Maher Athar Ilyas

Germany | linkedin.com/in/athar-ilyas | github.com/atherrrrr

Experience

Applied AI Engineer Jul 2023 – Present

TestFabrik AG (webMate automated QA platform) *Saarbrücken, Germany*

- Architected webMate's central LLM inference layer, now the single path for every AI feature: tool calling, persisted conversations, tenant rate limiting, and EU-hosted Azure OpenAI with per-customer provider selection.
- Achieved the **highest automated coverage** of any tool in the W3C ACT implementation report (**53 of 87** WCAG 2.2 rules) with an LLM- and ML-based accessibility analysis service built from scratch.
- Sped up the self-hosted LLM **1.64x** (single-stream) and raised KV-cache capacity 1.7x on vLLM via quantization, a prefix-cache fix, and async scheduling, with 0 of 1,173 high-confidence predictions changed.
- Built a benchmarking harness scoring candidate LLMs on accuracy, latency, and cost against ground-truth accessibility cases on **100+ websites**; used it to evaluate the OpenAI to self-hosted migration.
- Built a chatbot testing framework end to end in Scala, modeled on DeepEval and promptfoo concepts, with LLM-as-a-judge scoring on custom evaluation rubrics, red teaming, test datasets, and regression testing.
- Cut model cost roughly **3x** across a model migration through input preprocessing, request chunking, and inference tuning, re-validated against ACT test cases.
- Led AI feature development through planning, task delegation, and code review, owning technical decisions on model selection, inference providers, and GPU infrastructure.
- Led TestFabrik's side of the **LowBarrier** research consortium with DFKI, Fraunhofer IESE, and WSC, directing roadmap, requirements, and architecture for WCAG 3.0 conformance.
- Ran a client PoC for webMate, from requirements meetings to a live demo on the client's own product.

Full-Stack Developer Jun 2022 – Jul 2023

LAYERMARK *Remote / Washington, D.C.*

- Rebuilt backend services in Java Spring with cache management, reducing response times by 60%, and introduced a Drools-based rules engine enabling configuration-driven product offerings without code changes.
- Cut front-end development and maintenance time by 70% by rebuilding the UI as reusable React.js components on Ant Design and DevExtreme, backed by REST APIs and PostgreSQL.

Automated Trading Systems Developer Sep 2021 – Apr 2024

Freelance (Fiverr & Upwork) *Remote*

- Designed and deployed 75+ crypto trading bots at a 5/5 client rating, integrating exchange APIs (Binance, KuCoin, Telegram) with web scraping for portfolio management, real-time technical analysis, and back-tested strategy reports.

Projects

Life OS: Personal Analytics MCP Server & Reporting Agent | *MCP, Cloudflare Workers* 2025 – Present

- Built a production MCP server exposing calendar, habit, finance, and Garmin health data as typed tools, backed by a serverless ETL layer normalizing four sources into scored daily records.
- Automated a weekly and monthly review pipeline in which agent skills define scoring formulas and templates, then write structured reviews straight into Notion with zero manual data entry.

SkillSprint | *AWS Lambda, DynamoDB, SageMaker, LLMs, Next.js* | **Best Project Runner-up** 2024

- Serverless micro-learning platform generating personalized skill roadmaps with LLMs and web-scraping curated sources into daily quizzes and flashcard infobits; built through the **AWS University Engagement** program.

WePrep | *LLMs, CNNs, computer vision, speech synthesis, FastAPI* | **Best Data Science Project Award** 2024

- Built an end-to-end AI interview simulator: LLM-generated questions rendered as lip-synced synthetic interviewer video, CNN-based emotion and body-language scoring of the candidate's recorded answers, LLM evaluation of answer content, and automated scraping of study resources targeting their weakest areas.

Education

Bilkent University Jun 2024

B.Sc. Computer Engineering, CGPA 3.69, Magna Cum Laude, Full Merit Scholarship *Ankara, Turkey*

Saarland University Apr 2023 – Sep 2023

Exchange Semester: Machine Learning, High-Level Computer Vision, Web Security *Saarbrücken, Germany*

Technical Skills

AI & ML: LLMs, vLLM, llama.cpp, quantization, LLM evaluation, fine-tuning (LoRA), harness engineering, red teaming, function calling, agentic workflows, MCP, RAG, embeddings & vector search, Hugging Face Transformers, PyTorch
Backend & Data: Python, FastAPI, Java, Spring Boot, Scala, PostgreSQL, Elasticsearch, RabbitMQ, Celery
Cloud & DevOps: AWS, Azure (AI Foundry), Docker, Kubernetes, CI/CD, Grafana, Cloudflare Workers
Frontend & Testing: React.js, Next.js, Angular, TypeScript, Playwright, Selenium, Appium, OpenSTF
Languages: English (C2), Urdu (C2), Turkish (A2), German (A2)