

# RAFIQUE NAZIR

AI ENGINEER

✉ rafique.nazir@protonmail.com

in LinkedIn 🌐 rafiquenazir.de

🌟 Certifications & Badges

👤 Work Permit: Permanent Residence

📍 Location: Darmstadt, Hessen, Germany



## SUMMARY

- AI Engineer building production **Generative AI** and LLM systems in **Python** on **AWS**. Current work covers agent tool orchestration over the **Model Context Protocol (MCP)**, translating natural language into API calls, and LLM evaluation.
- More than ten years of backend experience across multiple roles, building scalable microservices for enterprise clients, mainly in **Python** (FastAPI, Rasa) and **Java**.
- **AWS Certified Solutions Architect** with hands-on experience in Bedrock, EKS, **Kubernetes**, **Helm**, **Terraform**, and GitLab CI/CD.
- Used to working independently. Currently the only engineer on an enterprise AI platform, responsible for architecture, implementation, quality, deployment, and scoping.

## PROFESSIONAL EXPERIENCE

### BITA GmbH

AI Engineer

Frankfurt, Germany

01.2026 – Present

- Only AI engineer on a project that turns a **fintech index-building platform** into a conversational system, where users create investment universes, define index methodologies, and run backtests in plain language through AI agents.
- Built a **Just-In-Time tool discovery** system on the **Model Context Protocol (MCP)**, so agents find and call the right platform API at runtime. Tool search runs on **ChromaDB** with Titan embeddings.
- Designed the layer that translates natural language into API calls on **AWS Bedrock**, using schema-driven parameter extraction, documented defaults, and live lookup of platform data catalogs instead of hardcoded identifiers.
- Built an **LLM-as-a-Judge** evaluation framework and a YAML story-based end-to-end suite that checks tool selection and payload correctness. A story only gates the build after its flake rate has been measured.
- Set up tracing and per-model cost monitoring with **Langfuse** and **OpenTelemetry**, and added **fallback chains and I/O guardrails** so the system stays predictable and returns useful error messages when calls fail.
- Own the delivery pipeline (**Helm** on **AWS EKS**, GitLab CI with security scans, OAuth2/OIDC with Cognito and Keycloak) and handle scoping, estimation, API documentation, and C-level demos.
- **Tech:** Python, FastAPI, MCP, AWS Bedrock, AWS EKS, Langfuse, OpenTelemetry, ChromaDB, Redis, Docker, Helm, GitLab CI/CD, Pydantic, DeepEval, Gradio, Keycloak

### Career Break (Professional Development)

Germany

09.2024 – 12.2025

- Took a planned break for full-time training. Passed the official German **B1 and B2 (CEFR)** exams, earned the **AWS Certified Solutions Architect (Associate)** certification, and completed LLM engineering coursework (RAG, agents, fine-tuning).

### axxessio GmbH

Software Engineer

Darmstadt, Germany

01.2020 – 08.2024

- Key contributor to a **voice AI assistant** for Deutsche Telekom using **Rasa** and **Python**, helping achieve a **60%** increase in bookings.
- Built microservices and REST APIs with FastAPI, and integrated ML modules and **spaCy**-based NLU into production Rasa pipelines.
- Improved performance with **Redis caching** and PostgreSQL indexing, cutting API response times by **40%**; implemented CI/CD and infrastructure with **Kubernetes**, **Terraform**, and **Helm**.
- Mentored junior engineers (documentation, pair programming) and collaborated with ML and product teams in a cross-functional setup.
- **Tech:** Python, FastAPI, Rasa, spaCy, Redis, Kubernetes, Terraform, Helm, GitLab CI/CD, Azure, Docker, Pytest

## Working Student Roles

Software / Web Developer

Germany

08.2016 – 11.2019

- At **IDAT GmbH**, built a data-migration tool for SQL Server and automated AutoCAD workflows (C#, MS SQL Server, SSIS, tSQLt).
- At **arago GmbH**, wrote test automation for the **HIRO AI middleware** (Java, JUnit, Oracle).
- At **asuro GmbH**, developed features for a hybrid mobile insurance app (Ionic, Angular, TypeScript).

