Viktor Moskvoretskii

vvmoskvoretskii@gmail.com · Google Scholar · Github

AI Researcher with focus on LLM Trustworthiness and Safety.

Education

Ph.D. in Computer Science – EPFL, Lausanne 2025 — Present

M.S. in Machine Learning (with honors) – HSE University, Moscow 2023 — 2025

CGPA: 4/4
Thesis: “Self-Improvement for LLM”
Advisors: Irina Nikishina and Chris Biemann

Graduate Program in AI and Applied Mathematics (with honors) – MSU, Moscow 2021 — 2023

CGPA: 3.65/4

B.S. in Neuroscience (with honors) – HSE University, Moscow 2018 — 2022

CGPA: 3.95/4
Thesis: "Advanced Emprical Research on Grapheme-Color Synesthesia Induction With V4 tDCS Stimulation"
Advisors: Oksana Zinchenko and Alexey Gorin
Link: Publication

Work Experience

Doctoral Researcher — EPFL, DLab 09.2025 — Present

- LLM Pretraining
- LLM Safety

Research Lead — Skoltech 10.2024 — 09.2025

- LLM Uncertainty Estimation
- Hallucinations Detection and Mitigation
- LLM Safety

Research Engineer — Skoltech 07.2023 — 10.2024

- LLM for Lexical Semantics
- Machine Translation
- LLM Efficiency: PEFT and Compression
- Multivariate Time-Series Unsupervised Learning

Intern Researcher — DeepPavlov.ai 08.2022 — 06.2023

- Image2Text Dialogue System Research

Publications

A*, Q1

- [EMNLP 2025]: Pletenev, S., Marina, M., Ivanov, N., Galimzianova, D., Krayko, N., Salnikov, M., ... & **Moskvoretskii, V.** (2025). Will It Still Be True Tomorrow? Multilingual Evergreen Question Classification to Improve Trustworthy QA. arXiv preprint arXiv:2505.21115.
- [EMNLP 2025]: Marina, M., Ivanov, N., Pletenev, S., Salnikov, M., Galimzianova, D., Krayko, N., ... & **Moskvoretskii, V.** (2025). LLM-Independent Adaptive RAG: Let the Question Speak for Itself. arXiv preprint arXiv:2505.04253.
- [ACL 2025]: **Moskvoretskii, Viktor**, et al. Adaptive Retrieval Without Self-Knowledge? Bringing Uncertainty Back Home. arXiv preprint arXiv:2501.12835 (2025).
- [ACL 2025]: Zhelnin, M., **Moskvoretskii, V.**, Shvetsov, E., Venediktov, E., Krylova, M., Zuev, A., & Burnaev, E. (2024). GIFT-SW: Gaussian noise Injected Fine-Tuning of Salient Weights for LLMs. arXiv preprint arXiv:2408.15300.
- [KDD 2025]: Osin, D., Udovichenko, I., **Moskvoretskii, V.**, Shvetsov, E., & Burnaev, E. (2024). EBES: Easy Benchmarking for Event Sequences. arXiv preprint arXiv:2410.03399.
- [ICLR 2025 Workshop]. **Moskvoretskii, Viktor**, Chris Biemann, and Irina Nikishina. "Self-Taught Self-Correction for Small Language Models." arXiv preprint arXiv:2503.08681 (2025).
- [ICLR 2025 Workshop]. **Moskvoretskii, Viktor**, and Narek Alvandian. "Challenges of Multi-Modal Coreset Selection for Depth Prediction." arXiv preprint arXiv:2502.15834 (2025).
- [EMNLP 2024]. **Viktor Moskvoretskii**, Nazarii Tupitsa, Chris Biemann, Samuel Horváth, Eduard Gorbunov, and Irina Nikishina. 2024. Low-Resource Machine Translation through the Lens of Personalized Federated Learning. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 8806–8825, Miami, Florida, USA. Association for Computational Linguistics.
- [ACL 2024]: **Viktor Moskvoretskii**, Ekaterina Neminova, Alina Lobanova, Alexander Panchenko, and Irina Nikishina. 2024. TaxoLLaMA: WordNet-based Model for Solving Multiple Lexical Semantic Tasks. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2331–2350, Bangkok, Thailand. Association for Computational Linguistics.
- Andreev, S., **Moskvoretskii, V.**, Gorin, A., & Zinchenko, O. (2024). Grapheme-color synesthesia induction with V4 transcranial direct current stimulation. Current Psychology , 1-6.
- Andreev, S., **Moskvoretskii, V.**, Gorin, A., & Zinchenko, O. (2023). Induction of grapheme-color synesthesia-like effects in non-synesthetes via offline anodal tdcs over visual cortex in area v4. Brain Stimulation: Basic, Translational, and Clinical Research in Neuromodulation , 16(1), 274.

