Hemanth Irivichetty

AI & MLOps Engineer
ihemanth.2001@gmail.com | +91 8500363606 | LinkedIn | GitHub

Professional Summary

Product-focused AI Engineer with 2+ years of experience architecting high-performance LLM inference engines and distributed RAG pipelines. Expert in reducing production latency by 40% and inference costs by 60% through advanced quantization and memory optimization. Proven track record of migrating legacy NLP systems to Transformer-based architectures and deploying scalable, secure AI microservices on Kubernetes (EKS).

Technical Skills

GenAI & LLM Inference: vLLM, CTranslate2, PagedAttention, Dynamic Batching, Quantization (Int8/AWQ), LoRA/QLoRA Fine-tuning
RAG & Retrieval: Hybrid Search (Vector + Knowledge Graph), FAISS, Cross-Encoders (BGE/Cohere), RAGAS Evaluation
Models: LLaMA 3.1, Qwen 2.5, FLAN-T5, LSTM, RNN
MLOps & Cloud: Kubernetes (EKS), Docker, AWS (EC2, S3), CI/CD (GitHub Actions), Prometheus, Grafana
Backend Engineering: Python (AsyncIO), FastAPI, Celery, SQLAlchemy (Async), Hybrid Encryption, JWT/OAuth2
Vision & OCR: Tesseract OCR, Document Layout Analysis, Object Detection (YOLO)

Professional Experience

LLM & Vision Infrastructure Engineer

EonForge (Logos Technologies LLC) · Dubai, UAE / Remote · July 2025 – Present
• Designed Hybrid RAG system combining FAISS vector search with Knowledge Graph traversal
• Built end-to-end document processing pipeline using Tesseract OCR with custom layout analysis
• Architected orchestration layer for 'LumenCipher' Insurance CRM with JWT/OAuth2 security
• Developed intelligent agents for automated claims processing using SQLAlchemy (Async)

Member Technical Staff (NLP & AI)

Zoho Corporation · Chennai, TN · June 2023 – June 2025
• **4x Throughput:** Migrated to vLLM, increasing throughput from 20 to 80 tokens/sec
• **40% Latency Reduction:** Implemented Int8 Quantization using CTranslate2, reducing P99 from 5s to 3s
• **60% Cost Reduction:** Achieved through CPU offloading and optimized GEMM kernels
• Fine-tuned FLAN-T5 and LLaMA 3.1 using LoRA/QLoRA for specialized tasks
• Deployed on Kubernetes (EKS) with HPA based on GPU metrics and Prometheus/Grafana monitoring

Project Trainee (AI/ML)

Zoho Corporation · Chennai, TN · Oct 2022 – May 2023
• Led migration from RNN/LSTM to Transformer-based pipelines
• Built data migration pipelines using Zoho Catalyst
• Enforced code quality with Poetry and Pytest

Education

B.Tech in Computer Science & Engineering
Sri Venkateswara College of Engineering · CGPA: 7.72