

Владимир Куляшов

MLOps / LLM Ops инженер

Пермь @kulyashov koolvn cv.hl59.ru

Работу не ищу, но всегда рад интересному разговору

О себе

Меня вдохновляет возможность применять машинное обучение для решения реальных задач. В ML с 2018 года, начинал с аналитики и компьютерного зрения, поэтому хорошо понимаю, что мешает моделям доехать до продакшена.

Желание закрыть разрыв между исследованиями и production привело меня в MLOps и DevOps. Строю CI/CD и GitOps, Kubernetes-инфраструктуру с GPU, инференс LLM, мониторинг и логирование ML-сервисов, внутренние инструменты для ML-команд.

Люблю инженерную сторону процесса: превращать ML-идеи в стабильно работающие продукты. Опыт в финансах и управлении проектами помогает смотреть на задачи комплексно. Постоянно развиваю навыки и в MLOps, и в ML.

Опыт

MLOps / LLM Ops инженер, ГНИВЦ

ноя 2025 – н. в.

Москва

Развиваю ИИ-Песочницу ГНИВЦ – внутреннюю среду для ИИ-экспериментов и прикладных ИИ-сервисов (более 1000 пользователей). Отвечаю за инференс моделей в контурах опытной (КОЭ) и промышленной (КПЭ) эксплуатации ФНС России, участвую в проектировании ИИ-Платформы ФНС России.

- Развернул и поддерживаю сервис инференса LLM с OpenAI-совместимым API на vLLM, SGLang и NVIDIA Dynamo (Qwen 3.x, DeepSeek V4 и другие). Переводил продовые модели между движками и планировщиками без потери доступности, устранял деградацию под нагрузкой и «залипание» очередей.
- Оптимизирую инференс: непрерывный батчинг, префиксный и FP8 KV-кэш, политики очередей и защита от перегрузки. Медианный TTFT ключевой модели улучшился примерно в 3 раза (~1,1 → 0,4 с), ёмкость KV-кэша за счёт FP8 выросла почти вдвое.
- Разворачиваю, обновляю и сопровождаю модели на серверах инференса в КОЭ и КПЭ ФНС России; провожу бенчмарки моделей на скорость и когнитивные способности.
- Участвую в проектировании ИИ-Платформы ФНС России: концепция и проектная документация, архитектура среды исполнения моделей (Prefill/Decode disaggregation), подбор и распределение GPU, требования к ресурсам.
- Спроектировал Kubernetes-кластер песочницы с пулом GPU-узлов (NVIDIA H100, H200, RTX 6000 Ada), GPU-оператором, KAI Scheduler и Grove, Longhorn, Calico, MetalLB и Envoy Gateway; конфигурацию узлов автоматизировал через Ansible.
- Разворачиваю и обновляю без простоя сервисы песочницы: Open WebUI, GPUStack, JupyterHub, Qdrant, RAG-стек, MinIO, ClearML, Langfuse, BI, автоматизации на n8n. Обновил Langfuse с 3.x до 4.x с переносом ~1,9 млн событий без потери данных.
- Выбрал и развернул в песочнице модель решений: сравнил открытые варианты (Kev, Laya, Clef и другие) на одном наборе заданий и остановился на decisions 3.0.
- Построил CI/CD и GitOps-контур: Argo CD с SSO через Keycloak, сборка образов на Argo Workflows и kaniko в Artifactory, события из Bitbucket через Argo Events, Vault и External Secrets Operator, политики Kyverno.
- Веду безопасность контура: довёл образы (vLLM, SGLang, DCGM Exporter и др.) до 0 исправимых CRITICAL/HIGH по Trivy, restricted-профили подов, сетевые политики, взаимодействие с ИБ.
- Отвечаю за наблюдаемость и эксплуатацию: VictoriaMetrics, Grafana, Loki, Vector, статус-страница сервисов, разбор инцидентов, планирование ёмкости GPU, памяти и хранилища.
- Разработал Model Router – собственный HTTP-роутер для vLLM/SGLang с балансировкой, отказоустойчивостью, повторными попытками и метриками.
- Сопровождаю сдачу песочницы: подготовка к приёмке и переносу на другую площадку, инструкции по развёртыванию для эксплуатирующей организации, поддержка пользователей.

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-инженер, Группа компаний «Русагро»

дек 2023 – окт 2025

Москва

Развивал практики MLOps и DevOps в подразделении машинного обучения и ИИ: поддерживал 3 команды и инфраструктуру полного жизненного цикла моделей.

- Перевёл ML-приложения на Docker, что стало основой для перехода на Kubernetes.
- Поднял два dev-кластера Kubernetes (1 и 6 нод) для сервисов компьютерного зрения и классического ML.
- Создал шаблоны GitLab CI (сборка, тесты, сканирование уязвимостей, деплой): запуск пилотов сократился с месяцев до нескольких недель.
- Внедрил SAST и сканирование контейнеров на Trivy.