A, B

- [LREC-COLING 2024]: **Moskvoretskii, V.**, Panchenko, A., & Nikishina, I. (2024, May). Are Large Language Models Good at Lexical Semantics? A Case of Taxonomy Learning. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (pp. 1498-1510).

Preprint and Misc.

- [Submitted to ARR 2025]: Kharinaev, A., **Moskvoretskii, V.**, Shvetsov, E., Studenikina, K., Mikhail, B., & Burnaev, E. (2025). Investigating the Impact of Quantization Methods on the Safety and Reliability of Large Language Models. arXiv preprint arXiv:2502.15799.

- [Submitted to ARR 2025]: **Moskvoretskii, Viktor**, et al. "Do I look like acat. n. 01to you? A Taxonomy Image Generation Benchmark." arXiv preprint arXiv:2503.10357 (2025).
- [Submitted to ARR 2025]: Nikishina, I., Anwar, S., Dolgov, N., Manina, M., Ignatenko, D., **Moskvoretskii, V.**, ... & Biemann, C. (2025). Argument-Based Comparative Question Answering Evaluation Benchmark. arXiv preprint arXiv:2502.14476.
- **Moskvoretskii, V.**, Frolov, A. & Kuznetsov, D. IMAD: Image-Augmented Multi-Modal Dialogue. J Math Sci 285, 72–87 (2024). https://doi.org/10.1007/s10958-024-07434-0
- [Submitted to ICDM 2025]: **Moskvoretskii, Viktor**, et al. "MLEM: Generative and Contrastive Learning as Distinct Modalities for Event Sequences." arXiv preprint arXiv:2401.15935 (2024).

Awards

Yandex Scholarship 2024 — Awarded for exceptional GPA and significant research contributions in 2024.

HSE Academic Scholarship 2024 — Awarded for outstanding academic performance in 2024.

HSE Best Paper 2024 — HSE 2024 best student paper award

HSE Best Paper 2023 — HSE 2023 best student paper award

Student Supervision

6. **Ekaterina Neminova** (Sep 2023—now, HSE)
   *BSc thesis co-supervision with **Irina Nikishina**: "Investigation of Large Language Models for the Taxonomy-related Tasks"*
   led to ACL paper: "TaxoLLaMA: WordNet-based Model for Solving Multiple Lexical Semantic Tasks"

5. **Alina Lobanova** (Sep 2023—now, HSE)
   *BSc thesis co-supervision with **Irina Nikishina**: "Multilingual Large Language Models for predicting IS-A relationships"*
   led to ACL paper: "TaxoLLaMA: WordNet-based Model for Solving Multiple Lexical Semantic Tasks"

4. **Vladimir Ganzhara** (Sep 2024—now, HSE)
   MS project supervision: "Aligning LLM Confidence with Truthfulness with Self-Play"

3. **Gleb Stenin** (Sep 2024—now, HSE)
   MS project supervision: "Aligning LLM Confidence with Truthfulness with Self-Play"

2. **Vlad Knyazhevski** (Sep 2024—now, HSE)
   MS project supervision: "Aligning LLM Confidence with Truthfulness with Self-Play"

1. **Luiza Nigogosova** (Sep 2024—now, HSE)
   MS project supervision: "Aligning LLM Confidence with Truthfulness with Self-Play"

Teaching Experience

2025 **Tensor Decomposition for DL (en)** — YSDA, Moscow
Role: Lecturer, Seminarian, Preparing Homework
Duration: Term 1
Audience: Senior CS Students
Links: Github

2024-2025 **Neural Natural Language Processing (en)** — HSE, Moscow
Role: Lecturer and Seminarian
Duration: Terms 1, 2 and 3
Audience: 4th year BSc students in Fundamental and Applied Linguistics
Links: Github

2023-2024 **Neural Natural Language Processing (en)** — HSE, Moscow
Role: Lecturer, Seminarian, Preparing Homework
Duration: Terms 1, 2 and 3
Audience: 4th year BSc students in Fundamental and Applied Linguistics
Links: Github

2024 **Tensor Decomposition for DL (en)** — Skoltech, Moscow
Role: Lecturer, Seminarian, Preparing Homework
Duration: Term 1

Audience: Industrial AI Researchers
Links: Github

2024 **Tensor Decomposition for DL (en)** — Deep School, Moscow
Role: Lecturer, Seminarian, Preparing Homework
Duration: Term 3
Audience: Industrial AI Researchers
Links: Github

Reviewing

ACL , ICLR , EMNLP , NAACL , COLING , IEEE , AINL ,

Additional Education

Aug 2023 AI Summer School by Skoltech

Jul 2023 Generative Modeling and RL Summer School by AIRI

May 2023 Generative Modeling and AI Theory Summer School by HSE

Feb 2023 Applied Informatics and Mathematics Winter School by HSE

Jul 2022 NLP and CV Summer School by AIRI

Aug 2020 Machine Learning and Applied Mathematics Summer School by MIPT and IITP RAS