## Previous Industry Roles

Software Engineer

Karachi, Pakistan

08.2011 – 09.2015

- At **Connectcloud**, as Senior Software Engineer, worked on backend development for **UCloud**, a software-defined cloud platform, and migrated a legacy desktop application to a modern web application (Java Spring, REST, AngularJS, MySQL).
- At **SofDigital Systems**, improved ERP backend logic and financial reporting in the SmartERP and SmartHCM products (ASP.NET, C#, NHibernate, Oracle PL/SQL, Crystal Reports).
- At **Game Ventures**, built backend services for Howzat, a real-time multiplayer cricket game (PHP, PostgreSQL, TDD).

## TECHNICAL SKILLS

|                                       |                                                                                                                                                                                                                                                                                                            |
|---------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Languages &amp; Frameworks</b>     | Python (FastAPI, Rasa, Gradio), Java (Spring Boot)                                                                                                                                                                                                                                                         |
| <b>Generative AI &amp; LLM Agents</b> | Model Context Protocol (MCP, FastMCP), AWS Bedrock, Anthropic Claude, OpenAI API, LangChain, Ollama, agent orchestration, function calling, prompt engineering & management (Langfuse), RAG, vector databases (ChromaDB, pgvector), embeddings, hybrid search & reranking, fine-tuning (QLoRA), guardrails |
| <b>LLM Evaluation &amp; LLMOps</b>    | LLM-as-a-Judge, DeepEval, Ragas, retrieval metrics (MRR, nDCG, Recall@k), hallucination checks, e2e test suites                                                                                                                                                                                            |
| <b>Async &amp; Testing</b>            | asyncio, httpx; Pytest, TDD, pytest-mock, Ruff, MyPy, SonarQube                                                                                                                                                                                                                                            |
| <b>DevOps &amp; Cloud</b>             | AWS (EKS, Bedrock, EC2, S3, Lambda, RDS, SQS, Cognito), Kubernetes, Docker, Terraform, Helm, GitLab CI/CD, GitHub Actions, Prometheus, Grafana, OpenTelemetry                                                                                                                                              |
| <b>Databases &amp; ORM</b>            | PostgreSQL, MongoDB, Redis; SQLAlchemy, Alembic, Hibernate/JPA                                                                                                                                                                                                                                             |
| <b>Data &amp; Messaging</b>           | Apache Kafka, RabbitMQ, Amazon SQS; Pandas, Polars, NumPy                                                                                                                                                                                                                                                  |
| <b>Architecture</b>                   | AI Solution Architecture, Agentic AI, Microservices, Event-driven, RESTful API Design, Cloud-native, Infrastructure as Code (IaC)                                                                                                                                                                          |
| <b>Security</b>                       | OAuth2, JWT, OpenID Connect, RBAC, Keycloak, AWS Cognito                                                                                                                                                                                                                                                   |
| <b>AI-Assisted Development</b>        | Daily work with Claude Code and GitHub Copilot                                                                                                                                                                                                                                                             |

## CERTIFICATIONS

- AWS Certified Solutions Architect, Associate (08.2025) [View Certificate](#)
- AI Engineer Core Track: LLM Engineering, RAG, QLoRA, Agents, Udemy (2025)
- Professional Scrum Developer I, Scrum.org (03.2023) [View Certificate](#)
- ForgeRock Certified Access Management Specialist (07.2020)

## EDUCATION

### M.Sc. Distributed Software Systems

Technische Universität Darmstadt

10.2015 – 07.2019

Darmstadt, Germany

### B.Sc. Computer Science

National University of Computer and Emerging Sciences

08.2007 – 06.2011

Karachi, Pakistan

**Award:** NICT Scholarship Program (2007), a four-year, fully funded national merit scholarship from the Ministry of IT, Pakistan.

## LANGUAGES

**English:** Full professional proficiency • **German:** B2 (Deutsch-Test für den Beruf, 08.2025)