

Annas Mustafa

+49 157 5727 4762 | annasmustafa77@gmail.com | Berlin, Germany
linkedin.com/in/annas-mustafa | github.com/AnnasMustafaDev

SUMMARY

AI/ML Engineer with over 3 years of experience building intelligent, agentic systems and RAG pipelines using large language models. Currently pursuing a Master's in Artificial Intelligence, with a focus on developing reliable, scalable, and production-ready AI solutions that bridge research and real-world applications. Passionate about innovation, automation, and creating AI systems that think and act autonomously.

TECHNICAL SKILLS

LLM Ecosystems: OpenAI, Gemini, Claude, Azure AI, LangChain, LlamaIndex, LangGraph, CrewAI
Programming & Backend: Python, FastAPI (Sync/Async), SQL, C++, TypeScript, JavaScript, Next.js
RAG Engineering: Vector Stores (FAISS, Pinecone, Azure Search), Embeddings, Chunking, Hybrid Retrieval (BM25), Indexing, Re-ranking, Caching, Streaming
Agentic AI & Tool Use: Reflexive Reasoning, Function Calling, Workflow Orchestration, A2A, MCP
MLOps & DevOps: Docker, Git/GitHub Actions (CI/CD), Cloud (GCP, AWS, Azure)
Other Tools: n8n, MongoDB, PostgreSQL, Streamlit, LoRA, QLORA, PEFT, W&B, Arize Phoenix, MLflow, Custom Evaluation Pipelines, Model Versioning & Governance

PROFESSIONAL EXPERIENCE

AI Product Engineer

BIK GmbH

Jan 2026 – Sep 2026

Nuremberg, Germany

- Built per-client multi-tenancy on Azure for an AI knowledge platform, isolating 10+ data domains (documents, chat, recordings, LLM memory) into dedicated tenant databases
- Implemented envelope encryption at rest with key-rotation tooling and a phased production rollout
- Built Excel-to-Neo4j knowledge-graph import with entity resolution and fuzzy matching, enabling natural-language queries over tabular business data
- Benchmarked Docling vs. Azure Document Intelligence and Graphiti alternatives to drive build-vs-buy decisions
- Built per-tenant LLM cost tracking and billing with automated reconciliation against Azure usage
- Delivered GDPR PII inventory and cross-store deletion; fixed credential leaks and added data-isolation alerting
- Shipped 30+ merged PRs across a FastAPI, PostgreSQL/PGVector, Neo4j, Redis, and Docker stack

AI/ML Developer

Stixor

Feb 2025 – Sep 2025

Islamabad, Pakistan

- Led the design and integration of agentic AI architectures for contract automation and self-editing agents
- Built a RAG-based legal chatbot system with model versioning, tracing, and evaluation pipelines
- Contributed to 'Malakah', a legal AI startup that secured \$600K in seed funding, developing backend APIs via FastAPI with Dockerized deployments
- Applied MLOps practices (monitoring, cost optimization, model iteration) and optimized inference latency through hybrid retrieval and caching

AI Engineer

Niblon

Mar 2024 – Jan 2025

Remote

- Designed and deployed 10+ multi-agent and RAG systems with adaptive memory and reflexive reasoning
- Created React-based AI agents using LangGraph for contextual orchestration and feedback loops
- Built LLM-as-Judge pipelines and Human-in-the-Loop evaluators, improving accuracy by 25%
- Enhanced RAG retrieval precision with embedding quantization, adaptive chunking, and hybrid ranking, reducing response latency by 40%
- Researched Agent-to-Agent and MCP-based reflexive reasoning for scalable automation

Associate ML Developer

Developers Den LLC

Dec 2022 – Jan 2024

Remote

- Architected and deployed hybrid RAG systems with multi-step retrieval pipelines, combining dense embeddings (FAISS) and sparse retrieval (BM25) to achieve 35% improvement in answer relevance
- Integrated Zapier workflows with LLM-powered automation agents, enabling seamless connections between 50+ SaaS tools and reducing manual task processing time by 60%
- Built custom evaluation frameworks using RAGAS metrics to monitor and improve RAG performance

- Developed voice-enabled AI agents using Whisper (STT) and ElevenLabs (TTS), processing 10K+ voice interactions monthly with 92% transcription accuracy
- Implemented multi-step reasoning chains with LangChain and custom tool-calling mechanisms, handling complex queries requiring 3-5 sequential API calls
- Optimized embedding generation and caching strategies, reducing API costs by 45%

PROJECTS

Production-Scale Text-to-SQL Agent with Reflexive Reasoning | GitHub

- Engineered an advanced agentic pipeline to translate complex natural language into accurate SQL queries over a large (UN, SDG, World Bank) database (350+ columns, 7M+ rows)
- Achieved 90% accuracy by implementing reflexive loops, memory, and structured tool-calling to manage the precision/groundedness trade-off
- Tools: LangGraph, OpenAI Function Calling, SQL, FastAPI, LangChain, Arize Phoenix, GitHub Actions

Multi-Agent Research Assistant with LangGraph | GitHub

- Built an autonomous multi-agent system using LangChain and LangGraph that orchestrates collaborative agents—Supervisor, Researcher, Writer, and Critiquer—to automatically gather information, generate content, and provide critical feedback
- Implemented agent-to-agent communication patterns and workflow orchestration for complex research tasks
- Tools: LangGraph, LangChain, OpenAI, Python, Multi-Agent Systems

Real-Time Custom Object Detection & Tracking with YOLOv8 | GitHub

- Designed and deployed a production-grade object detection and tracking pipeline using YOLOv8, fine-tuned on a proprietary dataset with 12 custom classes
- Achieved 92.7% mAP@0.5 and sustained 18ms latency per frame, supporting real-time video analytics
- Integrated DeepSORT for multi-object tracking
- Tools: YOLOv8, DeepSORT, Ultralytics, Python, OpenCV, PyTorch

LLMOps RAG Evaluation & Tracing Framework | GitHub

- Developed an end-to-end framework for managing, evaluating, and tracing Retrieval-Augmented Generation (RAG) pipelines using modern LLMOps practices
- Integrated LangChain, Arize Phoenix for monitoring, and custom evaluation metrics to track RAG performance, latency, and accuracy across production deployments
- Implemented automated testing pipelines and model versioning for continuous improvement
- Tools: LangChain, Arize Phoenix, Python, RAG, MLOps, Model Evaluation

EDUCATION

BTU Cottbus-Senftenberg

Master of Science in Artificial Intelligence

Cottbus, Germany

Sep 2025 – Present

- Coursework: Neural Networks, Mathematical Data Science, Information Retrieval, Image Processing & Computer Vision, Logic in Databases, Data Mining

HITEC University

Bachelor of Science in Computer Science – CGPA: 3.18/4.00

Taxila, Pakistan

Sep 2020 – Jul 2024

- Coursework: Machine Learning, Comparison of Learning Algorithms, Deep Learning, Big Data, Data Structures & Algorithms

CERTIFICATIONS

Generative AI with LLMs

Mar 2024

- Gained practical knowledge of Generative AI and LLMs through hands-on AWS training
- Studied GenAI research and real-world applications guided by AWS AI experts

LANGUAGES

English: C1 – Professional Working Proficiency

German (Deutsch): A2 – Elementary (Actively Improving)

