

# Andrei Panferov

ML RESEARCHER

+7 (926) 431-76-71 | [andrei@panferov.org](mailto:andrei@panferov.org) | [BlackSamorez](#) | [in BlackSamorez](#)

## Summary

I am set to graduate this summer with a Bachelor of Applied Mathematics and Physics from the Moscow Institute of Physics and Technology, where I am currently ranked in the top 15 out of more than 200 students in my cohort. I studied Machine Learning at the Yandex School of Data Analysis, and I have a keen interest in Natural Language Processing, efficient Deep Learning and Federated Learning. My achievements range from securing a gold medal at the International Physics Olympiad to making significant contributions to the open-source community, as well as acquiring work experience in both industry and academia.

## Experience

### Yandex Research

ML RESEARCH RESIDENT

[Russia](#)

November 2023 - Present

### 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 (see Publications)

### Eqvilent (High Frequency Trading Fund)

SOFTWARE ENGINEER

[Remote](#)

July 2022 - March 2023

### Yandex

ML ENGINEER INTERN (NLP)

[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 of the *map-reduce* inference interface from *0* to *85%* through rigorous unit testing
- Took part in developing a *universal LLM benchmarking solution* adapting *two datasets* for it

### Terra Quantum AG

RESEARCHER

[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*
- Working toward my thesis on distributed training of Large Language Models under the supervision of *Prof. Alexander Gasnikov*

### Yandex School of Data Analysis (YSDA)

POST-BACHELOR'S PROGRAM IN MACHINE LEARNING

[Moscow, Russia](#)

2021 - 2023

- Completed 12 *MSc* level courses. Specialized in *Deep Learning* and *Natural Language Processing*
- Contributed significantly to *Open-Source* projects (see Open-Source Contributions)
- Served as a *TA* for the *NLP* course. Prepared a seminar on *Model Compression*, challenged the students to implement *GPTQ*

## Publications

### Correlated Quantization for Faster Nonconvex Distributed Optimization

ANDREI PANFEROV, YURY DEMIDOVICH, AHMAD RAMMAL, PETER RICHTÁRIK

[KAUST, Saudi Arabia](#)

Under Review for AISTATS 2024

## Open-Source Contributions

---

### 🔗 **tensor\_parallel**

[GITHUB.COM/BLACKSAMOREZ/TENSOR\\_PARALLEL](https://github.com/BlackSmorez/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 stars on *GitHub*

### 🤖 **LLaMA implementation for transformers**

[HUGGINGFACE.CO/DOCS/TRANSFORMERS/MAIN/MODEL\\_DOC/LLAMA](https://huggingface.co/docs/transformers/main/model_doc/llama)

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

### 🤖 **HuYaLM-100B**

[HUGGINGFACE.CO/BLACKSAMOREZ/HUYALM-100B-FP16](https://huggingface.co/BlackSmorez/HuYaLM-100B-FP16)

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

### 🔗 **NLP Bot Project**

[GITHUB.COM/BLACKSAMOREZ/EBANKO](https://github.com/BlackSmorez/ebanko)

- 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 as 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* (IT social network) about the project, reaching the daily top-1 in the *ML* section