Andrei Panferov

ML Researcher

+7 (926) 431-76-71 | andrei@panferov.org | BlackSamorez | BlackSamorez

Experience

KAUST, Optimization and Machine Learning Lab — *Saudi Arabia*

Research Intern — *July 2023 - September 2023*

- Conducted research under the supervision of *Prof. Peter Richtarik*
- Derived theory and ran experiments on *distributed optimization*, focusing on *communication compression*
- Submitted a first-author paper titled "Correlated Quantization for Faster Nonconvex Distributed Optimization" to an upcoming conference, with the preprint soon to be on arxiv

Eqvilent (High Frequency Trading Fund) — *Remote*

Software Engineer — *July 2022 - March 2023*

Yandex — *Moscow, Russia*

ML Engineer Intern (NLP) — *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 — *Moscow, Russia*

Researcher — *July 2020 - July 2022*

- Researched *quantum algorithms* for business applications
- Developed an NMR spectra analysis toll, allowing for it's use for for quantum computations
- Optimized *LLM* dedployment for chat assistant applications, reducing latency by *40%*

Awards

International Physics Olympiad — *Israel*

Gold Medal — *Summer 2019*

Education

Moscow Institute of Physics and Technology (MIPT) — *Moscow, Russia*

Bachelor of Science in Applied Mathematics and Physics — *2020 - 2024*

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

Yandex School of Data Analysis (YSDA) — *Moscow, Russia*

Industry-Oriented Program in Machine Learning (MSc Level) — *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

Open-Source Contributions

tensor_parallel

- Developed an open-source **python library** for tensor parallel **PyTorch** models training and inference tightly inregrated with **Hugging Face transformers**
- Recieved more than *400* ⭐ on *GitHub*

🤗 HuYaLM-100B

- Adapted **YaLM-100B** LLM specifically for **Hugging Face transformers**, rewriting the officialy 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