

VLADISLAV ELDASHOV

✉ eldashovvv@yandex.ru |  linkedin.com/in/vladislav-eldashov |  t.me/velliev |  github.com/avolxn

SUMMARY

Senior Machine Learning Engineer with 4+ years of experience building production ML and LLM systems. Combines hands-on engineering with ownership of technical strategy, architecture, and end-to-end delivery, leading cross-functional teams from problem definition and evaluation through deployment and operation.

EXPERIENCE

Lead ML Engineer, Sandbox

Apr 2026 – present

- Owned AI strategy and roadmap for a medical information system serving **350+ clinics**, leading a cross-functional team from requirements through launch.
- Automated clinical workflows spanning transcription, visit documentation, medical record audits, patient summaries, discharge reports, pre-visit check-ins, follow-ups, and call analysis, with **6.3M visits** and **900K calls** processed.
- Built a multimodal AI agent combining RAG over a **900K-page** knowledge base with web search, sandboxed code execution, MCP integrations with dynamic tool discovery, skills, long-term memory, and automatic context compaction.
- Architected multi-provider infrastructure with circuit breakers, automated failover, persistent job queues, background workers, automatic recovery, tracing, and zero-downtime deployments, achieving **99.8% availability**.

ML Engineer, Mars

Sept 2023 – Apr 2026

- Built an enterprise Text-to-SQL platform with schema-aware retrieval, multi-step query planning, SQL validation, and role-based access control, automating **68% of analytical requests**.
- Architected an on-premises LLM inference platform serving 10+ models, reducing **p95 latency by 41%** through dynamic batching, request routing, KV-cache optimization, and workload-aware resource allocation.
- Developed an LLM-powered content generation pipeline with automated quality evaluation, accelerating content production **8x** and increasing marketplace CTR.
- Built a contract intelligence system combining entity extraction, cross-document relation linking, and discrepancy detection, achieving **92% field-level extraction precision** on held-out B2B contracts.
- Developed a review intelligence system combining aspect-based sentiment analysis, weak supervision, and topic discovery to identify emerging product defects and recurring customer issues.

Data Scientist, Kontur

Mar 2022 – Jan 2023

- Developed a B2B churn prediction system using CRM/ERP data and behavioral feature engineering, improving **PR-AUC from 0.71 to 0.81** on held-out data.
- Built a demand forecasting service with automated feature pipelines and backtesting, reducing **forecasting error by 30%**.

PROJECTS

TinyGPT

- Pretrained a 0.7B-parameter Transformer from scratch on 5.8B tokens using **8x A100 GPUs**, achieving **0.2626 CORE**.
- Distilled a 1.9B-parameter teacher into the student (offline soft-label), achieving **0.3168 ChatCORE**.
- Built an Airflow-orchestrated distributed training stack with PyTorch FSDP, vLLM inference, sequence packing, streaming datasets, resumable checkpoints, and MLflow tracking.

SKILLS

NLP: PyTorch, Hugging Face Transformers, FSDP, vLLM, TensorRT-LLM, Triton, ONNX Runtime, PEFT, LangGraph, LangChain, DSPy, MCP, Qdrant, FAISS, Elasticsearch, OpenSearch, Langfuse; LLMs, RAG, AI agents, ASR, distributed training, pretraining, fine-tuning (SFT, LoRA), distillation, quantization, inference optimization, LLM evaluation

ML: Scikit-learn, CatBoost, XGBoost, LightGBM, Optuna, SHAP, SciPy, NumPy, Pandas, PySpark, Polars; classification, regression, ranking, retrieval, metric learning, clustering, time series forecasting, A/B testing

Engineering: Python, SQL, FastAPI, PostgreSQL, Redis, Kafka, ClickHouse, S3, Airflow, Docker, Kubernetes, MLflow, Weights & Biases, Prometheus, Grafana, Git, CI/CD, Linux; MLOps, ML System Design

EDUCATION

NUST MISIS, B.Sc. in Computer Science, Intelligent Data Analysis

ADDITIONAL EDUCATION

Yandex School of Data Analysis (YSDA): Efficient Deep Learning, NLP, Deep Learning, Reinforcement Learning, A/B Testing

COMPETITIONS

- 2nd place**, PurpleHack (2026): Factual hallucination detector for GigaChat3-10B.
- 2nd place**, TulaHack (2025): Navigator for people with disabilities considering their limitations and infrastructure.