

Vladimir Kulyashov

MLOps / LLM Ops Engineer

Perm @kulyashov koolvn cv.hl59.ru

Not looking for a job, but always happy to have an interesting conversation

About

I'm driven by applying machine learning to real problems. I've been in ML since 2018, starting with analytics and computer vision, so I know first-hand what keeps models from reaching production.

Closing the gap between research and production brought me to MLOps and DevOps. I build CI/CD and GitOps, GPU Kubernetes infrastructure, LLM inference, monitoring and logging for ML services, and internal tooling for ML teams.

I love the engineering side of things: turning ML ideas into products that just work. A background in finance and project management helps me see the whole picture. I keep growing in both MLOps and ML.

Experience

MLOps / LLM Ops Engineer, GNIVC

Nov 2025 – present

Moscow

I develop the GNIVC AI Sandbox, the company's internal environment for AI experiments and applied AI services (1,000+ users). I'm responsible for model inference in the Federal Tax Service of Russia's pilot and production environments, and I take part in designing the FTS AI Platform.

- Deployed and run an LLM inference service with an OpenAI-compatible API on vLLM, SGLang and NVIDIA Dynamo (Qwen 3.x, DeepSeek V4 and others). Moved production models between engines and schedulers with no loss of availability; fixed degradation under load and stuck queues.
- Optimise inference: continuous batching, prefix and FP8 KV cache, queue policies and overload protection. Median TTFT of the key model improved about 3x (~1.1 → 0.4 s), and FP8 nearly doubled KV cache capacity.
- Deploy, update and maintain models on inference servers in the FTS pilot and production environments; benchmark models for speed and cognitive ability.
- Take part in designing the FTS AI Platform: concept and design documents, model runtime architecture (Prefill/Decode disaggregation), GPU selection and allocation, resource requirements.
- Designed the sandbox Kubernetes cluster with a GPU node pool (NVIDIA H100, H200, RTX 6000 Ada), GPU Operator, KAI Scheduler and Grove, Longhorn, Calico, MetalLB and Envoy Gateway; automated node configuration with Ansible.
- Deploy and upgrade sandbox services without downtime: Open WebUI, GPUStack, JupyterHub, Qdrant, a RAG stack, MinIO, ClearML, Langfuse, BI, n8n automations. Upgraded Langfuse from 3.x to 4.x, migrating ~1.9M events with no data loss.
- Selected and deployed a decision model in the sandbox: compared open options (Kev, Laya, Clef and others) on one task set and chose decisions 3.0.
- Built the CI/CD and GitOps setup: Argo CD with Keycloak SSO, image builds on Argo Workflows and kaniko into Artifactory, Bitbucket events via Argo Events, Vault with External Secrets Operator, Kyverno policies.
- Own platform security: hardened images (vLLM, SGLang, DCGM Exporter and others) to zero fixable CRITICAL/HIGH findings in Trivy, restricted pod profiles, network policies, work with the security team.
- Run observability and operations: VictoriaMetrics, Grafana, Loki, Vector, a service status page, incident reviews, capacity planning for GPU, memory and storage.
- Built Model Router, an in-house HTTP router for vLLM/SGLang with load balancing, failover, retries and metrics.
- Support the sandbox handover: preparing for acceptance and migration to another site, deployment guides for the operating organisation, user support.

Kubernetes, NVIDIA GPU Operator, KAI Scheduler, Grove, NVIDIA Dynamo, vLLM, SGLang, GPUStack, Helm, Argo CD, Argo Workflows, Argo Events, Vault, External Secrets, Kyverno, Keycloak, Longhorn, Calico, MetalLB, Envoy Gateway, VictoriaMetrics, Grafana, Loki, Vector, n8n, Langfuse, ClearML, MinIO, Qdrant, PostgreSQL, Ansible, Trivy, Go, Python

MLOps Engineer, Rusagro Group

Dec 2023 – Oct 2025

Moscow

Developed MLOps and DevOps practices in the ML & AI division, supporting 3 teams and the infrastructure for the full model lifecycle.

