

Mikhail Sharabarin

Machine Learning Engineer | Data Scientist

Saint Petersburg / Moscow, Russia · Citizenship: Kazakhstan

[Email](#) | [GitHub](#) | [LinkedIn](#) | [LeetCode](#) | [Portfolio](#) | [Resume](#)

SUMMARY

4th-year SPbU student specializing in NLP, LLMs, and Classic ML. Seeking an ML Engineer / Data Scientist internship focused on NLP, LLMs, Classic ML, or RecSys. Strong algorithmic background and hands-on end-to-end ML experience, from data preparation and analysis to ML experimentation, model integration, and inference. Conducts research experiments on model and KV-cache quantization using GPUs.

EDUCATION

Saint Petersburg State University

B.Sc. in Applied Mathematics and Computer Science

- 4th year | Full-time

Saint Petersburg, Russia

Expected 2027

PROJECTS

LLM KV-Cache Quantization

github.com/neuraCollab/kv-cache-quantization-reasoning-study

- Researched FP8 and HQQ INT4/INT2 quantization across 3 reasoning models and 80 AIME-24/MATH-500 tasks.
- Developed First Divergence Point, K-error, Attention-KL, and Logit-KL analyses to investigate reasoning degradation.
- Demonstrated architecture-dependent quality degradation and localized quantization error in outlier channels; investigated preservation of critical channels as a mitigation strategy.

AutoTopic — Topic Modeling

github.com/neuraCollab/ml-portfolio/AutoTopic

- Built a BERTopic pipeline using multilingual SentenceTransformers, UMAP, HDBSCAN, and c-TF-IDF; used pymorphy3 for Russian and spaCy for English.
- Compared BERTopic with LDA and automated hyperparameter optimization with Optuna using coherence and diversity.
- Identified 59 topics from 236K analyzed documents; achieved diversity of 0.966 vs. ~ 0.91 for LDA.

Distributed ML Inference

github.com/neuraCollab/ml-portfolio/cassandra-grpc-ml

- Implemented TF-IDF + One-vs-Rest Logistic Regression for 50-topic classification using AutoTopic data.
- Designed a FastAPI $\rightarrow$ Coordinator $\rightarrow$ gRPC workers architecture with Kubernetes, Cassandra for state, and MinIO for ML artifacts.
- Increased throughput from 87.1 to 346.2 req/s when scaling from 1 to 5 worker replicas, reducing p50 latency from 684 to 165 ms.

ECG Edge AI

github.com/neuraCollab/ml-portfolio/raspberry-pi-ecg

- Built an ECG acquisition system using $2 \times$ AD8232 + $2 \times$ Arduino Nano $\rightarrow$ Raspberry Pi 5; implemented 6-lead reconstruction and signal filtering.
- Integrated an existing PyTorch ECG model and prepared it for CPU-only inference with TorchScript.
- Implemented a FastAPI/WebSocket inference pipeline and signal-quality gate; achieved 5.18 ms per prediction on Raspberry Pi 5.

ACHIEVEMENTS

Gradient of Growth — Top 1%: Time-series forecasting competition; predicted next-month product price using ARIMA.

MTS Challenger — Top 15%: Worked in a 4-person team to develop and present a working Airtable-like product with AI integrations and RAG, with distributed responsibilities across the team.

SKILLS

Machine Learning: Classic ML, Deep Learning, NLP, LLMs, RecSys, Diffusion Models

Frameworks: PyTorch

Languages: Python, C++, SQL

Engineering: Docker, Git, React, Next.js

LANGUAGES

English — B2–C1