khaykingleb@gmail.com
https://github.com/khaykingleb
https://linkedin.com/in/khaykingleb

Gleb Khaykin

+995 585-889-511
https://khaykingleb.com
https://t.me/khaykingleb_blog

Summary

MLOps/DevOps Engineer with 4+ years of experience in software engineering and infrastructure. Specialize in building high-performance systems across cloud and on-premise environments. Built and deployed scalable production pipelines supporting 30k+ daily active users and real-time AI workloads. Strong background in speech, natural language processing, and finance.

Experience

Together AI — Amsterdam, Netherlands
MLOps Engineer — *July 2025 — Present*

Stealth Startup — Remote
MLOps/DevOps Engineer — *July 2023 — July 2025*

- Architected Tailscale-meshed K3s GPU clusters with the NVIDIA GPU Operator, achieving >50% reduction in GPU hosting costs
- Established IaC monitoring and alerting for >4 Kubernetes clusters, decreasing incident resolution time from 1 day to 1 hour
- Spearheaded the integration of an ensemble model for ASR and speaker diarization, achieving 11x speedup with improved accuracy
- Built a GPU-enabled inference pipeline with theNVIDIA Triton Inference Server, reduced per-hour meeting processing costs by 5x
- Delivered multi-stage financial parser with vision LLMs, automated 70% of manual analysis and secured early enterprise client

Myna Labs — Remote
ML/MLOps Engineer — *November 2022 — March 2024*

- Developed a priority-queuing REST/gRPC API serving 60k+ inference requests per hour for multi-modal speech tasks
- Researched and deployed state-of-the-art multilingual, emotional text-to-speech and voice conversion models in 44 kilohertz
- Engineered a speech model fine-tuning service, reducing voice cloning time to under 1 minute
- Terraformed and migrated speech infrastructure to Kubernetes on GKE with >4000 lines of IaC
- Mentored a team of engineers on MLOps/DevOps best practices, improving code quality across projects

Stealth Startup — Remote
Data Scientist — *September 2021 — July 2022*

- Designed a PySpark pipeline to scrape and process S&P 500 news to estimate investor sentiment with BERT-like models
- Developed a list-wise learning-to-rank system for stock selection, optimizing portfolio performance in long-short strategies
- Engineered cron-based ETL workflows to aggregate data from 150+ financial API endpoints

Projects

- **Research Playground**: ML/DL models collection in PyTorch Lightning with distributed training support on K3s cluster
- **Huffman Archiver**: C++ implementation of Huffman coding for lossless data compression with Conan package management
- **Stochastic Optimization**: zero-gradient methods (Particle Swarm, Genetic Algorithm) for optimization problems
- **Personal Website**: portfolio built with Next.js, Supabase, and Notion API
- **Dotfiles**: declarative system configuration with Nix for reproducible environments on macOS and NixOS

Education

NRU "Higher School of Economics" — Moscow, Russia
B.S., Computer Science and Finance, summa cum laude — *September 2018 — July 2022*

CFA Institute — Moscow, Russia
Level 1 passed, 90th percentile — *February 2021*

Skills

- **Languages**: Python, Rust, Go, C++, SQL, TypeScript
- **DevOps**: Terraform, Ansible, Nix, Docker, Kubernetes, ArgoCD, CI/CD, AWS, GCP, Grafana, Prometheus, Loki, Tailscale
- **MLOps**: DVC, NVIDIA Triton, Ollama, LangChain, LangGraph, ONNX, TensorRT, CoreML
- **R&D**: PyTorch, Lightning, Hydra, W&B, Gradio, Scikit-learn, Numpy, Pandas, SciPy
- **DE**: Hadoop, Spark, Polars, Kafka, RabbitMQ, Celery
- **Storage**: Redis, PostgreSQL, PGVector, S3, R2, Supabase, MongoDB
- **Backend**: gRPC, REST, Websocket, Webhook, PubSub, AsyncIO, FastAPI, Pydantic
- **Frontend**: React, Next.js, Remix, Tailwind, Vercel