- Moved ML applications to Docker, laying the groundwork for Kubernetes.
- Set up two Kubernetes dev clusters (1 and 6 nodes) for computer vision and classic ML services.
- Built GitLab CI templates (build, tests, vulnerability scanning, deploy), cutting pilot launches from months to a few weeks.
- Introduced SAST and container scanning with Trivy.
- Deployed the core MLOps stack: ClearML, MiniIO (S3), HashiCorp Vault, Prometheus, Grafana, Loki, Alloy.
- Configured infrastructure with Ansible together with DevOps engineers and the infrastructure team.
- Built and maintained an internal utility library that standardised ML development.
- Sat on architecture committees, reviewed solution designs and code, and wrote recommendations on MLOps practices.

Senior Computer Vision Engineer, Rusagro Group

Apr 2023 – Dec 2023

Moscow

Development automation and architecture of CV services.

- Built and maintained an internal computer vision library.
- Made architecture decisions for CV services and helped plan new projects.
- Deployed experiment tracking.
- Designed a storage and catalogue system (DWH) for labelled and unlabelled data.

Python, PyTorch, OpenCV, ClearML, Docker

Computer Vision Engineer, Rusagro Group

May 2022 – Apr 2023

Moscow

Projects for the meat business – counting livestock and spotting sick or dead animals.

- Together with the team, brought pig counting accuracy in pens to 99+%.
- Built data collection pipelines and trained models.

Python, PyTorch, OpenCV, TensorRT

Lead ML Engineer, Forpost

Jul 2021 – May 2022

Perm

Video analytics, shipping models to production and leading a group of ML engineers.

- Set up and ran cloud infrastructure for neural network training.
- Wrote custom layers and implemented algorithms from recent computer vision papers.
- Standardised ML pipelines and shipped models to production.
- Hired, mentored and led ML engineers, set and tracked tasks, and handled technical talks with clients.

Python, TensorFlow, PyTorch, Darknet, OpenCV, TensorRT, TF Serving, GCP, AWS

ML Engineer, Forpost

Aug 2020 – Jul 2021

Perm

Detection, segmentation and classification models for video analytics.

- Collected and labelled data, fine-tuned and improved models, prepared them for production.
- Wrote tools and scripts for data collection and processing.

Python, TensorFlow, Darknet, OpenCV

Product Analyst, Dom.ru Movix, ER-Telecom Holding

Mar 2020 – Aug 2020

Perm

Product analytics for the Movix TV service.

- Tested hypotheses, computed product metrics, segmented the customer base.
- Analysed user search queries and subscriber traffic.
- Designed and populated the analytics team's database and automated reporting.

Python, SQL, pandas, Statistics

Teaching

MLOps Lecturer, Perm State University

Sep 2024 – Jun 2025

Taught a master's course on MLOps and integrating machine learning into web applications.

- Designed lectures and labs on automating the model lifecycle: data, training, experiment tracking, deployment.
- Supervised the capstone: from data collection to a model behind an API with performance and quality metrics.
- Taught students tools for automating and monitoring ML projects, including experiment tracking and CI/CD.

Projects

This website

Oct 2026 – present

An interactive CV with a "System 1 / System 2" AI agent: a fast typed decision model picks the route and filters off-topic requests, and Claude via LiteLLM reasons and answers using tools and cards.

- A fast decision model as System 1, with a chain of fallback links.
- The answering model is chosen per route.
- LiteLLM proxy, rate limits and budgets, Langfuse tracing.
- RU/EN, an admin panel with AI profile import, GitLab CI/CD on a self-hosted runner.

cv.hl59.ru

Pydantic AI, FastAPI, Next.js, LiteLLM, Claude, Langfuse, PostgreSQL, Docker, GitLab CI

LLM inference service

Nov 2025 – present

LLM inference with an OpenAI-compatible API for GNIVC staff and in the Federal Tax Service environments.

- vLLM, SGLang and NVIDIA Dynamo; Qwen 3.x, DeepSeek V4 and other models; engine migrations with no loss of availability.
- Median TTFT of the key model improved about 3x (~1.1 → 0.4 s).
- FP8 KV cache nearly doubled cache capacity.

vLLM, SGLang, NVIDIA Dynamo, Kubernetes

Model Router

Nov 2025 – present

An in-house HTTP router for vLLM/SGLang.

- Load balancing, failover, retries and metrics.

vLLM, SGLang

AI Sandbox Kubernetes cluster

Nov 2025 – present

The GPU cluster behind GNIVC's internal AI sandbox, serving 1,000+ users.

