

BELCE KAYA

Lead AI/MLOps Engineer

belcekaya@gmail.com | +34 613 02 99 19 | Spain | [LinkedIn](#) | [Portfolio](#)

PROFESSIONAL SUMMARY

AI/MLOps Engineer with MSc in Big Data Analytics. Over six years of experience across the full AI/ML stack - from building multi-agent systems, advanced RAG pipelines, and ML/DL models to designing cloud infrastructure, LLM Ops platforms, CI/CD automation, and production observability. Currently leading multi-team AI engineering efforts at AstraZeneca, building shared platform capabilities and developer tooling that enable teams to ship AI products faster. Native Turkish, fluent in English and Spanish.

EXPERIENCE

Lead AI/MLOps Engineer | AstraZeneca

Barcelona, Spain • Aug 2024 – Present

- Leading R&D projects and multi-team AI engineering efforts, building the AI platform infrastructure that product and research teams depend on for developing AI applications, serving 3+ cross-functional teams and reducing time-to-production for new AI features by ~40%.
- Architected and developed a framework-agnostic agentic interface using Domain-Driven Design with hexagonal architecture, creating a reusable platform layer enabling rapid experimentation across multiple frameworks (DSPy, Anthropic SDK, LangGraph, Agno) with full streaming support - centralizing repeated components and cutting new agent development time by ~50%.
- Implemented LiteLLM as the unified LLM proxy layer, enabling centralized model access, cost tracking, and governance across 3+ LLM providers. Configured rate-limit-aware routing with TPM/RPM fallback strategies and guardrails for AI services in production.
- Designed and implemented a GitHub-based Ticket System for infrastructure provisioning on AWS using Terraform as IaC, enabling teams to self-provision platform resources through standardized templates via GitHub Actions. Set up CI/CD pipelines with template repositories and OIDC-based deployment roles, reducing provisioning lead time from days to under 1 hour.
- Built end-to-end evaluation and observability pipelines combining custom metrics (DCG) with DeepEval LLM-as-a-judge, improving retrieval accuracy by ~25%. Integrated LangSmith with OpenTelemetry for production monitoring of latency, throughput, and token usage across all AI services.
- Engineered scalable backend services using FastAPI with mounted MCP servers, deployed on AWS ECS. Built agent-as-a-service patterns enabling headless agent invocation via external API; implemented prompt optimization using DSPy and caching strategies (ElastiCache), reducing average chatbot response latency by ~35%. Studied different query techniques in Opensearch (e.g. bm25, hybrid, Neural Sparse, knn etc.).
- Developed production GenAI chatbots and advanced multi-agent RAG systems including clinical trial intelligence (ClinicalTrials.gov, PubMed, internal datasets), drug repurposing via Biomedical Knowledge Graph MCP server, and a document authoring portal for clinical protocols - adopted by 50+ scientists across the Rare Disease Unit.
- Led OCR/document parsing evaluation (AWS Textract, MinerU/VLM, Docling) for complex clinical tables; deployed reranker and VLM OCR models from Hugging Face as SageMaker endpoints using JFrog Artifactory and vLLM.

Senior Data Scientist | Mercedes-Benz AG

Madrid, Spain • Oct 2022 – Oct 2024

- Developed and deployed ML forecasting models using tree-based algorithms and Temporal Fusion Transformers in AWS SageMaker for multivariate time series prediction, managing the full model lifecycle from training to production.
- Built price optimization system with log-log linear regression to forecast demand elasticity, delivering Power BI dashboards for KPI monitoring in pilot markets.
- Led cloud migration by refactoring PySpark codebases and migrating data infrastructure from Azure to AWS, improving scalability and reliability.
- Established data quality framework defining testing blueprints for data pipeline transformations within Agile/Scrum methodology. Won Best Forecasting Award at Mercedes-Benz FST Hackathon.

Data Consultant | Satec Group

Madrid, Spain • Nov 2021 – Oct 2022

- Automated water management using XGBoost for event detection; built Apache Nifi/Solr data pipelines for HL7 health data processing.
- Implemented Neo4j graph databases with Cypher querying for healthcare data modeling and analysis.

Business Intelligence Developer | Yapı Kredi Technology

Istanbul, Turkey • Oct 2020 – Sep 2021

- Built enterprise BI dashboards in Power BI with advanced DAX queries; implemented end-to-end ETL processes using SSIS/SSMS with T-SQL/PL-SQL pipeline management.

TECHNICAL SKILLS

AI / GenAI	RAG, Multi-Agent Systems, LangGraph, Agno, DSPy, Anthropic SDK, LiteLLM, MCP, A2A, Claude Code, Agent Skills, Lovable, ElevenLabs, Pipecat, OCR (Textract, MinerU, Docling, PyMuPDF, PaddleOCR)
Cloud & Infra	AWS, Azure, Terraform, Docker, Kubernetes, MongoDB, DynamoDB, Neo4j, Elasticsearch, Apache Solr, Databricks
LLM/ML Ops	CI/CD Pipelines, GitHub Actions, AWS CodePipeline, LLM Evaluation Frameworks, DeepEval, LangSmith, Langfuse, OpenTelemetry, MLFlow, JFrog Artifactory
ML & Inference	Scikit-Learn, PyTorch, TensorFlow, Keras, MLlib, Hugging Face, vLLM, LLM-D, MLX, Ollama
Languages	Python, TypeScript, PySpark, R, SQL, JavaScript, Cypher, HTML, DAX
Web Frameworks	FastAPI, Flask, Streamlit, HTMX
BI & Viz	Power BI, Tableau, Qlik, Gephi
Version Control	Git, GitHub Actions

EDUCATION

MSc in Big Data Analytics (GPA: 8.7/10) — Universidad Carlos III de Madrid	2021 – 2022
BSc in Industrial Engineering (GPA: 3.62/4) — Yıldız Technical University, EUR-ACE® Accredited	2015 – 2020
Bachelor of Business Administration (Erasmus+) — Università degli Studi di Perugia	2020

LANGUAGES & CERTIFICATIONS

LANGUAGES	AWARDS & CERTIFICATIONS
Turkish — Native	Best Forecasting — Mercedes-Benz FST Hackathon
English — Fluent (IELTS 6.5)	Highest-Ranked Graduate
Spanish — Proficient	AWS Certified Data Analytics Specialty
German — Beginner	ML Model Deployment
Italian — Beginner	NLP & Deep RL