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

Gleb Khaykin

Amsterdam, The Netherlands

+31 06 26172415
https://khaykingleb.com
https://t.me/khaykingleb_blog

Summary

ML Infrastructure Engineer with 4+ years of experience architecting high-performance ML systems, multi-cluster platforms, and large-scale inference/training pipelines. Specialized in distributed systems, Kubernetes, Terraform, observability, and CI/CD. Delivered 3x CI/CD improvements, 50%+ GPU cost reductions, and inference systems serving 60k+ req/hour.

Experience

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

- Accelerated CI/CD pipelines 3x through BuildKit mount caching, Docker optimization, and shared-cache self-hosted runners architecture

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

- Architected K3s GPU clusters with NVIDIA GPU Operator, reducing hosting costs by >50% via hybrid cloud optimization
- Established Terraform-managed observability platform for K8s infrastructure with Grafana Cloud, decreasing MTTR from 1 day to 1 hour
- Built inference system with NVIDIA Triton, Celery, and RabbitMQ for ASR/LLM workload orchestration
- Optimized ensemble ASR/diarization pipeline achieving 11x speedup, reducing processing costs by 5x
- Delivered multi-stage financial document parser using vision LLMs, achieving 70% automation and securing venture fund as first client

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

- Developed multi-model FastAPI platform for speech tasks (emotional TTS, voice conversion, ASR), serving 60k+ req/hour
- Built voice cloning service using LoRA fine-tuning, reducing voice addition time from days to under 10 minutes
- Modernized and automated speech infrastructure, migrating from Docker Compose to Terraform-managed GKE with autoscaling

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

- Designed distributed PySpark pipeline for financial sentiment analysis, scraping and processing S&P 500 news data with BERT models
- Developed learning-to-rank system for quantitative stock selection in long-short trading strategies
- Engineered automated ETL workflows on Databricks to aggregate and analyze data from 150+ financial API endpoints

Projects

- **Research Playground**: Production-ready PyTorch Lightning framework with Hydra configs and distributed training on self-hosted K3s
- **Dotfiles**: Declarative system configuration with Nix flakes and home-manager for reproducible macOS environments
- **Huffman Archiver**: C++ implementation of Huffman coding for lossless data compression with Conan package management
- **Personal Website**: Full-stack blog and portfolio with Next.js frontend, Supabase backend, and live content sync via Notion API

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 — *February 2021*

Skills

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