**ANURAG KALAPALA**

anuragworkprof@gmail.com | +1 (978) 943-2495 | [Linkedin](https://www.linkedin.com/in/anurag-kalapala-a03a28355/) | Dallas, TX

**Professional Summary**

A hands-on Gen AI Engineer with over two years of software development experience, including a dedicated focus on building and deploying next-generation AI applications. I specialize in using Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), and agentic frameworks like LangChain. Proficient in Python and its core AI/ML libraries (Hugging Face, PyTorch), with proven experience developing REST APIs with FastAPI. I am passionate about contributing to real-world Gen AI solutions in a collaborative, fast-paced environment.

**Education**

**Master of Science in Computer Science** | University of Massachusetts Lowell | Jan 2023 - May 2024 *Relevant Coursework: Machine Learning, Deep Learning, Natural Language Processing, AI*

**Bachelor of Engineering in Electronics and Communication** | Chaitanya Bharathi Institute of Technology | June 2018 – May 2022

**Professional Experience**

**Gen AI Engineer** | HealthEdge | *July 2023 – May 2024*

* Assisted in building and deploying Gen AI applications using **Python** and **FastAPI**, containerized with **Docker** for scalable microservice delivery.
* Supported the development and optimization of **RAG pipelines**, leveraging a strong understanding of **LLMs, embeddings, and vector search** concepts to retrieve information from unstructured data.
* Contributed to **prompt engineering**, testing, and optimization for various LLM-based applications to enhance performance and reliability.
* Participated actively in **Agile development processes**, including code reviews and documentation, collaborating with senior engineers to deliver high-quality, production-ready code.
* Helped integrate Gen AI services from cloud platforms like **AWS** into the existing application stack.

**Graduate Research Assistant** | University of Massachusetts Lowell | *June 2023 - May 2024*

* Collaborated with senior data scientists to **fine-tune and evaluate LLMs** (BERT, RoBERTa) for domain-specific NLP tasks, improving model accuracy and efficiency.
* Preprocessed large volumes of unstructured text data for LLM consumption, laying the groundwork for advanced NLP applications.
* Wrote test scripts to validate model outputs and system performance against established benchmarks.

**Technical Projects**

**RAG-based Misinformation Detection System**

* Architected and built a complete **RAG pipeline** to verify information credibility in real-time.
* The system utilized **LLM embeddings** and **vector search** to retrieve relevant context from a knowledge base, which was then fed to a fine-tuned SBERT model for classification.
* Developed and deployed the application as a **REST API using FastAPI**, demonstrating strong hands-on API development skills.
* The entire project was containerized with **Docker** and deployed on AWS, showcasing familiarity with cloud environments and CI/CD pipelines.

**Visual Classification System with PyTorch**

* Developed an end-to-end deep learning application using **PyTorch** for image classification.
* Exposed the model's functionality through a REST API, further strengthening experience in API design and deployment.
* Gained experience with the full machine learning lifecycle, from data preprocessing to model evaluation and deployment.

**Technical Skills**

* **Gen AI & LLMs:** **LangChain**, **Hugging Face Transformers**, **RAG Pipelines**, **Prompt Engineering**, Vector Search (FAISS, Pinecone), LLM Fine-Tuning, OpenAI API, **AWS Bedrock**.
* **Programming & Frameworks:** **Python**, **PyTorch**, **FastAPI**, Flask, SQL.
* **ML Concepts:** Classification, Regression, Evaluation Metrics, Embeddings.
* **DevOps & Tools:** **Docker**, **Git**, CI/CD, Jupyter Notebooks, Postman.
* **Cloud Platforms:** AWS, GCP.