Praneeth Posina

praneethposina@gmail.com | **linkedin.com/in/praneethposina**
716-468-9996 | **github.com/praneethposina** | **Portfolio Link**

EDUCATION

University at Buffalo *Buffalo, NY*
Master's in Artificial Intelligence *Aug 2023 - Jan 2025*

KL University *Hyderabad, India*
Bachelor's in Electronics and Communication Engineering *Jun 2019 - May 2023*

Certifications: Azure AI fundamentals, Google Cloud Skill Badges, AWS Machine Learning Foundations, Oracle Architect Associate.

EXPERIENCE

AI Developer *Oct 2024 - Present*
Kubed Root *Buffalo, NY*

- Designed and deployed a scalable **RAG Chatbot** integrating crop datasets, farming data, real-time weather APIs, USDA soil surveys, and personal farm device data for tailored agricultural advice to farmers.
- Implemented a vector search system using ChromaDB, Qdrant **Vector Databases, Langchain and LlamaIndex** for embedding retrieval.
- Performed **Prompt Engineering** for precise, context-aware query responses achieving over **96%** factual accuracy.
- Fine-tuned a **LLaMA 8B** model with **5,000** labeled Q&A pairs using **LoRA & PEFT** techniques, optimizing performance for production.
- Developed a robust **FastAPI** backend with **REST** and **WebSocket** APIs for real-time communication and user session management.
- Containerized the backend using **Docker** and deployed it on **AWS Elastic Beanstalk**, ensuring scalability and seamless integration with a real-time frontend chat interface to support over **50,000** users with sub-second response time.

ML Research Assistant *Dec 2021 - Jan 2023*
KL University *Hyderabad, India*

- Led **Research** on mediastinal lymph node malignancy detection, developing models that increased diagnostic accuracy to **98.2%** using **Deep Reinforcement Learning**. This work showcased the potential of reinforcement learning in critical healthcare applications.
- Authored two **IEEE-Published Papers** [1] [2], introducing innovative methods that outperformed existing models by upto **4%** accuracy.
- Engineered **Custom Deep Learning Algorithms**, including a DQN policy and an ensemble learning strategy.
- Achieved a **98.56%** accuracy rate by setting a **New Benchmark** in medical diagnostic systems.
- Analyzed extensive **Medical Image Datasets**, ensuring robust model performance and contributing to significant advancements in medical diagnostics, with practical implications for real-world healthcare applications.

PROJECTS

Customer Support Chatbot [Link]

- Engineered and deployed an end-to-end customer support chatbot using the **LLaMA 3.1 8B** model with **LoRA Fine-Tuning** and 4bit, 8bit, and 16bit **Quantization** for optimized inference via **Ollama** and **Flask API**.
- Leveraged **Docker** for containerization and **AWS ECS** with Fargate to deliver a scalable, highly available solution.
- Automated **CI/CD Pipelines** with AWS CodePipeline for seamless deployment, while implementing comprehensive **Monitoring** and logging through AWS **CloudWatch** to ensure performance and reliability.

LLM Powered Mobile Assistant [Link]

- Developed an LLM mobile assistant using a fine-tuned **LLaMA** model with **Agentic Workflow** to achieve complex task execution.
- Leveraged Appium for real-time app UI analysis and **Action Automation**, enabling dynamic interaction and error recovery.
- The assistant adapts and navigates unfamiliar apps seamlessly with over **90%** accuracy.

Wikipedia Chatbot [Link]

- Built a Retrieval-Augmented Generation (RAG) chatbot integrating **Web Scraping, Indexing,** and **Query Handling**.
- Leveraged Sentence Transformers for embedding **60,000+** Wikipedia documents, TF-IDF and **Cosine Similarity** for retrieval and re-ranking, and **OpenAI GPT API** for precise, context-aware responses.

Text Generative AI [Link]

- Developed a **43 Million Parameter** text-generative AI model using **Transformers** from scratch to generate fictional stories.
- Trained the model with over **42000 Tokens** of vocabulary. Optimized the model architecture and hyperparameters to achieve a **4.02%** testing loss in generating coherent and diverse storylines.

Multi-Agent Reinforcement Learning System [Link]

- Architected a complex Multi-Agent RL game environment using **DQN** and **A2C** algorithms to train competitive AI agents.
- Achieved **99%** target rate in under **2000** training episodes with improvements in **Policy Initialization** and **Reward Design** strategies.

SKILLS

Languages: Python, C, C++, Java, SQL.
Frameworks & Libraries: PyTorch, TensorFlow, Keras, JAX, NLTK, OpenCV, Scikit-Learn, Hadoop.
AI/ML & Gen AI : Deep Learning, Computer Vision, Natural Language Processing, Computer Science, Data Analysis, Large Language Models (LLMs), LangChain, Quantization, Fine-tuning, PEFT, Distributed Computing, Inference.
Other Skills: MLOps, MLFlow, CUDA, CI/CD, Git, Version Control, CLI, Cloud Computing, Containerization, Docker, Kubernetes, Deployment, Problem-solving.