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

Augmented RAG based Fin-QA using Multihop Multi-agent Interactions (Advisor: Dr. Daphne Ippolito)
- Developed BNY Mellon's **Knowledge Graph** based **RAG** using **Neo4J** and **QDrant**, achieving a **10%** improvement in **reasoned retrieval** with novel **Contriver** embeddings augmenting the graph over regular vector databases using **LlamaIndex**.
- Enhanced Financial **Multihop QA** accuracy by **2%** through Embeddings Augmented Knowledge Graph
- Boosted accuracy by **20%** in automated financial responses through **Microsoft Autogen** for **multi-agent LLM interactions**.
- Utilized **Ragas** framework for system evaluation, improving **Context Precision** by **5%** and **Faithfulness** by **20%** over GPT 3.5.

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 Multimodal Multihop Question Answering (Advisor: Dr. Louis-Philippe Morency)
- Implemented a patch-focused **Vision-Language Transformer** and **T5** answer generator, resulting in an **8%** QA accuracy gain.
- Implemented a **hierarchical approach for dense data representation**, improving multimodal reasoning capabilities by **2%**.
- Engineered **MiniGPT-4** with Vision Transformer and Llama-V2 for advanced multimodal **Chain of Thought Reasoning**

Mini Projects
- Built and trained a **GPT** from scratch on **C4 corpus**, achieving **perplexity** of ~**32**.

End to End Movie Recommendation System (Advisor: Christian Kaestner)
- Developed a **Kafka**-integrated movie recommendation system, serving **60,000+** requests with **production level** Python code.
- Enhanced user click-through rate by **15%** using **Sklearn's Surprise** for recommendations streamlined using **MLFlow**.
- Reduced data drift by **40%** with **Jenkins** pipelines, triggered with every **Git** Pull Request.
- Streamlined deployment on CMU's VMs with **Docker**; achieved **25%** faster model monitoring via **Prometheus and Grafana**.

RELEVANT EXPERIENCE

Transatlantic Reinsurance Pvt. Ltd. — **New York, U.S.**
Machine Learning Intern — *June 2023 – Aug 2023*
- Executed a **RAG-based QA** bot deployment, improving response times by **30%** and managing **load balancing** for 150+ users.
- Successfully deployed the final application using **Llama 2** via HuggingFace, optimizing its performance and usability.
- Leveraged **Langchain** with specialized chains for **Summarization, Conversational Retrieval, and Document Comparator.**
- Integrated **ChromaDB** for efficient embedding data storage and retrieval of over **1500** proprietary documents.
- Pioneered the implementation of a feedback-mechanism to enforce **Parameter Efficient Fine Tuning**.

Carnegie Mellon University — **Pittsburgh, U.S.**
Research Assistant — *May 2023 – June 2023*
- Scraped information from websites of charitable organizations using **BeautifulSoup**.
- Performed **ETL** by chunking text documents and loading **GPT-4 embeddings** into a **manual vector database**.
- Added a **manual semantic search pipeline** to retrieve information from the vector database using **cosine similarity**.
- Integrated **prompt engineering** for **RAG** to address survey respondents' queries using real-time public data.
- Deployed a GPT-4 chatbot on **Google Cloud Platform**, serving **50+ concurrent users.**

SKILLS

- **Languages/ DBs**: Python, Java, C, C++, JavaScript, HTML/CSS, PHP, MongoDB, Firebase, ChromaDB
- **Libraries:** Langchain, Langsmith, LlamaIndex, Bitsandbytes, Pandas, Numpy, TensorFlow, PyTorch, NLTK, AutoXGBoost, Surprise, AutoGluon, Sklearn, HuggingFace, PEFT, Transformers, PySpark, Ray
- **Frameworks:** Github, Github Actions, NodeJS, ReactJS, Microsoft Azure, GCP, AWS, PowerBI, Cronjobs, Nohup, AirFlow, Docker, Kubernetes, Neo4J, Databricks