Malinovskii Vladimir

ML Researcher · Software Engineer

galqiwi@gmail.com | galqiwi | galqiwi

Work Experience

Yandex Research

ML Resident *Feb. 2024*

- Working on quantization algorithms for LLMs

Yandex (Infrastructure)

Software Engineer *Jun. 2022 - Feb. 2024*

- Was responsible for one of the key parts in an internal kubernetes-like system
- Participated in hiring process by conducting coding interviews

Yandex (YTsaurus)

Software Engineer Intern *Dec. 2021 - Jun. 2022*

- Wrote highload C++ code for tracking memory consumption during data delivery
- Since YTsaurus is opensource, my work is available here (*link*)

Terra Quantum

Part Time ML Engineer *Aug. 2020 - Jun. 2022*

- Developed a GPT-based pipeline for a dialog bot

Education

HSE(Higher School of Economics)

Master of Science in Applied Mathematics and Computer Science *Jun. 2023*

- Joint master's program with the Yandex School of Data Analysis
- Holding perfect 5/5 GPA

MIPT(Moscow Institute of Physics and Technology)

Bachelor of Science in Applied Mathematics and Physics *Sep. 2019 - Jun. 2023*

- Minor in data analysis
- Achieved 4.68/5 weighted GPA

Publications

PV-Tuning: Beyond Straight-Through Estimation for Extreme LLM Compression *NIPS, 2024*

Vladimir Malinovskii*, Denis Mazur*, Ivan Ilin*, Denis Kuznedelev, Konstantin Burlachenko, Kai Yi, Dan Alistarh, and Peter Richtarik

Oral

Awards

2019 **Gold medal,** International Physics Olympiad *Israel*

Pet projects

Llama3.1-8b Inference in browser (AQLM.rs)

Oct. 2024

- I implemented multithreaded CPU inference for quantized Llama 3.1-8b model.
- Demo is available here (link).

Telegram bot for downloading audio

Jul. 2022 - Jul. 2024

- 200k users overall
- 9k daily active users

Technologies

For industrial purposes I prefer to write code in Go. For data science projects I use Python. When I need to develop something close to the hardware, I opt for C++ or Rust.