

Ömer Faruk Ballı

Senior Data Engineer & AI/ML Engineer | GenAI & LLM Specialist

omerfarukballi@gmail.com • +90 510 223 3977 • linkedin.com/in/omerfarukballis • github.com/omerfarukballi

PROFILE

Senior Data Engineer & AI/ML Engineer with 7+ years shipping production GenAI systems and enterprise data platforms — largely through **fixed-scope contract engagements: build, deliver, hand over**. Specialized in LLM application development (RAG/GraphRAG, multi-model orchestration, LLM security), production ML (churn, recommendation, pricing), and end-to-end data infrastructure (Snowflake, Airflow, ETL/ELT). Every engagement closed with a working system: two zero-downtime migrations, a live enterprise LLM chatbot, a published PyPI SDK, and two consumer apps shipped solo. **Now moving to a long-term in-house role to grow what I build.**

EXPERIENCE

Data Excellence Lead (Contract), Air Liquide

Apr – Sep 2026 · *contract completed*

- Led data strategy across Turkey, Bulgaria, and Romania for the Align program (global SAP S/4HANA implementation), translating business requirements into master data specifications across manufacturing and retail operations
- Architected master data preparation across ERP and satellite systems (Wise, Concur, Workday, Salesforce); engineered ETL workflows and SQL extraction pipelines with data-quality governance through SIT, UAT, and cutover phases
- **Completed fixed-scope engagement with structured handover** of governance framework, cutover documentation, and executive analytics to central teams

Senior Data Scientist (Contract), BGTS

Dec 2025 – Apr 2026 · *delivered & handed over*

- Built and operated AutoData, a full-stack ML data-preparation platform (Next.js + Flask + PostgreSQL/Redis on Azure, H100 GPU workers): 6-stage pipeline, 19+ data connectors (Snowflake, BigQuery, Databricks, S3/GCS/Blob, MongoDB, Kafka), published PyPI SDK with REST API · datatoolpack.com
- Led zero-downtime production migration via live public-IP cutover (31 GB → 48-core/992 GB VM), eliminating OOM failures; hardened security with LLM prompt-injection guards, CSRF, Fernet credential encryption, and read-only SQL validation
- Delivered 4Global POI data-cleansing platform (Streamlit) for Middle East Google Maps datasets: rule-based deduplication (lat/lon/name entity resolution), automated re-categorization, and QA validation for the client's planning tool

Senior Data Engineer (Contract via Adesso), Philip Morris International

Jan – Nov 2025 · *delivered & handed over*

- Led data-platform migration from legacy AWS (MS-SQL, SSIS) to Snowflake/Matillion architecture with zero downtime, ensuring uninterrupted business continuity throughout the transition
- Architected the full data warehouse in Snowflake — dimensional models (SCD Type 2), stored procedures, and views across stage, integration, and presentation layers on medallion architecture
- Designed multi-tenant Matillion ETL environments with custom PySpark components; deployed Power BI dashboards on Snowflake and handed the platform over to the internal team at contract end

AI/ML Engineer (Contract), Multi

Aug 2024 – Jan 2025 · *chatbot live at handover*

- Architected a production enterprise LLM assistant with multi-model orchestration (LLaMA, Gemma, Mistral) on open models for data privacy — WhatsApp integration, REST APIs, multi-format document processing (PDF, DOCX, images)
- Implemented advanced RAG pipeline with ChromaDB vector store, semantic search, and context-aware prompt templates, achieving high accuracy across diverse enterprise query types
- Designed secure multi-tenant architecture — session management, per-tenant document repositories, prompt versioning — delivered live to production and documented for the client team

Data & AI Engineer (Contract), Warner Music Group

Jul 2023 – Aug 2024

- Developed GenAI content-safety system combining Azure Content Safety with custom LLM workflows, automating manual review across global artist platforms while maintaining high accuracy
- Led end-to-end data migration for the Songkick app — rebuilt all pipelines, dimensional models, and ETL workflows with data catalog and documentation; redesigned the Looker ecosystem on Snowflake with clustering-based user segmentation for high-profile artists including Dua Lipa

