




Arina Puchkova

 [rinapch](#) |  arina.pchkva@gmail.com |  [rinapch](#)

EXPERIENCE

Ex-Human, Inc

Aug. 2021 – Oct. 2023

Machine Learning Engineer

San Francisco, CA (remote)

- Trained generative models up to 34B in size implementing techniques for optimized GPU training. Employed TPUs to achieve a 2x reduction in training time for GPT-J and T5.
- Deployed and benchmarked models up to 70B using vLLM, TGI, and NVIDIA Triton-LLM frameworks. Achieved a seamless one-liner deployment configuration and maintained <1.5 sec. latency for each model instance.
- Prepared training datasets using both open source and proprietary data (i.e. user feedback and upvotes, regenerated responses) for various instruction-tuning techniques.
- Led the development of safeguards for open-domain conversation. Introduced a set of classifiers aimed at preventing unsafe content like hate-speech, and streamlined safety alignment of the main dialog model. Successfully reduced the share of unsafe messages by 52%.
- Used Python, PyTorch, Flask, FastAPI, Docker, ONNX, TransformerDeploy, vLLM, TGI, Triton, Google Cloud Platform, Runpod, Firebase, Qdrant, Redis

Higher School of Economics

Jan. – July 2022

Big Data and Information Retrieval School

Moscow, Russia

- Introduction to Python Teaching Assistant

Department of Higher Mathematics

Sept. 2019 – May 2021

- Mathematics and Statistics Teaching Assistant

Leroy Merlin

May 2021 – Oct. 2021

Data Analyst Intern

Moscow, Russia

- Worked with the Search and Recommendations Team; received a return offer for a full-time position.
- Prepared reports on search algorithm quality and A/B tests results, conducted CustDev.
- Helped developing language and recommender models. Assisted with a model for keyword extraction, resulting in 11% increase in search accuracy.
- Used Python, SQL, DVC, Airflow, Clickhouse, Greenplum

EDUCATION

Ludwig Maximilian University

Oct. 2023 – Present

MSc Data Science

Munich, Germany

Got accepted into one of the 20 places in a cohort (competition being 25 people per position)

Higher School of Economics

Sept. 2018 – June 2022

BA Political Science, with honors

Moscow, Russia

- * Major in Political Science, GPA 9.36/10
- * Minor in Intellectual Data Analysis, GPA 10/10

Top-1 Student of [2019/2020](#) and [2018/2019](#) academic years, graduated in Top-3 of the class

CONTINUING EDUCATION

Advanced Language Processing Winter School

Jan. 2022

Workshops on Neuro-Symbolic Commonsense Knowledge and Reasoning,
Text Detoxification, Machine Translation, Data and Model Scaling

Grenoble, France (remote)

IPSA-HSE Summer School

Aug. 2019

Attended applied regression course with [Prof. Ponarin](#)

NRU HSE, St. Petersburg, Russia

PUBLICATIONS

1. Alexandr, N., Irina, O., Tatyana, K., Inessa, K., **Arina, P.** (2021). Fine-Tuning GPT-3 for Russian Text Summarization. In: Silhavy, R., Silhavy, P., Prokopova, Z. (eds) Data Science and Intelligent Systems. CoMeSySo 2021. Lecture Notes in Networks and Systems, vol 231. Springer, Cham.
https://doi.org/10.1007/978-3-030-90321-3_61

PROJECTS

Neural Pushkin

- * Parsed and cleaned all Alexander Pushkin's writings, fine-tuned ruGPT-3 on the extracted data.
- * Deployed model as a service, maintained a Telegram bot and a dedicated website.
- * Received over 100 daily visitors (500+ during highload).
- * Featured on [tgproger](#), [exploit.media](#) (in Russian).

SKILLS

Languages : English (C1), Russian (native), German (B2)

Programming Languages : Python, R, JAX, SQL

NLP Stack : Huggingface, TGI, vLLM, trl, DeepSpeed, accelerate, Triton Inference Server, TensorRT-LLM, peft

MLOps : Docker, Flask, FastAPI, Google Cloud Platform, Qdrant, Grafana

Tools : Git, \LaTeX , WandB, Notion