ANISH VIJAYVERGIYA

■ +91 8005708487 | ■ anishvijayvergiya1010@gmail.com
■ LinkedIn: linkedin.com/in/anish-vijayvergiya-a599b4190 | ■ GitHub: github.com/Anishvjy

PROFESSIONAL SUMMARY

- AI Engineer with 4 years and 10 months· experience designing and deploying production-grade Generative AI, NLP, and ML solutions.
- Specialized in RAG pipelines, transformer fine-tuning, and scalable MLOps on AWS/GCP.
- Proven track record in delivering latency-optimized, high-accuracy AI models for enterprise use cases across customer service, document automation, and knowledge management.
- Built enterprise-grade RAG systems reducing query latency by up to 50%.
- Fine-tuned Whisper-large-v3 & transformer models achieving >95% transcription accuracy.
- Designed automated ML workflows cutting deployment cycles by 40%.
- Hands-on expertise in LLMs (GPT, LLaMA, Mistral), vector databases, and cloud-native deployments.

CORE SKILLS

- ML/AI Frameworks: TensorFlow, PyTorch, Hugging Face, BERT, GPT-3, LLaMA, Mistral, XGBoost, CNNs, LSTMs, Llama-3.2, Qwen-2.5, Phi-3, Mistral Large-2, QLoRA, Agentic RAG, GraphRAG, hybrid (BM25+dense)
- Data & Pipelines: ChromaDB, MongoDB, Airflow, Vector Search, OCR (Tesseract, OpenCV)
- Cloud & Deployment: AWS, GCP, Docker, Kubernetes, CI/CD · Serving: vLLM, TGI/Ollama · rerankers (BGE/Cohere)
- Programming: Python, FastAPI, Flask · Vector DB: pgvector, Milvus, Weaviate, Pinecone (HNSW/IVF-PQ)
- Other Tools: Git, REST APIs, Microservices, LangGraph, CrewAI, GGUF

CERTIFICATIONS

- Deep Learning Specialization – Coursera
- Machine Learning Specialization – Coursera

PROFESSIONAL EXPERIENCE

AI Engineer – **Globiva** (Bengaluru | Jan 2025 – Present)
- Designed and deployed a RAG pipeline (ChromaDB + LLaMA/Mistral) enabling 50% faster document query resolution.
- Fine-tuned Whisper-large-v3 for call-center transcription, boosting speech-to-text accuracy from 89% → 95%+.
- Built scalable MLOps pipelines with Docker, Kubernetes & FastAPI; implemented CI/CD workflows reducing deployment time by 40%.
- Automated HR data synchronization via Airflow DAGs with ZingHR APIs, processing 10k+ records daily.

Independent AI/ML Engineer – **Consultant** (Remote | Jun 2022 – Jan 2025)
- Delivered custom RAG-based knowledge assistants for SMEs, improving query accuracy by 35%.
- Developed OCR-based document digitization system using CNN + LSTM, reducing manual data entry time by 70%.
- Fine-tuned BERT and GPT-3 for document classification & summarization, achieving F1-scores > 0.92.
- Provided MLOps consulting on AWS/GCP deployments for 5+ clients across finance, HR, and healthcare sectors.

Machine Learning Engineer (NLP) – **Tata Consultancy Services** (India | Jan 2021 – May 2022)
- Developed sentiment analysis models using LSTMs, improving classification consistency by 15%.
- Built OCR pipeline for scanned document processing, increasing extraction accuracy from 78% → 91%.
- Implemented data preprocessing workflows reducing model training times by 30%.

EDUCATION

B.E. in Electronics & Computer Engineering – MBM Engineering College, Jan 2020