khaykingleb@gmail.com
khaykingleb
khaykingleb

Gleb Khaykin

+995 585-889-511
khaykingleb.com
@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. Reduced infrastructure costs by over 30% and deployed production pipelines that support real-time AI and data-intensive workloads for 30k+ daily active users. Strong background in speech, natural language processing, and finance.

Experience

MLOps/DevOps Engineer at iClerk — San Francisco, United States (remote)
Silicon Valley startup automating tedious tasks for B2B, $3.4M pre-seed investment — *July 2023 — Present*

- Architected a K3s GPU cluster with NVIDIA GPU Operator, achieving >50% reduction in GPU hosting costs
- Established IaC-driven monitoring and alerting across 4 K8s clusters, decreasing downtime and incident resolution time
- Spearheaded integration of an ensemble model for ASR and speaker diarization, achieving 11x speedup with improved accuracy
- Built a GPU-accelerated inference pipeline with NVIDIA Triton, reducing per-hour meeting processing costs by 5x
- Delivered a RAG pipeline for financial statements parsing, automated 70% of manual analysis and secured early enterprise client

ML/MLOps Engineer at MynaLabs — Tbilisi, Georgia
GenAI FunTech startup, founded by a sole investor with 2 prior $150M+ exits — *November 2022 — March 2024*

- Developed a priority-queuing API (REST/gRPC) serving 60k+ inference requests/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 speech model fine-tuning service, reducing voice cloning time to under 1 minute
- Terraformized and migrated speech infrastructure to Kubernetes with >4000 lines of IaC
- Mentored a team of engineers on MLOps/DevOps best practices, improving code quality across projects

Data Scientist at DiviAI — San Francisco, United States (remote)
Silicon Valley startup building a financial data aggregator for the US market — *September 2021 — July 2022*

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

Projects

- **Research Playground** — collection of ML/DL model implementations with distributed training 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 Remix (SSR/CSR), Supabase, and Notion API
- **Dotfiles** — Nix-based declarative system configuration for reproducible environments on macOS and NixOS

Education

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

CFA Institute — Moscow, Russia
Level 1 Passed — *February 2021*

Skills

- **Languages**: Python, Rust, Go, C++, SQL, TypeScript
- **DevOps**: Terraform, Ansible, Nix, Docker, Kubernetes, AWS, GCP, Grafana, Prometheus, Loki, Tailscale, CI/CD, ArgoCD
- **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
- **Backend**: gRPC, REST, Websocket, Webhook, PubSub, AsyncIO, FastAPI, Pydantic
- **Frontend**: Remix, React, Tailwind, Vercel