

ÖMER FARUK KOÇ

MLOps & AI Platform Engineer

Istanbul, Türkiye • +90 507 219 2310 • omerfkoc98@gmail.com
linkedin.com/in/omerfkoc • github.com/negativexq • omerfkoc.dev

SUMMARY

MLOps and AI Platform Engineer with 3+ years of experience building and operating production ML, data, and Generative AI systems in financial services. Owned the lifecycle of **10+ production ML models**, engineered pipelines processing approximately **9M records daily**, and reduced a core feature engineering workflow from **120 to 30 minutes**. Recent engineering work focuses on reliable agentic systems, evaluation, observability, and failure recovery alongside ML platform engineering and on-premises LLM infrastructure.

EXPERIENCE

MLOps & Analytics Engineer

Fibabanka — Analytics Center of Excellence

Mar 2023 – Mar 2026

Istanbul, Türkiye

- Re-architected sequential Oracle feature-engineering workflows into parallel, modular **dbt + Airflow** pipelines, reducing runtime by **75%** from 120 to 30 minutes while improving lineage, dependency visibility, and maintainability.
- Engineered batch and near-real-time pipelines processing approximately **9M records daily** for feature generation, model training, batch inference, and downstream analytics.
- Owned deployment and lifecycle management for **10+ production ML models** on **MLRun (Iguazio)**, covering versioning, validation, production promotion, retraining, monitoring, and containerized FastAPI serving across Docker and Kubernetes environments.
- Built a **centralized feature platform** and dimensional data models shared by 10+ production ML models, replacing duplicated project-specific feature pipelines with reusable production-ready features.
- Prototyped ML and LLM workloads in Azure and migrated validated workloads to **on-premises GPU infrastructure**, serving quantized open-source models with llama.cpp and Ollama while preserving internal data residency.
- Delivered **RAG, document-processing, and structured information extraction** workflows and built a call-center intelligence pipeline processing approximately **9,000 recordings daily** with WhisperX transcription, timestamp alignment, speaker diarization, LLM summarization, and KPI extraction.
- Implemented data-quality, drift, and inference monitoring with **Great Expectations and Deepchecks**; automated model/data workflows with Airflow and CI/CD with GitHub Actions.

TECHNICAL SKILLS

Programming & Backend: Python, SQL, FastAPI, REST APIs, Pydantic, SQLAlchemy

MLOps & AI Platform: MLRun, Docker, Kubernetes, GitHub Actions, Model Serving, Feature Stores, Canary Deployment, Model Monitoring, Drift Detection

Generative AI & Agentic Systems: LangGraph, LangChain, RAG, Hybrid Retrieval, Qdrant, Deterministic Guardrails, Human-in-the-Loop, Evaluation, MCP, llama.cpp, Hugging Face Transformers, WhisperX

Data & Streaming: Apache Airflow, dbt, Apache Kafka, Redis, PostgreSQL, Oracle, Trino/Starburst, ETL/ELT

Observability & Infrastructure: OpenTelemetry, Jaeger, Prometheus, Grafana, Great Expectations, Deepchecks, Azure, On-Premises GPU, Linux, Git

SELECTED PROJECTS

Agentic Customer Service Platform [GitHub]

2026

Built a production-oriented agent platform where **LLM outputs are treated as untrusted proposals rather than execution authority**. Introduced deterministic grounding, target validation, policy enforcement, persisted confirmation, idempotency, and audit controls before business mutations. Designed a bilingual model evaluation framework across **540 executions** and an operational release gate validating concurrency, recovery, and observability; achieved **0 unsafe executions, 0 confirmation bypasses, and 0 unauthorized or duplicate mutations**, while driving unsafe executable candidates through a deterministic hardening path from **15 → 3 → 0 → 0 → 0**.

Knowledge Base RAG [GitHub]

2026

Engineered a multi-source RAG platform with incremental synchronization, **dense+sparse retrieval, RRF fusion, cross-encoder reranking, citation integrity, Qdrant reconciliation, OpenTelemetry, and DeepEval**. Hardened versioned re-indexing against partial writes, cancellation, schema migration, and registry/index drift with a **448-test suite**; controlled bilingual experiments isolated cross-lingual mismatch as a reranking weakness.

ModelOps Control Plane [GitHub]

2026

Built an ML release control plane with **weighted canary routing, policy-driven evaluation, automated promotion/rollback, fault injection, and auditable deployment timelines**. Added optimistic concurrency controls and end-to-end validation for worker-driven rollback and **10% → 25% → 50% → 100%** progressive promotion across a multi-service deployment stack.

Real-Time Commerce Platform [GitHub]

2026

Designed a Kafka platform with **at-least-once processing, idempotent consumers, transactional persistence, Redis coordination, bounded retries/DLQ, and a transactional outbox**. Profiled and optimized Kafka/PostgreSQL hot paths: bounded offset-commit batching and query-plan-driven indexing reduced a critical lookup from approximately **10.9 ms to 0.25 ms** and increased sustained throughput in controlled local benchmarks from approximately **750 to 1,050 events/s (+40%)** with three workers and three partitions.

EDUCATION

B.Sc. in Electronics Engineering, Gebze Technical University

2026