Ajaz Ahmad

AI Engineer | GenAI, ML, AWS
khanajaz8395@gmail.com | +353-0899655863
[LinkedIn] | [Portfolio]

Executive Summary

AI Engineer with 5+ years of experience in building **AI focused products** in fast-paced, innovation-driven environments. Proven track record of designing and deploying production-grade ML/GenAI systems at scale, driving automation and analytics in regulated domains such as digital pathology and manufacturing. Skilled in **LLM integration**, **Computer Vision Models**, and full ML lifecycle ownership—from ideation and prototyping to scalable cloud deployment. Adept at aligning technical strategies with product goals in modern, data-centric organizations.

Core Competencies

SaaS & AI Product Development

- Scalable AI systems for cloud-native, production environments
- E2E ML model lifecycle: exploration → validation → deployment
- Product-focused ML development in regulated environments

GenAI & LLM Engineering

- GPT-4, LLaMA 2/3, LangChain Agents, OpenAI Functions
- Retrieval-Augmented Generation (RAG), structured prompting, fallback strategies
- Knowledge-grounded chatbots and semantic search over enterprise data

ML & Data Engineering

- Classical ML + Deep Learning for CV & NLP
- Multimodal AI: Image, text, and metadata fusion
- Semantic search: FAISS, Qdrant, hybrid retrieval

Infrastructure & Deployment

- FastAPI, Docker, Kubernetes, AWS (Lambda, S3, EC2), Azure
- CI/CD pipelines, MLOps best practices, observability, model versioning
- Backend: Python, PySpark, PostgreSQL, Elasticsearch, SQL, Tensorflow, PyTorch

Professional Experience

AI Engineer

Deciphex, Dublin, Ireland | Nov 2019 – Present

Product- AI Development

- Built **enterprise-grade Clinical RAG Assistant** for 50k+ pathology slides using GPT-4 and LangChain; achieved 95%+ metadata validation accuracy extracted from images using OCR.
- **Fine-tuned open-source LLMs** (e.g., LLaMA 2) with techniques like LoRA/PEFT for enterprise RAG pipeline integration, achieving improved factual grounding and response reliability.
- Created **production-ready LLM agents** with prompt fallback strategies, reducing metadata QA effort by 75%.

GenAI & LLM Systems

- Architected **hybrid retrieval systems** combining FAISS/Qdrant with traditional filters for high-precision GenAI results.
- Enabled semantic querying of medical databases through LangChain Agents + OpenAI Functions, improving query resolution efficiency by 60%.

Engineering for Scale

- Deployed microservices handling millions of images using **Docker, Kubernetes, and AWS**, ensuring sub-second latency.
- Led adoption of CI/CD pipelines and version-controlled MLOps for robust model releases.
- Deployed scalable GenAI model, cutting latency by 30% and costs by 20% via optimized orchestration and lightweight APIs.
- Strategic migration of AI compute jobs from DataCrunch to Oracle Cloud Infrastructure (OCI). Mentored junior engineers; championed team-wide code quality and reproducibility.

Software Engineer – AI & Automation (Manufacturing SaaS)

Volkswagen IT Services, Pune, India | Feb 2017 – Aug 2018

- Built a conversational AI assistant using AWS Lex for real-time factory reporting, reduced downtime and query time significantly.
- Designed predictive maintenance models and integrated them into real-time Azure dashboards.
- Developed scalable ETL pipelines in Snowflake for backend APIs servicing 1M+ transactions daily. Used SQL for KPI analytics and historical equipment failure pattern analysis.

Highlighted SaaS AI Projects

- **Clinical RAG Assistant** – GPT-4, LangChain, FAISS
  Automated metadata QA for 50k+ clinical slides; deployed as a cloud-native AI feature
- **Semantic Query System** – Qdrant, Streamlit, OpenAI
  Enabled free-text medical querying with real-time GenAI backend
- **Multimodal QA Pipeline** – CV + LLM + FastAPI + Kubernetes
  Integrated image analysis and GenAI for clinical diagnostics
- **Manufacturing Intelligence Bot** – AWS Lex, Lambda, Data Bricks
  Delivered to 500+ users with 60% faster query resolution
