EMMANUEL ONWUBUYA

AI Engineer

Hannover, Germany · emyraeleson@gmail.com · +4915739814757 · LinkedIn · Portfolio

PROFILE

AI/LLM Engineer with 4+ years building production-grade data and AI systems. Specialises in agentic AI architecture, designing multi-agent workflows, production RAG systems, and LLM pipelines that ship to real users. Experienced in LLMOps across Azure, AWS, and GCP, with a track record of reducing cost, latency, and failure rates in live inference systems. Communicates system design decisions clearly to both engineering and executive stakeholders.

LLM Frameworks: LangChain, LangGraph agentic workflows, multi-agent systems, RAG pipelines

AI Systems: Agentic RAG, GraphRAG, tool-calling pipelines, structured output generation, LLM evaluation, MCP, OpenAI APIs, Azure AI Foundry, HuggingFace Transformers

Data & Cloud: Databricks, Snowflake, PySpark, DBT, Airflow, Azure, AWS, GCP

Engineering: Python, FastAPI, Docker, Terraform, SQL, CI/CD, MLflow

WORK EXPERIENCE

Data & AI Engineer — Wefra Life | **2024 – Present**

Frankfurt, Germany

- Architected a multi-agent LLM system using LangGraph with each agent handling a distinct task (anomaly detection, metric validation, insight generation) with shared state and fallback routing to prevent silent failures.
- Built production RAG pipelines with hybrid retrieval (dense + sparse search), chunking strategies, and cross-encoder reranking cutting manual analysis time by 40% while maintaining retrieval precision.
- Implemented structured output generation (JSON schemas via OpenAI function calling) and multi-intent classification to handle ambiguous user queries reliably across production workloads.
- Designed LLMOps observability layer: latency tracking, token cost monitoring, and automated regression tests to catch prompt drift and model degradation post-deployment.
- Migrated 200+ ETL models from dbt/Snowflake to Databricks using incremental loads and CDC cutting pipeline runtime by 50% and cloud costs by 25%.
- Designed Medallion (Bronze–Silver–Gold) architecture with Data Vault modeling, providing a reliable, lineage-tracked data foundation for downstream LLM workflows.

Data Engineer — Accenture | **2022 – 2024**

Hamburg, Germany

- Designed and built scalable data warehouse solutions using PySpark and SQL transforming heterogeneous source data into structured, schema-enforced datasets ready for downstream ML feature pipelines.
- Automated ETL orchestration using Skywise and Palantir, applying idempotent pipeline design and data quality checks reducing manual processing by 30% and improving data freshness SLAs.
- Optimised critical query paths via partitioning, predicate pushdown, and caching strategies reducing load times by 40% and improving system responsiveness under peak load.
- Reduced cloud infrastructure costs by 20% through right-sizing compute, implementing dynamic scaling policies, and consolidating redundant pipelines across hybrid Azure/GCP environments.

Junior Analytics Engineer — Domicil Real Estate Group | **2021 – 2022**

Munich, Germany

- Automated recurring ETL workflows using Python and shell scripting, reducing manual intervention and accelerating business insight delivery.
- Developed API-integrated data pipelines feeding structured data into cloud warehouses, supporting acquisition modelling for the finance team.
- Managed cloud VM provisioning and environment consistency via Linux and Bash scripting.

SKILLS

LLM & AI Engineering: LangChain, LangGraph, multi-agent architectures, agentic RAG, GraphRAG, tool-calling, structured output generation, prompt engineering, LLM evaluation (automated + human-in-the-loop), hallucination mitigation, OpenAI APIs, Azure AI Foundry, Amazon Bedrock, HuggingFace Transformers, fine-tuning, vector databases, hybrid search, reranking, Text-to-SQL, inference optimisation, MCP

MLOps / LLMOps: Model deployment, CI/CD, Docker, Terraform, Airflow, Dagster, MLflow, Grafana, Streamlit, FastAPI, latency/cost monitoring

Data Platforms & Cloud: Databricks, Snowflake, DBT, PySpark, Neo4j, Supabase, Postgres, Data Vault, Medallion Architecture, AWS, Azure, GCP, Tableau, Power BI

Programming: Python, SQL, JavaScript, React.js, Node.js, Bash, Cypher, REST APIs, HTML/CSS

Strengths: System design, stakeholder communication, trade-off analysis, agile delivery, data architecture, pipeline optimisation

PROJECTS

Ginja — Production AI Productivity System ginja.io | iOS & Android | React Native, FastAPI, LangGraph, LLM APIs, Supabase

End-to-end AI system converting unstructured user input into structured, actionable to-dos plans. Core pipeline: multi-intent classifier identifies input type (task capture, planning, reflection), routes to specialised LangGraph agents, and generates structured JSON outputs via OpenAI function calling. Designed for real-time inference, optimised prompt chains for sub-2s response times. Backend on FastAPI/Vercel with Supabase (pgvector) for user context and semantic memory. Implemented LLM evaluation loops and fallback logic for low-confidence outputs.

Ufindar — Semantic University Search Engine | LangChain, OpenAI, FastAPI, Streamlit, Supabase (pgvector), Docker, GCP

AI search engine enabling natural-language university queries via semantic retrieval over embedded institutional data. Deployed as a full-stack application on GCP with containerised FastAPI backend and Streamlit frontend.

EDUCATION

M.Sc. Data Analytics & Machine Learning — Universität Hildesheim | **Germany**

B.Eng. Information Systems & Technologies — Voronezh State University | **Russia**

LANGUAGES

English (Native) · German (Conversational) · Russian (Conversational)