Aadit Deshpande

linkedin.com/in/aadit-deshpande/

Email : aadit3003@gmail.com

Education

Carnegie Mellon University Pittsburgh, PA

*Master's Degree, ML, NLP (MIIS), School of Computer Science — **GPA: 3.9/4.0*** *Dec 2024*

- **Courses** : Machine Learning, Advanced NLP, Information Retrieval, 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 Engineer, AI Platform *Feb 2025 - Present*

- Building AI features for NX, Siemens's flagship CAD tool, including an Agentic Copilot and generative models.
- Spearheaded migration of core agent service from LangChain to LangGraph, cutting task failure rate by **70%**.
- Led research and authored articles on multi-agent systems, synthetic data, and generative CAD, engaging **50+ monthly internal readers** across engineering teams.
- Automated fine-tuning of domain-specific code LLMs with GitLab CI/CD and Weights & Biases, saving **4 hrs/week** and doubling iteration speed.

Software Development Intern *May 2024 - Aug 2024*

- Developed a Video Retrieval Augmented Generation (RAG) chatbot for Siemens's flagship CAD offering - NX.
- Improved Vector DB ranking (Pinecone) 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 top-3 RAG failure modes with **RAGAS** and **GPT-4** (Azure), enabling user experience improvements.
- Optimized Azure GPT-4 inference by **40%** using streaming, dynamic batching, and generation size control.

American Express, AI Labs Bangalore, India

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.
- Consolidated the Reddit insights pipeline into single internal tool for CFR, through cross-functional collaboration.

Projects

Efficiently training a Multilingual Diffusion model Nov 2024 – Dec 2024

- Developed a multilingual TTI system (EN, FR, DE) by fine-tuning **Stable Diffusion v2.1** with LoRA adapters and **multilingual CLIP** encodings, enabling global access to image generation for non-English users.
- Achieved **27% FID** and **8.5% IS** gain over cascaded baselines; preserved scene-level detail in FR/DE prompts.
- Built a multilingual benchmark from WIT (8k/2k) with CLIPScore filtering for better caption-image alignment.
- Reduced training compute by **30%** via **LoRA**-based tuning (rank-4), updating <**10%** of model parameters.

Evaluating the rhyming capabilities of Large language models (Prof. D Mortensen) Mar 2024 – May 2024

- Mined 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.

Technical Skills

Languages: Python, Java, C/C++, SQL, TypeScript, HTML, CUDA

Deep Learning: PyTorch, TensorFlow, Keras

Libraries: NumPy, Pandas, ESPnet, scikit-learn, NLTK, spaCy, Huggingface Transformers, React, vLLM, LangChain, LangGraph, LangSmith, Autogen

Tools: Azure ML, AWS Sagemaker, Slurm scheduler (GPU management), Wandb, Git, LaTeX, GitHub Actions