Michael Diskin

michael.s.diskin@gmail.com Europe yhn112 yhn112

Experience

Wildberries, Head of LLM R&D — Moscow, Russia
2024 – present
2 years

- Built LLM R&D org from 0 to 30+ engineers (4–5 teams); manage a 500+ GPU cluster and define platform strategy for LLM and embeddings company-wide.
- Shipped universal text embeddings (retrieval, ranking, classification) and marketplace-scale MT translating tens of millions of product listings into 10+ languages, including low-resource.
- Deployed optimized LLM serving with use-case routing, batching, and quantization — cut GPU costs by 40%+ while meeting latency SLAs under peak traffic.
- Built RAG-powered assistants (internal and seller-facing) with safety guardrails, evaluation framework, and gold-standard regression datasets.
- Established research-to-production operating model: experiment standards, evaluation gates, reliable releases — reduced iteration cycles from weeks to days.

Brask AI, Senior Research Engineer — Tbilisi, Georgia
2022 – 2023
1 year

- Led R&D on a lip-sync model for AI video dubbing: research, model iteration, production integration.
- Improved robustness across diverse speakers and conditions; defined quality metrics and failure-analysis pipeline with product and engineering.

Yandex Research, Research Scientist — Moscow, Russia
2021 – 2022
1 year

- Research on efficient and distributed training for large models; 5 papers at NeurIPS, ICML, and ICLR.
- Led pre-training of a fully open-source Russian BERT-class language model; co-authored the Hivemind decentralized training library.
- Created an evaluation benchmark for graph neural networks under heterophily (460+ citations, ICLR 2023).

Huawei, Research Engineer — Moscow, Russia
2020 – 2021
1 year

- Computer vision research: optical flow and depth estimation for image-based localization.

Early-stage startups, ML / Software Engineer — Moscow, Russia
2019 – 2020
1 year

- Built ML-powered backend services and data pipelines from scratch in small teams; end-to-end ownership from prototyping to deployment.

Yandex, Software Engineering Intern — Moscow, Russia
2017 – 2018
1 year

- Large-scale analytics over multi-terabyte log data using internal MapReduce infrastructure (YT).

Education

HSE University, Computer Science — Moscow, Russia
2022 – 2024

Yandex School of Data Analysis, Machine Learning (Graduate program) — 2019 – 2021

HSE University, Computer Science — Moscow, Russia
2014 – 2019

Skills

ML & AI

Frameworks

Data & Eval

Infra & MLOps

Languages

Awards

HSE FCS Scholarship for Research Excellence 2024
HSE University

Tinkoff Education Scholarship 2023
Tinkoff

Xeek.ai "Put it on the Map!" — 2nd place (100+ teams) 2020
Cash prize
Xeek.ai

Kaggle "Recursion Cellular Image Classification" — 13th of 800+ teams 2019
Kaggle

Huawei Image Inpainting Hackathon — 2nd place (100+ teams) 2019
Cash prize
Huawei

International Data Analysis Olympiad — 16th of 1000+ participants 2019
IDAO

Languages

Russian
Native speaker

English
Fluent