

# Bekhruz Kasymov

## LEAD PLATFORM ENGINEER - AI INFRASTRUCTURE & RELIABILITY

Tbilisi, Georgia | [bkasymov@outlook.com](mailto:bkasymov@outlook.com) | [LinkedIn](#) | [GitHub](#) | [bkasymov.info](http://bkasymov.info)

### PROFESSIONAL SUMMARY

Lead Platform Engineer with 5+ years of experience building reliable infrastructure for international AI, HealthTech, TravelTech, InsurTech, and banking products. Operated 8 Kubernetes clusters and 30+ microservices across five continents, supporting products used by 333K+ active users and approximately 400 clinics while reducing release time by 80%, cloud spend by 40%, and incident response time by 50%.

### CORE SKILLS

**Infrastructure:** Kubernetes, GCP, Azure, Linux, Docker, PostgreSQL, hybrid and multi-region systems

**Delivery:** Terraform, Terragrunt, Ansible, Helm, ArgoCD, GitLab CI/CD, Jenkins, GitHub Actions, GitOps

**Reliability:** Prometheus, Grafana, Loki, Alertmanager, Sentry, Trivy, SAST/DAST, RBAC, ingress and certificates

**AI and data:** vLLM, Ollama, ClearML, KubeRay, GPU workloads, Python, MongoDB, Redis, Kafka, RabbitMQ

### PROFESSIONAL EXPERIENCE

#### Confidential AI Platform | Lead Platform Engineer

2026 - Present | International | Hybrid cloud

- Own a hybrid Kubernetes platform across self-hosted development, Azure AKS production, and Nebius GPU infrastructure for AI workloads.
- Operate production infrastructure serving 333K+ active users across a large portfolio of AI generation and productivity workflows.
- Completed the Bitbucket-to-GitLab migration, including repositories, backups, protected reviews, CI/CD rewiring, and Azure Kubernetes runners.
- Delivered highly available PostgreSQL after an Azure database outage caused mass 5xx responses and authentication failure.
- Hardened production ingress against spoofed client-IP headers and investigated WAF and trusted-proxy boundaries.
- Resolved recurring GPU/NVML and AI-generation incidents and operated observability across Prometheus, Grafana, Loki, Sentry, Graylog, and Fluent Bit.

#### Diagnocat | Senior DevOps Engineer

2025 - 2026 | International | Multi-region production

- Managed three outsourced DevOps engineers and coordinated infrastructure strategy across frontend, backend, DevOps, and MLOps teams.
- Architected and operated 8 Kubernetes clusters on GCP and Yandex Cloud, running 30+ microservices for approximately 400 dental clinics across five continents.
- Reduced release time from 4-6 hours to under one hour (80%) and incident response time by 50% through delivery automation and observability.
- Cut cloud spend by 40% through resource cleanup and spot autoscaling; built 30+ GitLab CI/CD pipelines with security and GitOps controls.
- Deployed 4+ production AI models with KubeRay, Ollama, vLLM, ClearML, and GPU autoscaling, reducing external AI API costs by 4x.

#### WeGoTrip | Senior DevOps Engineer

2023 - 2025 | International | Remote

- Owned infrastructure for 8+ services across four cloud providers, supporting 10+ engineers and approximately 6,000 customers with zero unplanned downtime.
- Built and maintained 15+ CI/CD pipelines and reduced cloud costs by up to 70% through right-sizing, provider selection, Terraform, and Terragrunt automation.

#### Qantev | DevOps Engineer - Contract

2024 | International | Paris, France

- Delivered an AWS-to-VK Cloud migration in four months while maintaining delivery continuity.
- Built Azure infrastructure with Terraform and Terragrunt for an international AI health-insurance platform trusted by 20+ insurer clients worldwide; reduced build time by 30%.

#### Sberbank | DevOps Engineer / Software Engineering Intern

2020 - 2022 | Moscow, Russia

- Built and maintained 8+ Jenkins pipelines using GitLab, SonarQube, Python, and Shell automation.
- Trained two teams (8+ engineers) and reduced deployment errors by 30% through repeatable automation and quality controls.

## SELECTED ADDITIONAL ENGAGEMENTS

**WeChargers | Infrastructure consulting.** Production reliability, Docker-based service operations, observability, deployment safety, and database/application performance investigation for a power-bank rental platform.

**WeGoTrip | Post-employment consulting.** Returned for a separate project-based infrastructure engagement, providing delivery continuity in the company's multi-cloud production environment.

## SELECTED PRODUCT

**FailUp | Founder and Product Engineer.** Built a full-stack AI personal-growth platform across React, Node.js, Flutter, MongoDB, Redis/BullMQ, Qdrant, and OpenAI APIs, with semantic search, asynchronous AI workflows, encryption, data export, and self-hosted K3s operations.

## VOLUNTEER EXPERIENCE

**Legal Consultant | Kutafin Moscow State Law University (MSAL)** | Sep 2015 - Dec 2017 | Human Rights - Provided pro bono legal assistance to socially vulnerable communities.

**Assistant to an Associate Director | Beeline Russia** | Sep 2017 - Jan 2018 | Science and Technology - Created, maintained, and processed technical documentation.

**Technical Consultant | Gareni / Pamix / Tabler** | Sep 2021 - Present | Science and Technology - Advise colleagues on programming languages, databases, automation, and optimization.

**Distributor | Linkee** | May 2023 | Social Services

**Administrative Assistant | STATION F** | Jun 2023 | Arts and Culture

**Volunteer Staff | Viva Technology** | Jun 2023 | Science and Technology

## EDUCATION

**42 Paris** | RNCP Level 7, Expert in IT Architecture - Information Systems & Networks | 2022 - 2025

**School 21 by Sber** | Software Engineering Program | 2019 - 2022

**Kutafin Moscow State Law University** | Law | Graduated 2018

## LANGUAGES

Russian | Tajik | English