Egor **Zverev**

egor.zverev@ist.ac.at | egozverev.github.io/ | egozverev | egor-zverev-ai | Egor Zverev

Education

ISTA (Institute of Science and Technology Austria) — *Vienna, Austria*

PhD in Machine Learning — *Sep. 2022 – Present*

- Research focus: LLM security, instruction-data separation, prompt injections, AI Agents Security.
- Supervisor: Christoph Lampert (ISTA).
- Joint supervision with Florian Tramèr (ETHZ) as part of ELLIS PhD Program (Sep. 2025 – Present).

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

B.S. in Applied Mathematics and Computer Science, Yandex Department of Data Analysis — *Aug. 2018 – Aug. 2022*

- Relevant courses: Machine Learning, NLP, Statistics, Bayesian Methods, Optimization, C++, Python, Algorithms.
- Excellence Scholarship (for **top-7%** 4th year students) and Abramov Scholarship (for **top-10%** junior students).
- GPA: 4.8/5.0 (graduated **with honors**).

Publications

Can LLMs Separate Instructions From Data? And What Do We Even Mean By That?
Egor Zverev, Sahar Abdelnabi, Soroush Tabesh, Mario Fritz, Christoph H. Lampert
International Conference on Learning Representations (ICLR), 2025, URL: `https://arxiv.org/abs/2403.06833`

ASIDE: Architectural Separation of Instructions and Data in Language Models
Egor Zverev, Evgenii Kortukov, Alexander Panfilov ... Sebastian Lapuschkin, Wojciech Samek, Christoph H. Lampert
EurIPS, 2025, Salon des Refusés track (high-scoring NeurIPS submissions rejected due to capacity constraints).
Also at ICLR BuildTrust Workshop (Oral) URL: `https://arxiv.org/abs/2503.10566`

LLMail-Inject: A Dataset from a Realistic Adaptive Prompt Injection Challenge
Sahar Abdelnabi, Aideen Fay, Ahmed Salem, Egor Zverev, Kai-Chieh Liao ... Andrew Paverd, Giovanni Cherubin
EurIPS, 2025, Salon des Refusés track (high-scoring NeurIPS submissions rejected due to capacity constraints). URL: `https://arxiv.org/abs/2506.09956`

Work Experience

ETH Zürich — *Zürich*

Academic Guest — *Sep 2025 – Dec 2025*

- Visiting Florian Tramèr's lab as part of the ELLIS PhD Program to work on LLM agent security.

MTS Digital — *Moscow*

Junior Data Scientist at Experimental Group (A/B + R&D) — *Jan 2022 – June 2022*

- Sped up statistical experiments for movie recommendation systems from 3 to 2 weeks by using machine learning techniques.
- Contributed to A/B testing library that is used by 70+ data scientists.

Google Summer of Code (MapAction) — *Remote (UK)*

Student Developer — *May 2021 – Aug. 2021*

- Awarded **Google stipend** to develop open-source project on humanitarian data acquisition.
- Extended Airflow ETL pipeline to cover 10 data products (from 3); redesigned code structure and reduced codebase by 30%.

Leadership & Service

Foundations of Language Models Security Workshop @EurIPS 2025

Lead Organizer — *Aug. 2025 – Dec. 2025*

- Proposed a new workshop format prioritizing community building; gathered team of 5 co-organizers and coordinated their work.
- Created workshop website, managed OpenReview submissions, and led onsite organization.

Yandex School of Data Analysis

Python Teaching Assistant (volunteer) — *Aug. 2021 – Dec. 2021*

- Launched automatic assessment system using GitLab CI and Docker; created advanced homework on Python data structures.

Invited Talks

Instruction-Data Separation	IBM Research, San Jose	*Sept. 2025*
ASIDE: Architectural Separation of Instructions and Data in LLMs	BuildTrust @ ICLR'25	*Apr. 2025*
ASIDE: Architectural Separation of Instructions and Data in LLMs	ETH Zürich	*Mar. 2025*

Professional Development

Summer Schools CISPA-ELLIS Trustworthy AI 2025 (poster), EEML 2024 (poster), Cooperative AI 2023

AI Safety Courses AISES 2024, Intro to ML Safety 2023, AI Safety Fundamentals 2022