Muhammad Murtaza

Senior Python AI Engineer | AI Agents & LLM Systems | RAG Pipelines

muhammadmurtaza6211@gmail.com | +92 317 0924890 | linkedin.com/in/muhammad-murtaza-144a3a223 | Kohat, Pakistan (UTC+5)

PROFESSIONAL SUMMARY

Senior Python AI Engineer with 4+ years of production experience designing and deploying AI agent systems, LLM-powered applications, and RAG pipelines that automate real business workflows at scale. Proven track record building multi-agent architectures with LangChain, LangGraph, and LlamaIndex; fine-tuning LLMs (Llama 3, GPT-4) with LoRA/QLoRA; and designing RESTful microservice APIs (FastAPI, Flask) serving 100K+ daily requests at sub-200ms latency. Delivered systems handling 50K+ monthly users and 2M+ document pipelines. Reduced inference latency by 25% and cloud costs by 30% through quantization and infrastructure optimization. Proficient with DeepEval and Ragas for LLM evaluation, Docker/Kubernetes for containerization, and AWS/Azure/GCP for cloud deployment. Strong communicator comfortable collaborating across engineering, product, and data science teams.

TECHNICAL SKILLS

Python & Frameworks	Python (5+ yrs), FastAPI, Flask, Django, Node.js, NestJS
AI Agents & Orchestration	LangChain, LangGraph, LangSmith, LlamaIndex, Multi-Agent Orchestration, Tool Integration, Stateful Memory, Human-in-the-Loop
LLMs & Generative AI	GPT-4/GPT-5, Claude (Anthropic), Llama 3, Gemini, Mistral, AWS Bedrock, Groq, Azure OpenAI
LLM Fine-Tuning	LoRA, QLoRA, Unsloth, 4-bit Quantization, Hugging Face Transformers, PEFT
RAG & Retrieval	RAG Pipeline Design, Semantic Search, Hybrid Search, Pinecone, FAISS, ChromaDB, Weaviate, Milvus
LLM Evaluation	DeepEval, Ragas, Prompt Engineering, Structured Output (JSON mode, Function Calling)
ML & Deep Learning	PyTorch, TensorFlow, scikit-learn, Hugging Face, Text Classification, NER, Summarization, Sentiment Analysis
Prototyping Tools	Streamlit, Gradio
Speech & Voice AI	Whisper ASR, ElevenLabs TTS, Deepgram, Voice Pipelines
Frontend	React, Next.js 15, TypeScript, TailwindCSS, Framer Motion, ReactFlow
Cloud & MLOps	AWS (SageMaker, Lambda, EC2, Bedrock), Azure (OpenAI, ML), GCP, Docker, Kubernetes, MLflow, CI/CD, Git
Databases & Infra	PostgreSQL, MongoDB, Redis, Elasticsearch, Neon, Convex, ImageKit
Security & APIs	REST APIs, Microservices, JWT Auth, HIPAA Compliance, OAuth, API Rate Limiting
Other Tools	Streamlit, Remotion, Inngest, Clerk, Twilio, Stripe, Git

PROFESSIONAL EXPERIENCE

Generative AI Engineer & Python Backend Developer *Jan 2023 – Aug 2025*

Vision Byte Technologies — Islamabad, Pakistan

- Designed and deployed production multi-agent AI systems using LangChain, LangGraph, LangSmith, and LlamaIndex — automating complex business workflows with parallel agent orchestration and tool integration, achieving 40% efficiency improvement.
- Built production RAG pipelines handling 2M+ documents at 92% retrieval accuracy using Pinecone, FAISS, ChromaDB, Weaviate, and Milvus with hybrid semantic search strategies.
- Fine-tuned LLMs (Llama 3, GPT-4, domain-specific models) using LoRA, QLoRA, and Hugging Face Transformers; achieved 35% performance improvement over baseline through custom training strategies and 4-bit quantization.
- Developed scalable FastAPI/Flask RESTful microservice APIs serving 100K+ daily requests at sub-200ms latency; designed secure endpoints with JWT auth and API rate limiting.
- Integrated and orchestrated cloud-hosted foundation models via AWS Bedrock, Azure OpenAI, Groq, and Anthropic Claude; reduced inference cost by 30% through model quantization and efficient batching.
- Built LLM evaluation pipelines using DeepEval and Ragas to monitor retrieval quality, answer faithfulness, and hallucination rates across deployed RAG systems.
- Designed and deployed end-to-end voice AI pipelines integrating Whisper ASR and TTS systems (ElevenLabs, Deepgram), enabling hands-free workflow automation and reducing manual tasks by 45%.
- Deployed AI workloads on AWS (SageMaker, Lambda, EC2) and Azure using Docker and Kubernetes — handled 10x traffic growth with 99.9% uptime and zero-downtime deployments.
- Established MLOps workflows with CI/CD pipelines, model versioning, and experiment tracking using MLflow, ensuring

reproducible and reliable model deployments.
- Built rapid prototypes using Streamlit and Gradio for internal demos, accelerating stakeholder feedback loops and product iteration cycles.
- Implemented HIPAA-compliant security best practices for sensitive healthcare data: encrypted data at rest and in transit, role-based access control, and audit logging.
- Built NLP pipelines for text classification, NER, and summarization processing millions of documents with 92% average accuracy across healthcare and enterprise use cases.
- Collaborated cross-functionally with data scientists, product managers, and frontend developers to deliver end-to-end AI products; explained technical concepts clearly to non-technical stakeholders.

