

ANSH MANGUKIYA

AI Engineer · Generative AI · LLM Systems · RAG · Multi-Agent Orchestration

Surat, Gujarat, India | +91 7069216178 | anshmangukiya.ai@gmail.com
linkedin.com/in/anshmangukiya | github.com/ANSH1370

SUMMARY

AI Engineer with 1+ year of hands-on experience designing and deploying production-grade Generative AI systems — RAG pipelines, fine-tuned LLMs, and multi-agent orchestration platforms. Skilled in vLLM-based multi-GPU inference optimization, LangGraph/LangChain agent design, and integrating .NET enterprise systems with modern Python AI stacks. Proven impact: **80%** reduction in customer onboarding time, **150+ tokens/sec** inference throughput, and **60–70%** drop in manual engineering effort across automation initiatives.

TECHNICAL SKILLS

AI / Machine Learning: Large Language Models (LLMs), Generative AI, Retrieval-Augmented Generation (RAG), Fine-Tuning, Model Distillation, RLHF, Prompt Engineering, Model Inference Optimization, Multi-Agent Systems, AI Agent Orchestration

LLM Frameworks & Tools: vLLM, LangChain, LangGraph, CrewAI, Hugging Face Transformers, PyTorch

Vector Databases & Retrieval: Pinecone, ChromaDB, FAISS, Embeddings, Semantic Search

Software Engineering: Python, FastAPI, .NET Core, C#, ASP.NET, REST API Design, Microservices Architecture, System Design

Cloud & Infrastructure: AWS (EC2, S3), Docker, Kubernetes, CI/CD Pipelines, MLOps, Multi-GPU Deployment & Orchestration

Data Engineering & Databases: PostgreSQL, MySQL, SQL Optimization, ETL, Data Pipelines, Data Warehousing, Power BI, Tableau

Specializations: Production LLM Deployment, Enterprise Software Integration, ERP/CRM Systems, Cross-Platform Development (.NET + Python)

PROFESSIONAL EXPERIENCE

AI Engineer — Commercient LLC (Hybrid)

January 2025 – Present

- ▶ Architected and deployed a no-code ingestion and chatbot deployment platform enabling enterprise customers to autonomously launch domain-specific GenAI assistants from unstructured data sources (PDFs, web pages, internal documents).
- ▶ Built a customer-facing LLM tuning platform enabling clients to fine-tune domain-specific models on proprietary documents using model distillation and reinforcement learning feedback loops.
- ▶ Integrated tuned models into production workflows, reducing manual ticket triage effort by **60%** and improving response accuracy and consistency.
- ▶ Engineered a custom multi-GPU LLM inference stack on vLLM with advanced batching, KV-cache optimization, and parameter tuning, achieving **150+ tokens/second** throughput under concurrent load across distributed GPU servers.
- ▶ Designed an intelligent SQL view generator using LLM pattern recognition to auto-template ERP-to-CRM data connectors, eliminating **80%** of manual integration coding during new-customer

onboarding.

- ▶ Built a conversational ERP-to-CRM sync agent using LangGraph, enabling non-technical users to configure complex integrations via natural language and reducing engineering support load by **70%**.
- ▶ Implemented RAG-based knowledge retrieval pipelines with vector embeddings and semantic search, powering context-aware responses across deployed chatbots.
- ▶ Bridged .NET enterprise applications (ASP.NET / C#) with Python AI services through FastAPI microservices and well-defined REST contracts, enabling seamless cross-platform data flow.
- ▶ Containerized and deployed AI services with Docker on multi-GPU infrastructure, applying CI/CD and MLOps practices for reliable, repeatable production rollouts.

EDUCATION

Bachelor of Engineering — Computer Science

2021 – 2025

Sarvajani College of Engineering & Technology, Surat, Gujarat

CGPA: 9.78 / 10

CERTIFICATIONS

- ▶ **Industrial Artificial Intelligence with Cloud Computing** — Microsoft [View Certificate ↗](#)
- ▶ **Introduction to NLP** — Infosys [View Certificate ↗](#)
- ▶ **Sequences, Time Series and Prediction** — Coursera (DeepLearning.AI) [View Certificate ↗](#)