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 boosting model performance, reducing costs by >$10k monthly, deploying scalable solutions for apps with >30k DAU, and optimizing workflow pipelines. Integrate cutting-edge models and build robust infrastructure across cloud and on-prem environments. Background in speech, natural language processing, and finance.

Experience

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

- Refined an ensemble model for speech-to-text and speaker diarization, achieving an 11x speedup and improved accuracy
- Developed a scalable inference pipeline with NVIDIA Triton, reducing costs by 5x for processing one-hour meetings
- Secured an early client with an advanced RAG to parse financial statements, decreasing manual workload by 70%
- Architected a K3s GPU cluster with NVIDIA GPU Operator for monitoring and validation, cutting GPU costs by >50%
- Established IaC monitoring and alerting for 4 K8s clusters with Grafana Alloy OTel Collector, reducing downtime and errors

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

- Built a REST/gRPC API for speech tasks with priority queuing, handling 60k+ inference requests per hour
- Researched and deployed state-of-the-art multilingual, emotional text-to-speech and voice conversion models in 44 kHz
- Created a service for fast speech model fine-tuning to any voice in under 1 minute
- Terraformized and migrated speech infrastructure to Kubernetes with >4000 lines of IaC
- Introduced MLOps/DevOps best practices to a 20+ developer engineering division

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

- Engineered a PySpark pipeline for scraping S&P 500 news to estimate investor sentiment with BERT-like models
- Developed ETL pipelines that aggregate data from 150+ financial API endpoints
- Designed a listwise learning-to-rank system for stocks recommendation, enhancing long-short strategy performance
- Adopted a Glicko rating system to rank Russell 3000 CEOs

Education

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

- Relevant coursework: C++, Python, R, Machine Learning 1, Machine Learning 2, Large Scale Machine Learning, Deep Learning, Deep Learning in Audio Processing, Reinforcement Learning, Calculus, Linear Algebra, Probability Theory, Mathematical Statistics, Stochastic Processes, Econometrics, Microeconometrics, Differential & Difference Equations

Projects

- **Research Playground**: Collection of ML/DL model implementations with distributed training on K3s clusters
- **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 various 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

Miscellaneous

- **CFA Level 1**: Passed in February 2021, scoring in the top 10% of candidates worldwide
- **Teaching Assistant (NRU HSE)**: Coordinated courses in Accounting, Probability, Statistics, and Machine Learning

Skills

- **Languages**: Python, Rust, Go, C++, SQL, TypeScript, JavaScript
- **DevOps/MLOps**: Terraform, Ansible, Nix, Docker, K8s, K3s, Helm, AWS, GCP, Grafana, Prometheus, Loki, Nginx, Kong, Tailscale, ArgoCD, CI/CD, DVC, NVIDIA Triton, Ollama, LangChain, LangGraph, TorchScript, ONNX, TensorRT, CoreML
- **R&D**: PyTorch, Lightning, Hydra, W&B, Gradio, Scikit-learn, Numpy, Pandas, SciPy
- **Backend**: gRPC, REST, Websocket, Webhook, PubSub, AsyncIO, FastAPI, Pydantic, JWT, Node.js
- **Frontend**: Remix, React, Tailwind, DaisyUI, Vite, Vercel
- **DE**: Hadoop, Spark, Databricks, Polars, Kafka, RabbitMQ, Celery
- **Storage**: Redis, PostgreSQL, PGVector, Sqitch, S3, R2, Supabase, Airtable