- NVIDIA H100, H200 and RTX 6000 Ada nodes with GPU Operator, KAI Scheduler and Grove.
- Longhorn, Calico, MetalLB, Envoy Gateway; nodes configured with Ansible.
- Zero-downtime services: Open WebUI, GPUStack, JupyterHub, Qdrant, ClearML, Langfuse; Langfuse 3.x → 4.x upgrade migrating ~1.9M events.

Kubernetes, NVIDIA GPU Operator, KAI Scheduler, Grove, Longhorn, Calico, MetalLB, Envoy Gateway, Ansible, GPUStack, Qdrant, Langfuse

Sandbox GitOps and security

Nov 2025 – present

A CI/CD and GitOps setup with secrets management, policies and image scanning.

- Argo CD with Keycloak SSO, builds on Argo Workflows and kaniko into Artifactory, Bitbucket events via Argo Events.
- Vault with External Secrets Operator, Kyverno policies.
- Images hardened to zero fixable CRITICAL/HIGH findings in Trivy.

Argo CD, Argo Workflows, Argo Events, Keycloak, Vault, External Secrets, Kyverno, Trivy

Federal Tax Service AI Platform

Nov 2025 – present

Designing a platform for tens of thousands of inspectors, hundreds of thousands of citizens and high-load automated systems.

- Concept and design documents.
- Model runtime architecture with Prefill/Decode disaggregation, GPU selection and allocation per model class.

NVIDIA Dynamo, Kubernetes

MLOps course for master's students

Sep 2024 – Jun 2025

A course on the full lifecycle of an ML system, with a capstone project.

Python, Docker, FastAPI, CI/CD

Kubernetes platform for ML

Dec 2023 – Oct 2025

From "it runs on my laptop" to containers and two Kubernetes dev clusters for ML services.

- One Docker-based runtime for all ML teams.
- 1-node and 6-node clusters for CV and classic ML.

Kubernetes, Docker, Helm, Ansible

CI/CD templates and security

Dec 2023 – Oct 2025

Reusable GitLab CI stages for ML projects with built-in vulnerability checks.

- Build, tests, SAST and container scanning with Trivy, deploy.
- Pilot launches went from months to a few weeks.

GitLab CI, Trivy, Docker

MLOps stack

Dec 2023 – Oct 2025

Experiment tracking, artifact storage, secrets and observability for ML services.

- ClearML for experiments, tasks and agents; MinIO for artifacts; Vault for secrets.
- Prometheus, Grafana, Loki and Alloy for ML service metrics and logs.

ClearML, MinIO, Vault, Prometheus, Grafana, Loki, Alloy

Computer vision on farms

May 2022 – Dec 2023

Counting livestock and spotting sick or dead animals from video.

- 99+% accuracy counting pigs in pens.
- Data collection, training and optimisation of detection models.
- A DWH for labelled and unlabelled data.

PyTorch, OpenCV, TensorRT

Video analytics

Aug 2020 – May 2022

Detection, segmentation and classification models for video analytics, from labelling to production.

- Implementing algorithms from fresh papers and custom layers.
- Cloud training infrastructure on GCP and AWS.

TensorFlow, Darknet, TensorRT, TF Serving, GCP, AWS

Skills

LLM & inference	vLLM, SGLang, NVIDIA Dynamo, GPUStack, NVIDIA GPU Operator, KAI Scheduler, Grove, LiteLLM, Pydantic AI
MLOps & infrastructure	Kubernetes, Docker, Helm, GitLab CI, Argo CD, Argo Workflows, Argo Events, Ansible, ClearML, MinIO, Vault, External Secrets, Kyverno, Keycloak, Longhorn, Calico, MetalLB, Envoy Gateway, n8n, Prefect, Trivy
Observability	Prometheus, Grafana, Loki, VictoriaMetrics, Vector, Langfuse
ML & computer vision	PyTorch, TensorFlow, OpenCV, TensorRT, TF Serving, Darknet, pandas, NumPy
Languages & fundamentals	Python, SQL, Bash, Go, Linux, Git, Statistics

Cloud & data

GCP, AWS, PostgreSQL, MongoDB, Qdrant

Education

- 2008 Real estate appraisal and management. Perm State Technical University
- 2021 Neural networks in computer vision. Perm State University, Regional Institute of Continuing Education
- 2020 Data Scientist program. Skillbox
- 2020 Professional Scrum Developer. Agilix consulting
- 2019 mlcourse.ai – machine learning. Open Data Science

Languages

Russian – native English – C1