

Vishnu Koraganji

Email: vishnukoraganji369@gmail.com | Mobile: +1(717) 686-6762 | Portfolio: 07VK.github.io/Resume | LinkedIn: linkedin.com/in/vishnukoraganji

Professional Summary

Founding AI Engineer building HIPAA-compliant medical RAG platforms. MS in Computer Science from Penn State University with specialization in AI/ML, NLP, Gen AI and Agentic systems, with 1-2 years of project based solid foundations in machine learning and hands-on experience deploying end-to-end ML solutions using Python, PyTorch, and TensorFlow. Demonstrated ability to solve complex problems and optimize performance in building scalable ML systems.

Technical Skills

Languages: Python, Java, C, JavaScript, SQL

Machine Learning & Deep Learning: PyTorch, TensorFlow/Keras, Pandas, Scikit-learn, NLTK, MLFlow, DVC; Classification, Clustering, Regression, NLP, Computer Vision, CNN, RNN, YOLO

Gen AI Stack: Transformer models (BERT, GPT, Gemini, Llama), RAG Pipelines, PEFT (LoRA/QLoRA), LangChain, CrewAI, Hugging Face, FAISS, Vertex AI Vector Search, Cohere Rerank

MLOps & Cloud: Docker, MLFlow, GCP (Vertex AI, Document AI, Cloud Run), AWS, HIPAA Compliance

Databases & Tools: PostgreSQL, Firebase, Firestore, ChromaDB, Ollama, Ngrok, ArcGIS, CI/CD, FastAPI, Git

Experience

Founding AI Engineer | ClearChartAI, Remote

Aug 2025 - Present

- Architected **HIPAA-compliant** medical RAG platform on GCP using **Document AI + Gemini 2.5** hybrid approach for processing clinical patient records including lab reports, discharge summaries, and consultation notes.
- Built intelligent **document chunking** pipeline with **medical entity recognition**, dosage pattern extraction, and clinical abbreviation handling, optimizing retrieval for patient record queries.
- Configured full GCP backend infrastructure: **Vertex AI** Vector Search, **Document AI processors**, **Cloud Run** deployments, with proper dev/staging/prod environment separation and **BAA compliance**.
- Implemented hybrid retrieval using **FAISS + Vertex AI embeddings** with **Cohere Rerank 3**, achieving **0.85+** cosine similarity on medical queries with **< 3s response latency**.

Graduate Research Assistant | Penn State University, Middletown, PA

Aug 2023 - May 2025

- Conducted applied AI research in medical imaging; designed and implemented a novel **YOLO-SCSA attention module** pipeline for skin lesion detection, achieving **4% mAP improvement** over baseline.
- Benchmarked **LLM fine-tuning vs RAG** approaches using small language models (SLMs) for domain-specific question answering, documenting trade-offs in accuracy, latency, and computational cost.
- Mentored **10+** undergraduate ML/AI capstone projects; provided technical guidance on model architecture, hyperparameter tuning, and experiment tracking using MLFlow.

Cloud Architect Intern | AWS (Amazon Web Services), Vijayawada, IND

Oct 2021 - Dec 2021

- Provisioned and managed AWS services (EC2, S3, RDS, VPC) establishing robust cloud infrastructure with auto-scaling policies and load balancers for optimized resource distribution.
- Secured cloud environments through IAM access controls, encryption mechanisms, and CloudWatch monitoring.

Projects

AI Agentic Resume Tailor (CrewAI, FastAPI, React, LangChain, Gemini API, Scrapy)

Mar 2025 - May 2025

- Developed CrewAI multi-agent system to automate resume tailoring, boosting ATS scores by **15-20%**; engineered web scraping agent parsing **5+ job boards** with **95% accuracy**.
- Built document processing pipeline (PDF, DOCX, LaTeX) with **100%** format preservation, deployed via FastAPI.

RAG Based QA Bot (PyTorch, DistilBERT, FAISS, FastAPI, LangChain, Ollama)

Jan 2025 - Apr 2025

- Engineered RAG chatbot for Q&A on Amazon Reviews dataset (**213K products**) with FT-DistilBERT sentiment analysis achieving **89% accuracy** and FAISS + MMR retrieval with **95% off-topic rejection**.
- Achieved **0.85** cosine similarity on product queries; deployed with FastAPI delivering responses in **2-3 seconds**.

Skin Cancer Detection System (Python, OpenCV, TensorFlow, PyTorch, YOLOv8)

Aug 2024 - Dec 2024

- Developed YOLOv8-based detection system with novel **SCSA Attention Module**, achieving **4% mAP@50** improvement on HAM10000 dataset.
- Integrated preprocessing pipeline (artifact removal, CLAHE, color normalization) enhancing precision/recall by **+8%**.

Landslide Estimation & Analysis (Python, TensorFlow, HDF5, ArcGIS)

Jan 2023 - May 2023

- Modeled landslide susceptibility using U-Net with DEM, NDMI, slope, elevation, and rainfall data, achieving **98% accuracy** (+**3%** over baseline).
- Generated landslide susceptibility map for SR-530 with **93% accuracy** using novel estimation method.

Education

Penn State University

Aug 2023 – May 2025

Masters in Computer Science

Middletown, PA

VR Siddhartha Engineering College

Aug 2019 - May 2023

Bachelors in Computer Science Engineering

Vijayawada, India