khaykingleb@gmail.com
/khaykingleb
/khaykingleb

GLEB KHAYKIN
📍 Tbilisi, Georgia

+995 585-889-511 📞
khaykingleb.com 🌐
@khaykingleb_blog ✈

SUMMARY

MLOps/DevOps Engineer with 3.5+ years of software engineering experience. Specialized in boosting model performance, reducing costs by >$10k monthly, deploying scalable solutions to apps with >30k DAU, and optimizing data and workflow pipelines. Integrated state-of-the-art technologies **across cloud** and **on-premise** environments. Background in **distributed systems**, **speech processing**, **speech synthesis**, **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
July 2023 — Present

○ Refined the ensemble model for transcription and speaker diarization, achieving an **11x speed increase** & **improved accuracy**
○ Innovated a scalable cost-efficient inference pipeline, **reducing** operational **costs by 5x** for one-hour meeting processing
○ Architected a K3s-based GPU cluster that **cut GPU** hosting **costs by >50%**
○ Established **IaC monitoring & alerting** for **4 K8s clusters** with Grafana Alloy OTel Collector, reducing downtime and errors

ML/MLOps Engineer at NeiroAI
Tbilisi, Georgia

Generative AI startup with $150M and $200M B2C app exits
November 2022 — March 2024

○ Built a REST/gRPC API for speech, handling **60k+ inference** and **fine-tune requests** per hour with **priority queuing**
○ 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** critical speech infrastructure 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 **recommendation system** for stocks, optimizing long-short strategies
○ Adopted a **Glicko rating** system to rank Russell 3000 CEOs

PROJECTS

Deep Learning for Audio
○ Implemented **QuartzNet**, **FastSpeech**, and **HiFi-GAN** from scratch based on arXiv papers
○ Applied **compression techniques** to keyword spotting models, achieving **11x size reduction** and **9x speed increase**

Canonical Huffman Archiver
○ Developed an **ASCII archiver** in **C++** with **Conan-based** auto-build system

Stochastic Optimization Methods
○ Constructed **Particle Swarm Optimization** for performance testing on Rosenbrock and Ackley benchmark functions
○ Applied **Genetic Algorithm** for solving the NP-hard Traveling Salesman Problem

EDUCATION

NRU "Higher School of Economics"
Moscow, Russia

B.S. with summa cum laude in Economics and Data Science
September 2018 — July 2022

○ Top 3% among 450 students
○ 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, IFRS, Corporate Finance

MISCELLANEOUS

CFA Level 1
○ Passed in February 2021, scoring in the **top 10%** of candidates **worldwide**

Teaching Assistant at NRU "Higher School of Economics"
○ Facilitated course coordination in **Probability**, **Statistics**, and **Machine Learning** from 2020 to 2022

SKILLS

○ **Languages**: Python, Rust, Go, C++, SQL, TypeScript, JavaScript
○ **Frontend**: Remix, React, Tailwind, Vercel
○ **Backend**: gRPC, REST, AsyncIO, FastAPI, JWT, Node.js
○ **Builds**: Make, CMake, Conan, Poetry, Pnpm, Npm
○ **DevOps**: Terraform, Ansible, Docker, Kubernetes, K3s, AWS, GCP, Grafana, Prometheus, Loki, Nginx, Acme, ArgoCD, CI/CD
○ **MLOps**: DVC, NVIDIA Triton, Ollama, LangChain, TorchScript, ONNX, TensorRT, CoreML, Gradio
○ **DE**: Hadoop, Yarn, Spark, Databricks, Polars, Kafka, Celery, RabbitMQ
○ **Storage**: Redis, PostgreSQL, PGVector, Sqitch, S3, R2, Supabase, Airtable
○ **R&D**: PyTorch, Lightning, Hydra, W&B, Scikit-learn, Numpy, Pandas, SciPy