- Развернул ключевую MLOps-инфраструктуру: ClearML, MinIO (S3), HashiCorp Vault, Prometheus, Grafana, Loki, Alloy.
- Настраивал инфраструктуру через Ansible вместе с DevOps-инженерами и службой инфраструктуры.
- Разрабатывал и поддерживал внутреннюю библиотеку утилит, стандартизирующую ML-разработку.
- Участвовал в архитектурных комитетах, проводил ревью архитектуры и кода, готовил рекомендации по MLOps-практикам.

Kubernetes, Docker, Helm, GitLab CI, Trivy, ClearML, MinIO, Vault, Prometheus, Grafana, Loki, Ansible, PostgreSQL, MongoDB, Python, Bash

Computer Vision Engineer (Senior), Группа компаний «Русарго»

апр 2023 – дек 2023

Москва

Автоматизация разработки и архитектура CV-сервисов.

- Разрабатывал и поддерживал внутреннюю библиотеку для компьютерного зрения.
- Принимал архитектурные решения для CV-сервисов и участвовал в планировании новых проектов.
- Развернул трекинг экспериментов.
- Спроектировал систему хранения и учёта размеченных и неразмеченных данных (DWH).

Python, PyTorch, OpenCV, ClearML, Docker

Computer Vision Engineer (Middle), Группа компаний «Русарго»

май 2022 – апр 2023

Москва

Проекты мясного направления — подсчёт животных и поиск заболевших и павших животных.

- Вместе с командой довели точность подсчёта свиней в загонах до 99+%.
- Разрабатывал алгоритмы сбора данных и обучал модели.

Python, PyTorch, OpenCV, TensorRT

Ведущий ML-инженер, Форпост

июл 2021 – май 2022

Пермь

Видеоаналитика, вывод моделей в продакшен и руководство группой ML-инженеров.

- Развернул и администрировал облачную инфраструктуру для обучения нейросетей.
- Писал кастомные слои и реализовывал алгоритмы из свежих статей по компьютерному зрению.
- Стандартизировал ML-пайплайны и выводил модели в продакшен.
- Нанимал, обучал и руководил группой ML-инженеров, ставил и контролировал задачи, вёл технические переговоры с заказчиками.

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

ML-инженер, Форпост

авг 2020 – июл 2021

Пермь

Модели детекции, сегментации и классификации для видеоаналитики.

- Собирал и размечал данные, дообучал и улучшал модели, готовил их к продакшену.
- Писал утилиты и скрипты для сбора и обработки данных.

Python, TensorFlow, Darknet, OpenCV

Аналитик Дом.ру Movix, ЭР-Телеком Холдинг

мар 2020 – авг 2020

Пермь

Продуктовая аналитика ТВ-сервиса Movix.

- Проверял гипотезы, считал продуктовые метрики, сегментировал клиентскую базу.
- Анализировал поисковые запросы и интернет-трафик абонентов.
- Спроектировал и наполнил базу данных отдела аналитики, автоматизировал отчётность.

Python, SQL, pandas, Statistics

Преподавание

Преподаватель курса MLOps, Пермский государственный национальный исследовательский университет

сен 2024 – июн 2025

Читал магистрантам курс по MLOps и интеграции машинного обучения в интернет-приложения.

- Разработал теорию и практику по автоматизации жизненного цикла моделей: данные, обучение, трекинг экспериментов, деплой.
- Курировал итоговый проект: от сбора данных до модели за API с метриками производительности и качества.
- Учил студентов инструментам автоматизации и мониторинга ML-проектов, включая трекинг экспериментов и CI/CD.

Проекты

Этот сайт

окт 2026 – н. в.

Интерактивное резюме с ИИ-агентом по схеме «System 1 / System 2»: быстрая типизированная модель решений выбирает маршрут и отсекает запросы не по теме, Claude через LiteLLM рассуждает и отвечает с помощью инструментов и карточек.

- Быстрая модель решений как System 1 с цепочкой резервных звеньев.
- Модель для ответа подбирается под маршрут.
- LiteLLM-прокси, лимиты и бюджеты, трейсинг в Langfuse.
- RU/EN, админка с AI-импортом профиля, CI/CD в GitLab на своём раннере.

cv.hl59.ru

Сервис инференса LLM

ноя 2025 – н. в.

Инференс LLM с OpenAI-совместимым API для сотрудников ГНИВЦ и в контурах ФНС России.

- vLLM, SGLang и NVIDIA Dynamo; модели Qwen 3.x, DeepSeek V4 и другие; переезды между движками без потери доступности.
- Медианный TTFT ключевой модели улучшился примерно в 3 раза (~1,1 → 0,4 с).
- FP8 KV-кэш почти удвоил ёмкость кэша.

vLLM, SGLang, NVIDIA Dynamo, Kubernetes

Model Router

ноя 2025 – н. в.

