Mehmet Taha Bağatur

Platform Engineer · DevOps / MLOps · AI in production

Bursa, Türkiye (UTC+3) · mt.bagatur@gmail.com · +90 507 438 83 51 · linkedin.com/in/mtbagatur · github.com/mtbagatur

SUMMARY

Platform engineer working end to end across models, infrastructure, edge hardware and live users at a large retail group's R&D center. As the unit's sole DevOps engineer, I run a ~166-server estate and own deployment, monitoring, alerting and logging for AI products used by live retail operations.

I started in computer vision, NLP and LLM/RAG, then moved into production delivery; I also own tech decisions, vendor evaluations and management presentations.

CORE SKILLS

- **MLOps & AI:** MLflow, Vertex AI Pipelines (KFP), containerised model serving, vLLM, Triton (pre-production), drift monitoring, model explainability (SHAP); PyTorch, TensorFlow, Hugging Face, OpenCV, scikit-learn; CV, NLP, LLM/RAG, VLM
- **CI/CD & IaC:** Azure DevOps Server & Services (YAML/classic pipelines, agents, administration, REST API), GitLab CI, GitHub Actions, Cloud Build, Terraform, Ansible, Helm
- **Infrastructure:** on-prem Kubernetes (kubeadm), Docker, VMware, Azure, Google Cloud (Vertex AI, Cloud Run, BigQuery — project), HPE Private Cloud AI/AI Server (GreenLake), NVIDIA Jetson (Orin NX, CUDA, TensorRT); Linux and Windows Server
- **Observability:** Prometheus, Alertmanager, Grafana (as code), Loki, Grafana Alloy, ELK/Kibana, OpenSearch, Fluent Bit, blackbox/node/Windows/DCGM exporters, DORA
- **Security, networking & data:** least privilege, secrets management, Let's Encrypt (HTTP-01/DNS-01), Nginx, IIS, TLS, DNS, Cloudflare, DMZ/VLAN · PostgreSQL, SQL Server, MongoDB, Redis, RabbitMQ, MinIO
- **Languages & ML:** Python, Bash, PowerShell, C++; object detection, image classification, intent recognition, signal analysis

EXPERIENCE

Özdilek Holding — Özveri R&D Center Jul 2022 – present
Bursa, Türkiye

DevOps & MLOps Engineer Jun 2024 – present

AI PRODUCTS IN PRODUCTION

- To make AI products dependable in live operations, I own deployment, monitoring, alerting and logging across the R&D center's seven products.
- **ShopperLine:** retail vision with 10 CV modules serving 500+ cameras on 50+ Jetson edge devices; I own deployment and monitoring and contributed to model work.
- **Enterprise RAG:** deployed and operate an air-gapped retrieval platform, keeping inference and document processing on-prem.
- **ATLAS:** deployed and operate an air-gapped multimodal RAG platform and its five domain assistants (94% recall, 91% faithfulness).
- **Demand forecasting:** platform serving 19 venues with MAPE below 10%; I contributed MLOps including retraining, drift monitoring and MLflow versioning.
- **ÖzdilekWallet:** face-recognition payment pilot operating below 300 ms; I contributed to model work and supported platform infrastructure.
- **Smart Shelf:** VLM planogram/price-tag checks across a 40-camera pilot on HPE Private Cloud AI (4× RTX PRO 6000); I contributed platform/MLOps and connected it to central monitoring.
- **TagSpot:** mobile price-tag checks across 19+ stores, processing 500K+ scans/month at 96%+ accuracy and 15–20 FPS on-device; I supported deployment, monitoring and model optimisation.
- **Fun Station:** DevOps across backend, frontend and AI services in 19 venues, plus MLOps (retraining, drift, MLflow); also contributed LLMOps to agentic workflows and the meeting assistant.

PLATFORM & RELIABILITY

- To make delivery repeatable across a broad estate, I run ~166 servers and on-prem Kubernetes as sole DevOps engineer, administer Azure DevOps Server (~80 projects, ~460 build and release definitions), and moved ~110 pipelines to a shared template with approved deploys, health checks and automatic rollback.
- Own DevOps process decisions; guide colleagues; evaluate products with vendors and present recommendations. Weekly live-data management reports combined a roadmap, decisions needed, a 10-area (1–5) maturity model (1.1→2.4 in three weeks) and a risk register; built a live-data report generator and tracked DORA metrics weekly.
- Won approval for a zero-downtime Cloudflare DNS migration by presenting three options; restructured the DevOps wiki (52 pages) with a standard adopted by 8 project wikis, plus runbooks, postmortem template, RACI and onboarding docs; led English-language vendor GPU incident RCA through disk replacement and coordinated cross-team production changes.
- Expanded monitoring to 121 endpoint checks (~7,700 services), cut noisy alerts ~45→0 and centralised logs from 64 servers (Loki/Alloy, environment retention, 5xx/OOM/SSH alerts); reduced OpenSearch/Fluent Bit shards 1,009→599.
- Stopped alert loss and reduced operational noise: restored delivery for 151 missing notifications, cut noisy alerts ~45→0, and grew actionable rules 51→133; put a runbook behind each remaining alert.
- Reduced access risk by auditing 17 PostgreSQL instances/355 databases, where 92% of connections used superuser; designed least-privilege roles, rotation and rollback, with previews passing in three environments.

