Nikita Pavlichenko

nikita.v.pavlichenko@gmail.com ❖ LinkedIn ❖ Google Scholar ❖ GitHub ❖ Website

WORK EXPERIENCE

JetBrains AI **March 2024 – Present**
Senior Research Engineer *Berlin, Germany*

- **Developing LLMs for Cloud Code Completion** in JetBrains AI: lead the LLM training project, responsible for data processing, **pre-training, fine-tuning, and alignment**. Beating GitHub Copilot.
- Mentoring junior engineers, making architecture and product decisions, establishing development guidelines. Responsible for ML engineers hiring process in the company.
- Trained and deployed LLMs for Code Completion, resulting in a 2x improvement in metrics compared to the baseline OpenAI-based solution. These models are deployed for use in JetBrains' high-selling IDEs, serving millions of users worldwide, and are a core feature of the JetBrains AI product, contributing over $1 million in revenue.

Yandex LLC **September 2020 – March 2024**
AI Research Scientist *Belgrade, Serbia*

Research Contributions:
- Lead deep learning research for Toloka AI, resulting in 6 publications on Natural Language Processing, Computer Vision, and generative AI at NeurIPS, SIGIR, and HCOMP.
- Autored the ICML 2023 tutorial grounded in my original research on RLHF.

Engineering and Development:

- Lead alignment (SFT and RLHF) in Toloka's LLM training project: responsible for dataset collection and trained a model that surpassed Falcon 40B Instruct with >65% win rate on real user requests.
- Develop and maintain an AI-assisted annotation service that showed up to a 30% increase in revenue for text classification projects.
- Develop and maintain the Crowd-Kit Python package that implements Bayesian models for truth inference.
- Authored 3 patent applications.
- Mentored 2 interns and a junior that was promoted to a mid-level position, and conducted more than 120 interviews (more than 40 hired).

MIPT and Inria NANO/D **March 2020 – September 2020**
Research Intern *Moscow, Russia/Grenoble, France*

- Developed and implemented the S-GCN (deep learning model operating on molecular graphs) method for protein model quality prediction that achieved state-of-the-art on CASP MQA challenge.
- Authored a paper on S-GCN accepted to a high-impact journal.

Yandex LLC **July 2018 – November 2018 / July 2019 – October 2019**
Software Engineering Intern / Machine Learning Engineer Intern *Moscow, Russia*

- Developed and implemented a machine learning model for ranking video clips in web crawler schedules. The model allowed to remove more than 90 % of deleted videos from search results.
- Enhanced the performance speed of all search anti-spam systems by 5% by introducing the fastest C++ hash-table implementation in Yandex.

EDUCATION

Moscow Institute of Physics and Technology **September 2021 – June 2023**
Master of Science in Computer Science *Moscow, Russia*

Moscow Institute of Physics and Technology **September 2017 – June 2021**
Bachelor of Science in Computer Science *Moscow, Russia*

SKILLS

Languages: Python, C++, Rust, R, Bash. **Frameworks & Libraries:** PyTorch, TensorFlow, Keras, XGBoost CatBoost, Sci-Kit Learn, NumPy, Pandas, PyTorch Geometric. **Tools & Technologies:** Hadoop, Hive, Spark, AWS, Kubernetes, Docker, Git, Linux, Algorithms, Data Structures. **Domains:** Crowdsourcing, NLP, CV, RL, Graph Machine Learning

PUBLICATIONS

[1] **Pavlichenko, N.**, Stelmakh, I., & Ustalov, D. (2021). CrowdSpeech and Vox DIY: Benchmark Dataset for Crowdsourced Audio Transcription. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, *1*.

[2] Igashov, I., **Pavlichenko, N.**, & Grudinin, S. (2021). Spherical convolutions on molecular graphs for protein model quality assessment. *Machine Learning: Science and Technology*, *2*(4), 045005.

[3] **Pavlichenko, N.**, & Ustalov, D. (2023). Best Prompts for Text-to-Image Models and How to Find Them. *SIGIR '23: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2067–2071.

[4] Ustalov, D., **Pavlichenko, N.**, Losev, V., Giliazev, I., & Tulin, E. (2021). A general-purpose crowdsourcing computational quality control toolkit for python. *The Ninth AAAI Conference on Human Computation and Crowdsourcing: Works-in-Progress and Demonstration Track (HCOMP 2021)*.

[5] **Pavlichenko, N.**, & Ustalov, D. (2021). IMDB-WIKI-SbS: An evaluation dataset for crowdsourced pairwise comparisons. *NeurIPS 2021 Data-Centric AI Workshop*.

[6] Tseytlin, B., **Pavlichenko, N.**, Likhobaba, D., Smirnova, A., & Ustalov, D. *Toloka at LMNL Challenge: Learning from Noisy Labels with Annotator Reliabilities and Self-Supervised Training*.

[7] Ustalov, D., Fedorova, N., & **Pavlichenko, N**. (2022). Improving Recommender Systems with Human-in-the-Loop. *Proceedings of the 16th ACM Conference on Recommender Systems*, 708–709.
Ustalov, D., Pavlichenko, N., Likhobaba, D., & Smirnova, A. (2023). *WSDM Cup 2023 Challenge on Visual Question Answering*.

[8] Ustalov, D., **Pavlichenko, N.**, Stelmakh, I., & Kuznetsov, D. (2021). VLDB 2021 Crowd Science Challenge on Aggregating Crowdsourced Audio Transcriptions. *Proceedings of the 2nd Crowd Science Workshop: Trust, Ethics, and Excellence in Crowdsourced Data Management at Scale*, 1–7.

[9] Ustalov, D., **Pavlichenko, N.**, & Tseitlin, B. (2021). Learning from Crowds with Crowd-Kit. *ArXiv Preprint ArXiv:2109.08584*.

[10] Ustalov, D., Smirnova, A., Fedorova, N., & **Pavlichenko, N.** (2023). Crowdsourcing for Information Retrieval. *European Conference on Information Retrieval*, 357–361.