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/blog_khaykingleb

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 ML platform automation. Delivered 50%+ GPU cost reductions, 4x CI/CD improvements, and inference systems serving 60k+ req/hour.

Experience

Together AI — Amsterdam, Netherlands
ML Platform Engineer — *July 2025 — Present*

- Migrated model development workloads off a saturated shared EKS to a dedicated one with Karpenter, eliminating 1h manual node scaling
- Engineered model sync pipeline and integration tests for 90+ models, unlocking dedicated and serverless inference for fine-tuned models
- Unified observability across 6+ K8s clusters into a single Grafana Cloud stack and shipped 30+ IaC-defined dashboards
- Accelerated CI/CD pipelines 4x through BuildKit mount caching, Docker optimization, and self-hosted runners with shared cache

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

- Provisioned and operated hybrid-cloud K3s GPU clusters with Tailscale-meshed networking, reducing GPU costs by >50%
- Standardized K8s monitoring and alerting across 4 clusters in Grafana Cloud, reducing issue detection time from ad hoc to under 1 hour
- Optimized ASR/diarization with NVIDIA Triton BLS and model selection, achieving 11x speedup and cutting processing costs by 5x
- Delivered a multi-stage financial document parser using vision LLMs, reaching 70% automation and securing a venture fund as first client

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

- Owned a priority-queued speech inference API (TTS/VC/ASR) on FastAPI and NVIDIA Triton, serving 60k+ req/hour
- Built a voice cloning service using LoRA fine-tuning, reducing voice addition time from days to under 10 minutes
- Modernized and automated speech-related infrastructure by 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**: Go, Python, Rust, C++, SQL, TypeScript
- **DevOps**: Kubernetes, Terraform, AWS, GCP, Docker, Helm, ArgoCD, Ansible, Nix, Grafana, Prometheus, Loki
- **MLOps**: NVIDIA Triton, TensorRT, ONNX, Volcano, Flyte, DVC
- **R&D**: PyTorch, Lightning, Hydra, W&B, NumPy, Pandas, SciPy
- **Data Engineering**: Spark, Polars, Kafka, RabbitMQ, Celery
- **Storage**: PostgreSQL, Redis, MongoDB, S3/R2, PGVector
- **Frontend**: React, Next.js, Tailwind