AI/ML Engineer (Contract), SmartIQ

Dec 2022 – Jul 2023

- Engineered hybrid LLM + classical-ML pricing system (XGBoost, Random Forest with generative components), improving price-estimation accuracy across extensive vehicle listings
- Built multilabel BERT text-classification for hierarchical listing attributes; architected Airflow orchestration on GCP from scratch with automated scheduling, failure recovery, and model quantization for mobile deployment

Data Engineer & Analytics, Rollic Games (Zynga)

Jul 2021 – Jul 2022

- Engineered ETL pipelines from game servers and third-party APIs into Snowflake; built real-time player segmentation serving millions of DAU and a Looker suite for monetization, retention, and game-economy analytics
- Ran weekly A/B testing and churn analysis through funnel and behavioral clustering, driving product decisions that improved long-term retention

Data Scientist, Getir

Jul 2020 – Jul 2021

- Led end-to-end ML projects to production: dimensional modeling, complex SQL transformations, classification & clustering models deployed via Airflow on AWS
- Engineered real-time recommendation engine and predictive churn models serving millions of users — collaborative filtering and automated campaign systems that increased basket size and retention

Data Scientist, Groupe Renault

Aug 2019 – Aug 2020

- Architected predictive-maintenance platform on GCP with real-time FANUC robot telemetry (torque, vibration) and anomaly detection to minimize unplanned downtime
- Built energy-forecasting system across 11 manufacturing departments, benchmarking 5 time-series models (ARIMA, Prophet, LSTM, XGBoost, LightGBM) with systematic tuning

INDEPENDENT PRODUCTS

Case-LLM — Local-first GraphRAG system: Whisper transcription → LLM entity extraction → Neo4j knowledge graph + ChromaDB hybrid retrieval, with intent routing and graph-path hallucination validation · github.com/omerfarukballi/Case-LLM

Cosmic Atlas — Consumer astronomy app on a 13k-line C ephemeris engine validated against NASA JPL Horizons (~1 arc-minute accuracy); 380+ on-device calculation modules (React Native) · kozmetikatlaser.com

Sprout Way — Child-development tracking app (ages 0–72 months), live on the App Store — solo end-to-end: product, backend, mobile, launch · sproutway.app

EDUCATION

Galatasaray University — M.Sc. in Data Science (2021–2022) · Thesis: graph network analysis of urban mobility (Uber dataset)

Istanbul Technical University — B.Sc. in Economics (2015–2019)

TECHNICAL SKILLS

GenAI & LLMs: RAG & GraphRAG, Knowledge Graphs (Neo4j, Cypher), LangChain, LangGraph, Agentic AI, Multi-Model Orchestration (GPT, Claude, LLaMA, Gemma, Mistral), Vector Databases (ChromaDB, Pinecone, Weaviate), Semantic Search & Embeddings, Whisper, Ollama, Prompt Engineering, Hallucination Control (graph-path validation), LLM Security (prompt-injection defense, multi-tenant isolation, credential encryption)

ML & MLOps: XGBoost, LightGBM, PyTorch, TensorFlow, Transformers & BERT (multi-label NLP), Time Series (Prophet, ARIMA, LSTM), Recommendation Systems, Behavioral Clustering & Segmentation, Anomaly Detection, Entity Resolution, MLflow, Model Monitoring & Data Drift, Quantization, GPU Workloads (H100), A/B Testing

Data Platforms: Snowflake (warehousing, SCD Type 2 modeling, stored procedures), BigQuery, Redshift, Databricks, PostgreSQL, MongoDB, Redis

Data Engineering: Airflow, Matillion, dbt, Kafka, PySpark, SSIS, ETL/ELT, Dimensional Modeling (Star Schema, Medallion), Data Quality & Governance, Master Data Management, Docker, Kubernetes

Enterprise Systems: SAP S/4HANA Master Data (Align S/4 migration), Salesforce, Workday, Concur & Wise integration, SIT/UAT & cutover management

Cloud & BI: AWS (EMR, Glue, SageMaker, Lambda), GCP (Vertex AI), Azure (ML, Content Safety), Looker, Power BI, Tableau

Mobile & Product: React Native (Expo, Skia, Reanimated), native module bridging (C/C++, JNI, Objective-C++), Supabase, App Store publishing & TestFlight distribution

Languages: Python (Pandas, NumPy, Scikit-learn, Flask, Streamlit), SQL (Snowflake, T-SQL), TypeScript/JavaScript (Next.js, React, REST APIs), C/C++ (native modules)