- Made recovery measurable with a SQL Server backup plan and weekly restore checks; brought 73 servers under Ansible and managed 29 Cloudflare DNS records plus 53 Grafana dashboards through Terraform plans and approved applies.
- Automated Let's Encrypt HTTP-01/DNS-01 certificate flows and moved DNS to Cloudflare with zero downtime; reduced OpenSearch shards from 1,009 to 599.
- Turned operational evidence into a management roadmap: maturity rose 1.1→2.4 in three weeks; initial DORA baseline was 2.2 deployments/week, ~2.5-hour lead time and 2.9% change failure rate.
- Restored a full Kubernetes control-plane outage and led a DMZ/VLAN migration for a production web platform; coordinated certificate, firewall and maintenance changes across systems, network, ERP and development teams.
- Introduced versioned images, an on-prem Nexus registry and GPU pre-checks for the store GPU fleet, stopping driver/library mismatches across 25 stores; administered ~166 Linux/Windows servers, DMZ web gateways and HPE AI infrastructure (4× RTX PRO 6000).

AI / ML Engineer (AI Engineer → ML Engineer) Jul 2022 – Jun 2024

- Shipped NLP, computer vision and LLM applications to production and edge, including KASED (internal AI assistant / corporate memory) and a RAG-based Q&A assistant over factory manuals.
- Built LLM workflow automations for reconciliation review, contract pre-approval and a RAG meeting assistant.
- Introduced MLflow for experiment tracking/model lifecycle; containerised ML workloads, set up CI/CD with DevOps and deployed lightweight LLMs on-prem and at the edge. Built Python/C++ pipelines for object detection, image classification, intent recognition and signal analysis.
- For ShopperLine retail vision, I collected data, trained custom YOLO models, adapted ByteTrack and optimised inference for Jetson as part of the team building the product.
- Edge AI optimisation with quantisation, pruning and TensorRT contributed to ~75% lower hardware cost; I contributed model optimisation for the edge devices.
- Contributed to model work for the face-recognition payment pilot and to edge AI optimisation.
- Contributed to projects that resulted in 4 patent applications.

SELECTED OPEN-SOURCE WORK

- **llm-prod-readiness-kit** — Prevents RAG regressions and data leaks with traceable eval gates, fallback and PII masking; layered defenses held successful red-team attacks to 0/10.
- **llm-serving-loadlab** — Pinpoints whether LLM latency comes from prompt prefill or load and tests serving capacity; a second replica raised throughput 1.49× at 8 concurrent requests.
- **redteam-check** — Surfaces staging-chatbot prompt leaks, unsafe replies and data exposure through 40 black-box attacks, with a readable report and JUnit results for CI.
- **open-ops-lab** — Turns Kubernetes alerts into grouped, read-only incident hypotheses with prompt-injection tests; kind verification folded 3 alerts into 1 incident.
- **gcp-mlops-capstone** — Blocks weak model releases with a Vertex AI quality gate and serves explained, drift-monitored predictions; test ROC AUC reached 0.9196.

PUBLICATIONS & SPEAKING

Presented the ICAIS 2024 paper and the ICOLES 2023 top-view paper; chaired all sessions at ICOLES 2023, Antalya.

- Preprocessing Optimization for Enhanced Real-Time Detection and Tracking in Vision Systems — *ICAIS 2024*, Istanbul, Nov 2024
- An Efficient Framework for Real-Time Age and Gender Detection Using Deep Learning Techniques — *ICOLES 2023*, Antalya, Nov 2023 (co-author; presented by a colleague)
- Design and Performance Analysis of Deep Learning-Based People Detection from Top-View Images — *ICOLES 2023*, Antalya, Nov 2023

EDUCATION & LANGUAGES

- B.Sc. Electrical and Electronics Engineering — Bursa Uludağ University
- Graduate coursework in Computer Engineering (M.Sc. program in progress) — began at Yıldız Technical University; continued at Bursa Uludağ University
- Turkish (native) · English (professional working proficiency)