David **Mikheev**

davidmiheev | davidmiheev | me@davidmiheev.life

Experience

dodil.io
Moscow

Founding Engineer
Dec 2025 – Present

- Designed and built *dodil-agent*, a Rust agent platform that turns a plain-language requirement into a deployed app: durable task orchestrator on NATS JetStream and S3, agent sandboxing migrated to NVIDIA OpenShell (Docker and Kubernetes compute drivers), per-run token budgets enforced at a metering proxy
- Co-built Ignite, a serverless platform in Rust on which LLM executors and agent sandboxes run; wrote the candle-based LLM executor (CUDA, quantization, HTTP/SSE serving); contributed to the team's ONNX Runtime executors (embeddings, text encoders, seq2seq)
- Optimized LLM inference in a candle fork: CUDA graphs, fused kernels, quantized Qwen3.6 / Gemma 4; 35B MoE decode 2.5× faster (19.6 → 48.9 tok/s)

Freelance
Moscow

Quantitative Researcher & Developer
Nov 2024 – Dec 2025

- Applied ML to signal discovery on recorded order books: built a markout-based spread rule that beat baseline strategies on held-out data across all strategy–asset pairs
- Implemented and backtested GLFT market making with hftbacktest on recorded Binance order books: calibrated fill intensity and volatility, modeled order queue, latency and inventory risk
- Built cross-exchange arbitrage scanners in Rust: real-time ingestion and processing of L2/L3 market data streams

Yandex
Moscow

Analyst Developer (Intern)
Jul 2024 – Oct 2024

- Built a dashboard analyzing dispatch timings and conversion to assignments across route types (Python, SciPy, YQL, DataLens)
- Ran several switchback experiments (an A/B-test analogue); the results changed assignment settings
- Proposed improvements to switchback and A/B methodology and results analysis (Nirvana, YT); found a bug in one assignment type

HSE Faculty of Computer Science; freelance tutoring
Moscow, Tbilisi

Teaching
2022 – 2024

- Adjunct instructor at HSE (Feb–Jun 2024): taught seminars on topological data analysis with a Python practical component
- Math and DSA tutor (Dec 2022–Jan 2024): taught mathematics, algorithms and data structures, and Python programming

Algebraic Topology and Applications Laboratory, HSE
Moscow

Research Internship
Sep 2021 – May 2022

- Developed GalaxyWitness, a Python package for topological analysis of astronomical data (on PyPI)
- Later supervised students who contributed to the project

Education

Lomonosov Moscow State University, Faculty of Mechanics and Mathematics
Moscow

Specialist Degree (MSc equivalent) in Mathematics, GPA: 4.9/5.0
Sep 2016 – Jun 2022

- Advisors: Dr. Victor Buchstaber, Dr. Anton Ayzenberg
- Thesis: Invariants of matroids defined by torus actions

Projects

Selected projects on GitHub (all highlighted elements are links)

CANDLE *2026*

- Performance fork of HuggingFace's Rust ML framework huggingface/candle: CUDA graph capture, fused kernels, static decode, load-time quantization; quantized weights published on Hugging Face
- Implemented Unlimited OCR (text decoder, vision encoder, crop mode); optimized or extended Qwen3.5/3.6, Gemma 4, DeepSeek-V2-Lite (fits a 16 GB GPU), PaddleOCR-VL, Whisper (graph decoding verified byte-identical to eager) and Flux

DS-AGENT *2026*

- Personal data-science and quant-research agent in the browser (FastAPI, WebSocket) on the Claude Agent SDK, with Kaggle, Google Colab, data-science, research and cross-session-memory MCP servers
- BYOK key vault (Fernet), Telegram bot with one-time-code login, one-command VPS deploy with GitHub Actions CD and end-to-end tests

RESEARCH-AGENT *2026*

- Personal research agent (*Atlas*): ReAct loop with arXiv search and paper ingestion, long-term semantic memory in a vector store (Dodil K3), React UI; deployable as a container on Dodil Ignite

GRACIO *2026*

- Single-threaded async runtime in Rust from scratch (executor, wakers, timers, bounded channels with backpressure), std only

Skills

Programming Languages	Rust (tokio, rayon, axum, tonic, sqlx, async-nats), Python, SQL, Go, CUDA
LLM inference and agents	vLLM, candle, CUDA graphs, GGML quantization, ONNX Runtime, MCP, tool-calling agents, Claude Agent SDK, Claude Code, NVIDIA OpenShell
Data science and ML	numpy, scipy, scikit-learn, jax, torch, huggingface (transformers, trl)
Backend and DevOps	NATS JetStream, S3, PostgreSQL, gRPC, SSE / WebSocket, Docker, Kubernetes, GitHub Actions, Linux
Languages	English, Russian