Annas Mustafa

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

Summary

Conversational AI Engineer with experience in 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.

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, Chunking, 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, Evaluation Pipelines, Model Versioning & Governance

Education

BTU Cottbus-Senftenberg — Cottbus, Germany
Master of Science in Artificial Intelligence — *Sep 2025 – Present*
- Coursework: Neural Networks, Mathematical Data Science, Information Retrieval, Image Processing & Computer Vision, Logic in Databases, Data Mining

HITEC University — Taxila, Pakistan
Bachelor of Science in Computer Science – CGPA: 3.18/4.00 — *Sep 2020 – Jul 2024*
- Coursework: Machine Learning, Comparison of Learning Algorithms, Deep Learning, Big Data, Data Structures & Algorithms

Professional Experience

Conversational AI Engineer — Feb 2025 – Sep 2025
Stixor Technologies — *Islamabad, Pakistan*
- Designed and integrated the **agentic AI architectures** for contract automation and self-editing agents.
- Built a **Self reflection AI system** with model versioning, tracing, and evaluation pipelines.
- Contributed to **'Malakah'**, a legal AI tech 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 — Mar 2024 – Jan 2025
Niblon — *Remote*
- Designed and deployed 10+ multi-agent and RAG systems with adaptive memory and reflexive reasoning
- 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%
- Developed voice-enabled AI agents using Whisper (STT) and ElevenLabs (TTS), processing 10K+ voice interactions monthly with 92% transcription accuracy
- Researched Agent-to-Agent and MCP-based reflexive reasoning for scalable automation

Associate ML Developer — Dec 2022 – Jan 2024
Developers Den LLC — *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
- Built custom evaluation frameworks using RAGAS metrics to monitor and improve RAG performance
- 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 multi-agent system using LangChain and LangGraph that orchestrates agents—Supervisor, Researcher, Writer, and Critiquer—to gather information, generate content, and provide critical feedback
- Implemented agent-to-agent 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
- 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 RAG pipelines
- Integrated LangChain, Arize Phoenix for monitoring, and custom evaluation metrics to track RAG performance, latency, and accuracy across production deployments
- **Tools**: LangChain, Arize Phoenix, Python, RAG, MLOps, Model Evaluation

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)