to-End (discrete unit E2E ESPNet model trained on CVSS-C)
- Enabled a fairer comparison by parameter-efficient fine-tuning (**LoRA**) of the Cascaded model on CVSS-C.
- Demonstrated the Cascaded model's superior ASR performance (**17.69** vs. **14.90**) on CVSS-C Spanish-English.

Maintaining Consistency in extended LLM medical note generation (Prof. C Rose) — Feb 2024 – May 2024
- Proposed a novel architecture for turn-wise text generation using a cascade of Llama2-7b-chat-hf generator and Maximal Marginal Relevance (**MMR**) summarizer models and proposed an error ontology of its failure modes.
- Performed prompt engineering with parametric personalized medical notes to evaluate turn-wise text consistency.
- Analyzed the generations (blogs) qualitatively, and quantitatively by proposing **5** new factual plausibility metrics.

Publications

[1] Evaluating the Multilingual Rhyming capabilities of Open-Source Large Language Models *[in preparation]*
[2] Key-phrase boosted unsupervised summary generation for FinTech organization *[preprint]*
[3] An ImageJ macro tool for quantitative analysis of Myopic Choroidal neovascularization [**PLOS ONE, 2023**]
[4] Population-based AI assessment of Diabetic Retinopathy risk factors [**Ophthalmic Epidemiology, 2023**]

Technical Skills

Languages: Python, Java, C/C++, MySQL, JavaScript, HTML
Deep Learning: PyTorch, TensorFlow, Keras
Libraries: NumPy, Pandas, ESPnet, Sci-kit learn, NLTK, Spacy, Huggingface transformers, LangChain, vLLM
Tools: Slurm scheduler (GPU management), Git, Unity, LaTeX, PowerBI, Data Studio, Wandb