Собственный HTTP-роутер для vLLM/SGLang.

- Балансировка, отказоустойчивость, повторные попытки и метрики.

vLLM, SGLang

Kubernetes-кластер ИИ-Песочницы

ноя 2025 – н. в.

GPU-кластер внутренней ИИ-песочницы ГНИВЦ для более чем 1000 пользователей.

- GPU-узлы NVIDIA H100, H200, RTX 6000 Ada с GPU-оператором, KAI Scheduler и Grove.
- Longhorn, Calico, MetalLB, Envoy Gateway; узлы настраиваются через Ansible.
- Сервисы без простоя: Open WebUI, GPUStack, JupyterHub, Qdrant, ClearML, Langfuse; обновление Langfuse 3.x → 4.x с переносом ~1,9 млн событий.

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

GitOps и безопасность песочницы

ноя 2025 – н. в.

CI/CD и GitOps-контур с управлением секретами, политиками и проверкой образов.

- Argo CD с SSO через Keycloak, сборка на Argo Workflows и kaniko в Artifactory, события из Bitbucket через Argo Events.
- Vault и External Secrets Operator, политики Kyverno.
- Образы доведены до 0 исправимых CRITICAL/HIGH по Trivy.

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

ИИ-Платформа ФНС России

ноя 2025 – н. в.

Проектирование платформы для десятков тысяч инспекторов, сотен тысяч граждан и высоконагруженных автоматизированных систем.

- Концепция и проектная документация.
- Архитектура среды исполнения моделей с Prefill/Decode disaggregation, подбор и распределение GPU под классы моделей.

NVIDIA Dynamo, Kubernetes

Курс MLOps для магистрантов

сен 2024 – июн 2025

Авторский курс о полном жизненном цикле ML-системы с итоговым проектом.

Python, Docker, FastAPI, CI/CD

Kubernetes-платформа для ML

дек 2023 – окт 2025

Путь от «запусков на ноутбуке» до контейнеров и двух dev-кластеров Kubernetes для ML-сервисов.

- Единая среда выполнения на Docker для всех ML-команд.
- Кластеры на 1 и 6 нод под CV и классический ML.

Kubernetes, Docker, Helm, Ansible

Шаблоны CI/CD и безопасность

дек 2023 – окт 2025

Типовые этапы GitLab CI для ML-проектов со встроенной проверкой уязвимостей.

- Сборка, тесты, SAST и сканирование контейнеров на Trivy, деплой.
- Запуск пилотов сократился с месяцев до нескольких недель.

GitLab CI, Trivy, Docker

MLOps-стек

дек 2023 – окт 2025

Трекинг экспериментов, хранилище артефактов, секреты и наблюдаемость для ML-сервисов.

- ClearML для экспериментов, задач и агентов; MinIO для артефактов; Vault для секретов.
- Prometheus, Grafana, Loki и Alloy для метрик и логов ML-сервисов.

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

Компьютерное зрение на фермах

май 2022 – дек 2023

Подсчёт животных и поиск заболевших и павших животных по видео.

- Точность подсчёта свиней в загонах — 99+%.
- Сбор данных, обучение и оптимизация моделей детекции.
- DWH для размеченных и неразмеченных данных.

PyTorch, OpenCV, TensorRT

Видеоаналитика

авг 2020 – май 2022

Модели детекции, сегментации и классификации для видеоаналитики, от разметки до продакшена.

- Реализация алгоритмов из свежих статей и кастомные слои.
- Облачная инфраструктура для обучения в GCP и AWS.

Навыки

LLM и инференс	vLLM, SGLang, NVIDIA Dynamo, GPUStack, NVIDIA GPU Operator, KAI Scheduler, Grove, LiteLLM, Pydantic AI
MLOps и инфраструктура	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
Наблюдаемость	Prometheus, Grafana, Loki, VictoriaMetrics, Vector, Langfuse
ML и компьютерное зрение	PyTorch, TensorFlow, OpenCV, TensorRT, TF Serving, Darknet, pandas, NumPy
Языки и основы	Python, SQL, Bash, Go, Linux, Git, Statistics
Облака и данные	GCP, AWS, PostgreSQL, MongoDB, Qdrant

Образование

2008	Экспертиза и управление недвижимостью. Пермский государственный технический университет
2021	Нейросетевые технологии в компьютерном зрении. ПГНИУ, Региональный институт непрерывного образования
2020	Профессия Data Scientist. Skillbox
2020	Professional Scrum Developer. Agilix consulting
2019	mlcourse.ai – машинное обучение. Open Data Science

Языки

Русский – родной Английский – C1