Egor **Zverev**

egor.zverev@ist.ac.at | https://egozverev.github.io/ | egozverev | Egor Zverev

Education

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

PhD in Machine Learning *Sep. 2022 – Present*

- Supervisor: Christoph Lampert (ISTA). Advisors: Florian Tramèr (ETHZ), Ahmad Beirami (Google DeepMind), Dan Alistarh (ISTA).

ELLIS PhD Program | ISTA & ETH Zürich *Vienna & Zürich*

PhD in Machine Learning *Sep. 2025 – Present*

- Jointly supervised by Christoph Lampert (ISTA) and Florian Tramèr (ETHZ).

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
ICLR, 2025

ASIDE: Architectural Separation of Instructions and Data in Language Models
Egor Zverev, Evgenii Kortukov, Alexander Panfilov, Soroush Tabesh, Sebastian Lapuschkin, Wojciech Samek, Christoph H. Lampert
ICLR BuildTrust workshop (Oral), 2025

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
arXiv:2506.09956 2025

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*

- Was awarded a **stipend from Google** to develop an open-source project on humanitarian data acquisition.
- Accelerated humanitarian response by extending an Airflow ETL-pipeline to cover 10 data products instead of 3.
- Redesigned the pipeline's structure completely and reduced the total amount of code by almost 30 % by optimizing legacy code.
- Created unit tests, 11 of which revealed bugs in legacy code.
- Wrote a blog post and detailed technical report on my work.

Organizational and Teaching Experience

Foundations of Language Models Security Workshop @EurIPS 2025

Lead Organizer Team Member *Aug. 2025 – Dec. 2025*

- Proposed a new workshop format that prioritizes community building over talk density.
- Gathered a team of 5 co-organizers and coordinated their work.
- Created workshop website, managed openreview double-blind submissions, created social announcements and led onsite organization.

LLM Safety and Security Workshop @ELLIS UnConference 2025

Lead Organizer Team Member *Oct. 2025 – Dec. 2025*

- Gathered organizer team, managed website and submissions, 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 allocation in memory.

MIPT.Stats

Statistics Teaching Assistant *Aug. 2021 – Feb. 2022*

- Conducted weekly one-on-one meetings with twelve 3rd year stats students and checked practical homework.

Talks

Instruction-Data Separation	IBM Research (San Jose)	*Sept. 2025*
ASIDE: Architectural Separation of Instructions and Data in Language Models	BuildTrust@ ICLR2025	*Mar. 2025*
Architectural Separation of Instructions and Data in Language Models	ETH Zürich	*Mar. 2025*
Instruction-Data Separation (Thesis Proposal Defense)	ISTA	*Jul. 2024*

Fellowships, Courses and Summer Schools

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

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