

Evgenii Kortukov

☎ +49 176 515 78226 | ✉ eeekortukov@gmail.com | 🌐 [Website](#) | [LinkedIn](#) | [GitHub](#) | 📍 Berlin, Germany

EDUCATION

University of Tübingen

M.Sc. in Machine Learning; GPA: 1.1/1.0

Tübingen, Germany

Oct 2021 – Mar 2024

- Thesis: Studying language model behavior under realistic knowledge conflicts.

Lomonosov Moscow State University

B.Sc. in Fundamental Informatics and Information Technology; GPA: 4.78/5.0

Moscow, Russia

Sep 2016 – Jun 2020

- Diploma with Honours. Thesis: Finding similar supercomputer applications using structural pattern recognition methods.

RESEARCH EXPERIENCE

Fraunhofer Heinrich Hertz Institute

Research Associate (PhD Candidate), [XAI group](#)

Berlin, Germany

Sep 2024 – Present

- Investigating actionable interpretability methods to advance the security and safety of LLM deployment. [1, 2, 3].
- Supervisors: Prof. Dr. Wojciech Samek, Dr. Sebastian Lapuschkin

Tübingen AI center

Student Research Assistant, [Scalable Trustworthy AI group](#)

Tübingen, Germany

Jun 2023 – Mar 2024

- Researched LLM knowledge updating behavior under knowledge conflicts. [4].
- Contributed to a qualitative study on ML practitioners' views of training data attribution methods. [5]
- Supervisors: Prof. Dr. Seong Joon Oh, Alexander Rubinstein and Elisa Nguyen

University of Tübingen

Student Research Assistant, [Decision Making Group](#)

Tübingen, Germany

Oct 2021 – Jan 2023

- Designed a numerical evaluation environment for applying multi-armed bandit algorithms to large-scale recommender systems [6].
- Worked on a project extending the contextual multi-armed bandit problem to the settings of distribution shifts and costly information acquisition [7].
- Supervisors: Prof. Dr.-Ing. Setareh Maghsudi and Dr. Saeed Ghoorchian

ENGINEERING EXPERIENCE

Yandex

Software Developer, [Storage Engineering Department](#)

Moscow, Russia

Feb 2019 – Jun 2021

- Automated and enhanced storage cluster management system, responsible for data recovery, load balancing, and background processes that maintain fault-tolerance, robustness, and availability. This reduced time of human intervention more than tenfold, making the storage more scalable.
- Co-developed a distributed system that makes storage more scalable through managing metadata in a separate database.
- Upgraded and extended the storage statistics collection process from backend to frontend (using C++, Python, and JavaScript).

PROJECTS

SPAR Fall 2025 Mentee

Remote, Global

Sep 2025 – Jan 2026

- Project Telos: A Behavioural and Representational Evaluation of Goal-Directedness in Language Model Agents
- Designed, implemented, and evaluated cognitive map probes on a GPT-OSS solving a navigation task. [8].

Tübingen AI Exhibition "Cyber and the City"

Tübingen, Germany

Aug 2022 – Feb 2023

- Developed and built an interactive museum exhibit showing and explaining an open-source text2image model to the Tübingen public as part of a team practical project (using Diffusers library, PostgreSQL, and a React frontend).
- The code is [available here](#). The exhibition won the [DFG Communicator award](#).

TEACHING EXPERIENCE

Trustworthy ML

Teaching Assistant

- Held tutorials and created homeworks for the Explainability part of the lecture.
- [Course website](#)

Tübingen, Germany

Oct 2023 – Apr 2024

SKILLS

Technical: LLMs, Interpretability tools, Deep Learning, NLP, RAG, Explainable AI, Classical ML, GNU/Linux, Databases, Git

Programming: Python, Pytorch, LLM stack (Huggingface, vllm, nnsight), C++

Languages: English (C2, IELTS 8.5), German (C1).

PUBLICATIONS

- [1] E. Kortukov, P. Komorowski, F. Klein, P. Engl, G. Sarti, S. J. Oh, S. Lapuschkin, and W. Samek, “[Predicting Future Behaviors in Reasoning Models Enables Better Steering](#),” in *Mechanistic Interpretability Workshop at ICML*, 2026.
- [2] A. Panfilov*, **Evgenii Kortukov***, K. Nikolić, M. Bethge, S. Lapuschkin, W. Samek, A. Prabhu, M. Andriushchenko, and J. Geiping, “[Strategic Dishonesty Can Undermine AI Safety Evaluations of Frontier LLMs](#),” in *ICLR*, 2026.
- [3] E. Zverev, **Evgenii Kortukov**, A. Panfilov, A. Volkova, R. Tabesh, S. Lapuschkin, W. Samek, and C. H. Lampert, “[ASIDE: Architectural Separation of Instructions and Data in Language Models](#),” in *ICLR*, 2026.
- [4] **Evgenii Kortukov**, A. Rubinstein, E. Nguyen, and S. J. Oh, “[Studying Large Language Model Behaviors Under Context-Memory Conflicts With Real Documents](#),” in *COLM*, 2024.
- [5] E. Nguyen, **Evgenii Kortukov**, J. Song, and S. J. Oh, “[Exploring Practitioner Perspectives On Training Data Attribution Explanations](#),” in *Neurips 2023 Workshop XAI in Action*, 2023.
- [6] S. Ghoorchian, **Evgenii Kortukov**, and S. Maghsudi, “[Non-stationary Linear Bandits with Dimensionality Reduction for Large-Scale Recommender Systems](#),” *IEEE Open Journal of Signal Processing*, 2024.
- [7] S. Ghoorchian, **Evgenii Kortukov**, and S. Maghsudi, “[Contextual Multi-Armed Bandit with Costly Feature Observation in Non-stationary Environments](#),” *IEEE Open Journal of Signal Processing*, 2024.
- [8] R. Arghal, F. Chen, N. Dalton, **Evgenii Kortukov**, C. McNamara, A. Nalmpantis, M. Nirvaan, G. Sarti, and M. Giulianelli, “[A Behavioural and Representational Evaluation of Goal-Directedness in Language Model Agents](#),” in *ICML*, 2026.