

Andrey Krisanov

Moscow, Russia | develop4net@gmail.com | github.com/akrisanov | linkedin.com/in/akrisanov | akrisanov.com

Software Engineer, LLM Inference & AI Infrastructure

PROFILE

Software engineer specializing in production LLM inference, GPU-backed infrastructure, and distributed systems. At Severstal, I lead the architecture and technical evolution of a shared inference platform running on Kubernetes and 24 NVIDIA H200 GPUs for enterprise AI products and coding agents.

My scope includes model serving, traffic management, performance and reliability, observability, capacity planning, model rollout and rollback, and multi-data-center resilience. Previously built and scaled production platforms across SaaS, fintech, data privacy, and high-traffic consumer systems.

SELECTED CAREER HIGHLIGHTS

- Lead the architecture and technical evolution of Severstal’s shared LLM inference platform, currently running on 24 NVIDIA H200 GPUs; designed the target topology and capacity model for a planned expansion to 48 H200 and 8 H100 GPUs across two data centers.
- Modernized a \$3M+ ARR SaaS platform, improving critical backend paths by 2x-10x, achieving 99.998% availability, and cutting mean time to recovery by 50%.
- Designed and launched X5 Media, a web platform that reached more than 20 million monthly active users within its first year.

WORK EXPERIENCE

Staff Software Engineer
Severstal | DaVinci GenAI Platform

Jan 2026 — Present
Moscow, Russia

- Lead architecture and technical direction for DaVinci, a shared GenAI platform serving enterprise AI products, assistants, and coding agents.
- Own production inference architecture for self-hosted DeepSeek, Kimi, Qwen, and other open-weight models on Kubernetes, vLLM, and 24 NVIDIA H200 GPUs, including model placement, parallelism, batching, KV-cache allocation, and request limits.
- Designed the target topology, capacity allocation, degradation policy, failover, and recovery strategy for a planned expansion to 48 H200 and 8 H100 GPUs across two data centers.
- Benchmark and tune model-serving configurations against production traffic, balancing time to first token, token throughput, concurrency, context length, and GPU memory.
- Design and develop the AI gateway and inference control plane for routing, aliases, quotas, rate limits, fallbacks, evaluation, and controlled rollout and rollback across self-hosted and external models.
- Build SLO-driven operations with Prometheus, Grafana, alerting, load testing, incident analysis, and GPU capacity forecasting.

Chief Technology Officer
PlanFact | B2B Accounting SaaS

Jan 2024 — Dec 2025
Moscow, Russia

- Served as the hands-on technical lead for modernization of a \$3M+ ARR SaaS product in a 100-person company while leading a 25-person IT organization.
- Modernized the core .NET platform and introduced a domain-oriented modular architecture, replacing legacy dependencies and overloaded layered patterns without disrupting business-critical services.
- Improved critical backend paths by 2x-10x and scaled read-heavy workloads through targeted refactoring, SQL and database tuning, strongly consistent replication, and horizontal read scaling.
- Introduced service-level observability and incident-management practices, achieving 99.998% availability and cutting mean time to recovery by 50%.
- Raised engineering standards through RFCs, architecture and code reviews, Terraform, Ansible, mentoring, and hiring.

Senior Software Engineer, Backend & Platform

Stattice.ai | Data Privacy Platform

Jun 2021 — Dec 2023

Berlin, Germany

- Built Python, FastAPI, and Celery backend services and Kubernetes (GCP) platform workflows for a cloud PII-protection product and privacy-preserving machine learning workloads.
- Modernized SDK architecture and cross-platform x64/ARM build pipelines; introduced type checking, static analysis, CI improvements, and security scanning while maintaining 99.99% availability.

Engineering Lead, Backend & Platform

X5 | Retail Technology

Jun 2019 — Jun 2021

Moscow, Russia

- Designed and launched X5 Media, a Python, FastAPI, and Next.js platform that reached more than 20 million monthly active users within its first year.
- Owned backend and platform architecture during the transition from outsourced to in-house engineering; built the first cross-functional teams and halved time-to-market for existing products.

Earlier Engineering Experience

Fintech, SaaS, and Startups

2012 — 2019

Germany, Russia, and Remote

- **Independent consultant:** Migrated a Frankfurt fintech from on-premises infrastructure to AWS, cutting infrastructure costs by 50%; designed cloud architectures for two early-stage startups.
- **Clark:** Built a secure server-to-server integration with N26 that enabled in-app insurance signing and created a new partnership revenue stream; modernized and hardened a Ruby on Rails platform to improve security, performance, maintainability, and support for white-label products.
- **Ridewithlocal:** Owned backend engineering, AWS infrastructure, and production reliability for the initial product launch, maintaining 99.99% uptime; delivered the MVP on schedule and helped the company secure a \$300K pre-seed round.
- **Payler:** Designed and launched a PCI DSS-certified payment gateway, helping transform the company from a services business into a product-focused fintech platform; scaled engineering from one developer to a team of approximately 30.

SKILLS

- **LLM Inference:** vLLM, NVIDIA A100/H100/H200, distributed model serving, continuous batching, tensor/data parallelism, KV-cache management, performance benchmarking, latency and throughput optimization
- **Distributed Systems & Reliability:** Kubernetes, AI gateways, model routing, rate limiting, fallbacks, multi-data-center resilience, SLOs, capacity planning, load testing, Prometheus, Grafana, DCGM
- **Programming & Data:** Python, FastAPI, C# (.NET), TypeScript, PostgreSQL, Redis, ClickHouse
- **Infrastructure:** Linux, Docker, Ansible, Terraform, AWS, GCP

EDUCATION

Bryansk State University

Master's in Applied Mathematics & Information Science

Bryansk, Russia

2006 — 2011