Vinay Premchandran Nair

Pittsburgh, U.S. | [LinkedIn](#) | (347) 291 3124 | vinaynair@cmu.edu

EDUCATION

Carnegie Mellon University (CMU)   **Pittsburgh, U.S.**
Master of Science in Artificial Intelligence and Innovation (MSAII)   *Aug 2022 - May 2024*
- Advanced NLP, Multimodal ML, Intro to Deep Learning, Large Language Models, ML in Production

Pune Institute of Computer Technology (PICT)   **Pune, India**
Bachelor of Engineering in Information Technology (Honors in AI/ML)   *Aug 2018 - July 2022*
- Deep Learning, Database Management Systems, Theory of Computations, Data Structures

RELEVANT PROJECTS

Tip Of the Tongue (ToT) Retrieval using LLMs (Advisors: Dr. Chenyan Xiong and Dr. Daphne Ippolito)
- Designed a GPT-4 based Query Decomposition framework that improved query performance for movie search engines.
- Implemented a **Dense Retrieval Model** to increase search efficiency and accuracy in matching movie names.
- Developed a **Pointwise Reranking Mechanism** with GPT-4, independently scoring retrieved results for optimized ranking.
- Improved key metrics, raising NDCG from **10% to 23%** and Precision from **4% to 20%** in search results.

Augmenting Visual QA using Advanced Vision Language Techniques (Advisor: Dr. Louis-Philippe Morency)
- Leveraged cross attentional learning for enhanced image understanding for reasoning in VQA.
- Implemented a **hierarchical approach for dense data representation**, improving multimodal reasoning capabilities.
- Developed a technique for creating dense query-image representations for improved answer generation in a Language Model.
- Established a joint training pipeline for query tokens and image patches, improving cross-modal understanding via a Vision-Language Transformer.

Attention-Based AI Learning Companion (Advisor: Dr. Bhiksha Raj)
- Designed an attention-based **knowledge tracing** encoder to track the user's knowledge state, facilitating personalized language learning using **Duolingo's SLAM dataset** to employ enhanced User Behaviour Modeling.
- Developed an RL decoder trained on Bayesian KT to **generate customized exercises** based on the user's performance and knowledge state to achieve an RMSE score of 0.43 and an AUC-ROC of 0.77 .
- Trained a T5 LLM to generate dialogue with users, leveraging their previous history and performance, for an improved Interactive Conversational Agent that could recommend relevant exercises for the user to learn a new language.

Mini Projects
- Built and trained a GPT from scratch on C4 corpus, achieving perplexity of ~32.
- Developed and trained Generative Imaging models such as a GAN, VAE, and a Diffusion achieving FID of 31.4
- Trained ResNet and ConvNext Models (from scratch) for face classification (VGGFace dataset), achieving ~91% test accuracy and ~93% test accuracy by ensembling them.
- Trained an encoder-decoder model using CTC decoding for Speech Transcription (achieving a test Levenshtein Distance of ~3 on the LibriSpeech dataset).

EXPERIENCE

Transatlantic Reinsurance Pvt. Ltd.   **New York, U.S.**
Machine Learning Intern   *June 2023 – Aug 2023*
- Spearheaded the development of a sophisticated in-house Large Language Model QA bot, assisting the risk assessment process for underwriters.
- Conducted extensive testing and evaluation of multiple open-source models, including Falcon 7B Instruct, MPT 7B, RedPajamas, and Llama 2, to determine the most suitable model for the project's requirements.
- Successfully deployed the final application using Llama 2 via HuggingFace, optimizing its performance and usability.
- Leveraged Langchain to augment risk-assessment related data-pipelines with specialized chains for Summarization, Conversational Retrieval, and Multiple Document retrieval, resulting in reduced inference time.
- Pioneered the implementation of a feedback-mechanism to enforce PEFT, facilitating the fine-tuning of the Llama 2 model to continuously improve the bot's performance and accuracy.

Carnegie Mellon University   **Pittsburgh, U.S.**
Research Assistant   *May 2023 – June 2023*
- Developed a Retrieval QA chatbot to assist survey respondents using GPT-4, Whisper and semantic search functionality.
- Incorporated prompt engineering to retrieve public information in real time to solve doubts that survey respondents have.
- Implemented efficient data collection and preprocessing, supporting diverse formats (PDFs, audio, text, web scraping) for optimized embeddings.
- Successfully deployed the chatbot on Google App Engine with comprehensive technical documentation for easy access and maintenance.

SKILLS

- **Languages/ DBs**: Python, Java, C, C++, JavaScript, HTML/CSS, PHP, MongoDB, Firebase, ChromaDB
- **Libraries:** Langchain, Langsmith, Bitsandbytes, Pandas, Numpy, TensorFlow, PyTorch, NLTK, AutoXGBoost, Surprise, AutoGluon, Sklearn, HuggingFace, PEFT, Transformers, Ray
- **Frameworks:** Github, Github Actions, NodeJS, ReactJS, GCP, AWS, Excel, Cronjobs, Nohup, Jupyter NB, Tableau, MLFlow