ML & Deep Learning Engineer *Jan 2021 – Nov 2022*
Dot Coder — KPK, Pakistan
- Built and deployed machine learning and NLP systems in production using Python, TensorFlow, and PyTorch for text classification and sentiment analysis, achieving 89% accuracy on real-world datasets.
- Developed end-to-end ML pipelines covering data collection, preprocessing, model training, evaluation, and deployment with automated monitoring and alerting.
- Created scalable Flask REST APIs exposing ML models to frontend applications, handling reliable JSON-based communication across services.
- Built rapid prototypes with Streamlit dashboards to visualize model outputs and communicate results to stakeholders.
- Optimized model deployment using Docker containerization, ensuring consistent cross-platform performance and simplified CI/CD integration.
- Applied security best practices for model APIs including input validation, authentication, and protection against prompt injection patterns in NLP systems.

KEY PROJECTS

AI Travel Chatbot *LangGraph · FastAPI · Next.js · AWS Bedrock · Groq · Llama 3.1 70B*
- Built an AI-powered travel planning chatbot with 9+ specialized parallel agents orchestrated via LangGraph: flight resolution, hotel search, activity planning, budget analysis, visa/weather advisory, and itinerary generation.
- Implemented Human-in-the-Loop Gen-UI with interactive cards inside the chat interface — planning engine fires only after full context is captured via structured JSON outputs.
- Stack: Next.js, TailwindCSS 4, Framer Motion, Clerk, Convex (frontend) + FastAPI, LangGraph, Redis, Docker, PostgreSQL (backend) with streaming SSE responses.

Healthcare Conversational AI Platform *LLM Fine-Tuning · NER · FastAPI · RAG · DeepEval*
- Built end-to-end conversational AI using fine-tuned LLMs, NER, and intent classification for patient engagement — processing 50K+ monthly interactions with 4.6/5 satisfaction rating.
- Integrated RAG pipeline over healthcare knowledge base with 92% retrieval accuracy using Ragas evaluation; implemented HIPAA-compliant data handling with encryption and audit logs.

Voice Medical Assistant — MediSynapse *Whisper ASR · RAG · Vision · TTS · HIPAA*
- Developed HIPAA-compliant voice AI assistant combining RAG, computer vision, and speech recognition with custom TTS for hands-free medical workflow automation.
- Reduced manual clinical tasks by 45% through intelligent voice-driven interaction and document retrieval with real-time streaming responses.

AI Course & Video Generator *Next.js 15 · FastAPI · LangChain · Remotion · Deepgram · AWS*
- Engineered a SaaS platform that converts any topic into a complete video course with animated slides, TTS narration, and auto-synced captions — fully automated end-to-end.
- Implemented parallel processing with ThreadPoolExecutor for concurrent slide, audio, and caption generation; used Inngest for background jobs and Redis for rate limiting.
- Built Streamlit admin dashboard for internal monitoring of job queues, user activity, and model performance metrics.

AI Career Assistant Platform *LangChain · LlamaIndex · NestJS · Next.js · Gemini AI · Inngest*
- Built full-stack SaaS with AI career counselor, resume analyzer (PDF parsing + scoring), ReactFlow-based roadmap generator, and personalized cover letter generator powered by Gemini AI.
- Used LlamaIndex for document ingestion and retrieval; Inngest for background AI agent orchestration; structured JSON outputs for consistent LLM responses.

LLM Fine-Tuning Pipeline *Llama 3 · LoRA · QLoRA · Unsloth · 4-bit Quantization · Hugging Face*
- Built end-to-end fine-tuning pipeline for Llama 3 using LoRA/QLoRA adapters and 4-bit quantization via Unsloth,

enabling high-performance training on consumer hardware.
- Evaluated fine-tuned models using DeepEval and Ragas benchmarks, achieving 35% accuracy improvement over baseline with automated regression tracking.

EDUCATION

Bachelor of Science in Information Technology *Oct 2021 – Jun 2025*

Kohat University of Science & Technology — Kohat, KPK, Pakistan

Relevant Coursework: Machine Learning, Natural Language Processing, Deep Learning, Data Science, Healthcare Informatics