

EMMANUEL ONWUBUYA

Forward Deployed AI Engineer

Germany · emyraeleson@gmail.com · +4915739814757 · [LinkedIn](#) · [Portfolio](#) · No Visa Sponsorship Required

PROFILE

Forward Deployed AI Engineer with 6+ years of experience across software, data, AI, and customer-facing solution delivery, owning problems from discovery and scoping through architecture, deployment, iteration, and production support. Delivered client-facing AI and analytics systems in regulated pharmaceutical environments, built and refined solutions for consulting clients, and independently built and operate Ginja, a production AI product with 3,000+ users. Strong in Python, LangGraph, Databricks, Palantir Foundry, and FastAPI.

WORK EXPERIENCE

Senior Data & AI Engineer — Wefra Life 2024 – Present

Frankfurt, Germany

- Led discovery with external clients, analysts, and internal teams to uncover reporting needs, access constraints, and workflow pain points, then translated them into deployed AI and analytics solutions.
- Built client-facing RAG and natural-language analytics systems over pharmaceutical and HCP data, combining role-aware access control, hybrid/vector retrieval, semantic business definitions, and tool calling for governed self-service insights.
- Diagnosed uncontrolled token costs in the initial RAG prototype, which called a single hosted model for every query regardless of complexity; designed and implemented a policy-based LLM gateway with smart routing across hosted and self-hosted models based on task complexity, cost, latency, and data sensitivity, cutting token cost by 12%.
- Owned the delivery cycle from requirements workshops and architecture through demos, user validation, rollout, and production iteration, refining workflows based on client feedback and observed failures.
- Implemented LangSmith observability across retrieval, model, and tool workflows, using traces to optimize caching, retrieval strategy, and model selection, reducing end-to-end AI response latency by 15%.
- Re-engineered 200+ Snowflake/dbt models on Azure Databricks and Delta Live Tables with CDC and incremental processing, reducing runtime by 50% and platform costs by 5–10% while powering downstream AI workloads.

Founder & Principal AI Engineer — AIXELAR Labs 2025 – Present · aixelar.io

Consulting Practice

- Lead client discovery and technical scoping, turning ambiguous business problems into proof-of-concepts, architecture decisions, and deployable AI, data, and software solutions.
- Build LLM-powered applications with tool calling and smart routing through an LLM gateway, optimizing model selection, caching, and execution paths to reduce token cost and latency by 10%.
- Own engagements through deployment and post-launch iteration, validating solutions with users, resolving production issues, and adapting workflows as client data and operational needs evolve.

Data Engineer — Accenture | Client: Airbus 2022 – 2024

Hamburg, Germany

- Worked directly with Airbus engineering and business teams to discover data needs, refine requirements, and iteratively deliver data products using PySpark, SQL, Palantir Foundry, and Skywise.
- Built and evolved enterprise-scale pipelines and transformation workflows across operational and engineering datasets, adapting implementations as client requirements and downstream use cases changed.
- Reduced distributed processing time by 40% through partitioning, predicate pushdown, caching, and query optimization.
- Reduced cloud infrastructure costs by 20% through workload optimization and more efficient resource utilization across Azure and GCP.

Junior Analytics Engineer — Domicil Real Estate Group 2020 – 2022

Munich, Germany

- Worked with real-estate acquisition and finance teams to understand reporting needs and translate fragmented property, acquisition, and financial data into usable analytics workflows.
- Built automated Python and SQL pipelines and API integrations that improved data availability and reduced manual preparation across real-estate reporting processes.
- Integrated data from multiple business systems into centralized reporting workflows, giving teams a more consistent view of portfolio and acquisition performance.

SKILLS

AI & Agent Systems: Python, RAG, LangGraph, LangChain, Tool Calling, LLM Gateways, Structured Outputs, MCP, LangSmith

Data Platforms: Palantir Foundry, Azure Databricks, Snowflake, dbt, PySpark, SQL, CDC, Delta Live Tables

Application & Deployment: FastAPI, TypeScript, React Native, Expo, Next.js, Vercel AI SDK, REST APIs, Docker, CI/CD

Cloud & Retrieval: AWS, Azure, GCP, PostgreSQL/pgvector, Vector Search, Hybrid Search

PROJECTS

Ginja — AI-Powered Personal Operating System | ginja.io | iOS & Android | 3,000+ users

- Independently built, shipped, and operate a production AI application serving 3,000+ users, owning architecture, backend systems, deployment, observability, and ongoing production support.
- Architected Brain Dump to convert unstructured text and voice into structured to-dos and plans using semantic retrieval, validated outputs, and deterministic confirmation before data is committed.

Number-it — AI-Powered SME Financial Platform | In Development

- Building an AI-enabled financial application that converts natural-language inputs into structured sales and expense records with explicit user confirmation before changes are committed.
- Implementing contextual conversations and tool-enabled workflows for transaction capture, classification, reporting, invoicing, and reminders.

EDUCATION

M.Sc. Data Analytics & Machine Learning — Universität Hildesheim, Germany | 2020 – 2023

B.Eng. Information Systems & Technologies — Voronezh State University, Russia | 2016 – 2020

LANGUAGES

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