

Yehor Brodskiy

Senior Developer

Kyiv, Ukraine | yehor@yeskiy.com | yeskiy.com | github.com/yeskiy | linkedin.com/in/yehor-brodskiy

Summary

DevOps and SRE engineer with more than 5 years building and operating production systems, from bare-metal Kubernetes to self-hosted LLM inference. Recent work: self-hosted, OpenAI-compatible LLM serving on open-weight models - a Rust batch-API gateway in front of vLLM, served on Kubernetes and across multiple GPUs over NVLink, hardened, monitored and handover-ready, including a fully air-gapped deployment for a confidential defense-sector client. I own the whole path, from hardware sizing to the inference engine, the API and the reliability layer, and write the application code around it too, including retrieval-augmented, natural-language-to-SQL assistants over complex retail schemas.

Experience

Senior Developer, The Parker Avery Group

Mar 2022 - Present

Enterprise systems integration for US retail and CPG: the Unified Demand Intelligence platform and Ask Udi.

- Delivered a fully air-gapped, self-hosted deployment for a confidential defense-sector client (name withheld under NDA), with no public inbound exposure.
- Served open-weight LLMs on Kubernetes and across multiple GPUs over NVLink.
- Deployed and load-tested vLLM serving a quantized open-weight model (Gemma, 12B, 4-bit and FP8 variants) on a single 24 GB GPU. Tuned KV-cache dtype, sequence concurrency, GPU memory utilization and the structured-output backend to reach about 335 output tokens per second at 16-way concurrency with zero request preemptions.
- Hardening and observability: bearer-token authentication with constant-time comparison, loopback-only exposure, Prometheus metrics and structured JSON logging.

Senior Technical Consultant, Mobilauto

2022 - 2024

I'd already moved to The Parker Avery Group when Mobilauto came back to life. I owned a lot of its parts, so I pretty much got pulled into a 2-year knowledge transfer.

Full Stack Engineer / DevOps (incl. SRE), Team Lead, Mobilauto

Jan 2020 - Mar 2022

Ran the taxi platform end to end: the Cabinet24 CRM, the Express Soft dispatch platform, Mobilauto Web, the gateway and the partner bridge. prpcow and rust-wasmpack-loader came out of this work as open source. It ended in spring 2022, with the full-scale invasion.

- Designed low-cost infrastructure on Kubernetes (k0s) for high availability of every critical service.
- Built a large-scale k0s environment supporting multiple deployment options, including automated setups in dedicated client VMs.
- Built a program set for taxi services on Kubernetes (k0s) for deployment orchestration. Node.js backend (including worker threads), Kafka for driver-point delivery, Redis for cache, history and pub/sub, PostgreSQL, React frontend, full Asterisk and A2Billing telephony integration.
- Led a team of six.

Full Stack Engineer / DevOps (incl. SRE), Qberry

Apr 2019 - Dec 2019

My first salaried job, at an outsourcing firm. Production reliability and product work.

- Built many customer-requested environments: Docker-based, bare-metal, several cloud platforms, and Kubernetes clusters including air-gapped setups.
- Customer projects (React and Vue frontends, Node.js including worker threads, MariaDB, MySQL and MongoDB), plus support for bare-metal servers including PowerDNS.
- Built gamification add-ons (Node and RabbitMQ on the backend, React on the frontend).
- Led 2 or more developers.

Full Stack Engineer / DevOps (incl. SRE), Tech Lead, Cropman.io

Jan 2019 - Apr 2019

A multi-image editor right in the browser. Built it with classmates from DAN.IT and sold it as a product.

- Used Firebase Cloud Functions as compute for user requests, with everything else on Docker Compose.
- An online batch photo editor (layers, crop, filters, watermarks). Node and Sharp (libvips) with Firebase for auth and storage, MongoDB, and a React frontend.
- Led a team of four.

Freelance, Self-taught builds

2015 - 2018

Skills

Orchestration

Kubernetes, K8s on bare metal (Talos, k0s), ArgoCD, Devtron, Custom CSI / Operators, k3s, Docker / Podman, CI/CD automation, Terraform, Ansible

LLM serving

vLLM, Ollama, Quantization (w4a16, FP8), Multi-GPU serving over NVLink, OpenAI-compatible APIs (Chat and Batch), Throughput and latency load-testing

Observability

SigNoz, Datadog, Prometheus, Grafana, OpenReplay, Microsoft Clarity

Cloud & hosting	Self-hosted infrastructure, Bare metal, Major clouds, Private, air-gapped and hybrid environments, EDGE computing
Languages	Python, Go, TypeScript / JavaScript, Rust, Java, C#
Runtimes	Edge runtimes, Bun, Node, WASM in the browser
Frontend	Next.js, React, Tailwind CSS, Angular, Vue, Nuxt.js
Backend & messaging	Node, prpcow, REST, RPC, GraphQL, WebSockets, gRPC, Protobuf, Spring, Microservices
Multimedia	nginx-vod-module, HLS, Streaming pipelines
Auth & identity	Keycloak, Zitadel, OIDC / OAuth2, SSO / IdP federation
CMS	Strapi, Custom headless
Data & AI	pgvector, qdrant, OpenAI APIs, Langfuse, LangFlow, LangGraph, RAG / embeddings, LangChain, ClickHouse, Apache Spark, DuckDB
Queues & gateways	Higress, KubeMQ, Kafka, Valkey / Redis, RabbitMQ
Realtime & telephony	Custom TURN / STUN with strong auth, Asterisk, A2Billing, WebRTC softphones
Databases	PostgreSQL, MySQL / MariaDB, SQL Server, Oracle DB, MongoDB, Cassandra, Couchbase, SQLite
Storage	S3 / MinIO, GCS, Azure Blob, B2, NFS / SMB, Ceph
Security	Network hardening, VPN-only and no-public-inbound topologies, Secrets management, Vulnerability scanning, Compliance controls

Education

Kyiv Aviation Institute (National Aviation University), Bachelor's, IoT	2021 - 2025
Faculty of Electronics, Robotics, Monitoring and IoT Technologies. Average mark 89/100. Qualification work: Network Protocol for Remote Function Calls (RPC-like) as the primary seamless connection between EDGE devices in IoT.	
Kyiv Aviation Institute (National Aviation University), Master's, Electronics, Robotics and IoT	2025 - 2026
Faculty of Air navigation, Electronics and Telecommunications.	
DAN.IT Education, Full stack course	2019 - 2021
Average mark 80/100.	

Languages

Ukrainian Native | **English** C1/C2 | **Russian** Native