

Andrei Panferov

ML RESEARCHER

☎ +7 (926) 431-76-71 | ✉ andrei@panferov.org | 🐙 BlackSamorez | 🌐 in BlackSamorez

Experience

KAUST, Optimization and Machine Learning Lab

RESEARCH INTERN

[Saudi Arabia](#)

July 2023 - September 2023

- Conducted research under the supervision of *Prof. Peter Richtárik*
- Derived theory and ran experiments on *distributed optimization*, focusing on *communication compression*
- Submitted a first-author paper to an upcoming conference, with the preprint soon to be on arxiv

Eqvilent (High Frequency Trading Fund)

SOFTWARE ENGINEER

[Remote](#)

July 2022 - March 2023

Yandex

ML ENGINEER INTERN (NLP)

[Moscow, Russia](#)

March 2022 - July 2022

- Refactored and optimized an *LLM inference framework* enabling abstract *tabular data* insertion for efficient *map-reduce* inference
- Increased test coverage from 0 to 85% through rigorous unit testing
- Took part in developing a *universal LLM benchmarking solution* adapting 2 datasets for it

Terra Quantum AG

RESEARCHER

[Moscow, Russia](#)

July 2020 - July 2022

- Researched *quantum algorithms* for business applications
- Developed an NMR spectra analysis tool, allowing for its use for quantum computations
- Optimized *LLM* deployment for chat assistant applications, reducing latency by 40%

Awards

International Physics Olympiad

GOLD MEDAL

[Israel](#)

Summer 2019

Education

Moscow Institute of Physics and Technology (MIPT)

BACHELOR OF SCIENCE IN APPLIED MATHEMATICS AND PHYSICS

[Moscow, Russia](#)

2020 - 2024

- Achieved a perfect 5.0/5.0 GPA
- Second minor in *Teaching Methods and Pedagogy*

Yandex School of Data Analysis (YSDA)

INDUSTRY-ORIENTED PROGRAM IN MACHINE LEARNING (MSC LEVEL)

[Moscow, Russia](#)

2021 - 2023

- Specialization in *Deep Learning* and *Natural Language Processing*
- Rigorously contributed to *open-source* (see the following section)
- Served as a *TA* for the *NLP* course. Designed a homework on *Model Compression* and graded a class of 200+ students on it

Publications

Correlated Quantization for Faster Nonconvex Distributed Optimization

ANDREI PANFEROV, YURY DEMIDOVICH, AHMAD RAMMAL, PETER RICHTÁRIK

[KAUST, Saudi Arabia](#)

Under Review

Open-Source Contributions

🔗 tensor_parallel

- Developed an open-source *python library* for tensor parallel *PyTorch* models training and inference tightly integrated with *Hugging Face*
- Received more than 400 🌟 on *GitHub*

🤖 HuYaLM-100B

- Adapted *YaLM-100B* LLM specifically for *Hugging Face transformers*, rewriting the officially published *Megatron-LM* implementation

🧐 LLaMA implementation for transformers

- Took part in adapting the *LLaMA* model to the *Hugging Face transformers* library, fixing the positional embedding errors and optimizing past key-value handling

🤖 NLP Bot Project

- Designed an automatic data collection system to extract thousands of dialogues from internet forums, refined the collected data using a pretrained sentiment analysis *BERT* model and published them it a dataset
- Fine-tuned a *GPT-2* model for *chatbot* purposes on the refined dataset and deployed it as a Telegram *bot*
- Published an article on *Habr* about the project, reaching the daily top-1 in the *ML* section