Evgenii Kortukov

+49 176 515 78226 | eekortukov@gmail.com | Website | LinkedIn | GitHub | Berlin, Germany

Education

University of Tübingen — Tübingen, Germany
M.Sc. in Machine Learning; ***GPA: 1.1/1.0*** — *Oct 2021 – Mar 2024*
- Thesis: Studying language model behavior under realistic knowledge conflicts.

Lomonosov Moscow State University — Moscow, Russia
B.Sc. in Fundamental Informatics and Information Technology; ***GPA: 4.78/5.0*** — *Sep 2016 – Jun 2020*
- Diploma with Honours. Thesis: Finding similar supercomputer applications using structural pattern recognition methods.

Research Experience

Fraunhofer Heinrich Hertz Institute — Berlin, Germany
Research Associate (PhD Candidate), ***XAI group*** — *Sep 2024 – Present*
- Investigating actionable interpretability methods to advance the security and safety of LLM deployment. [1, 2].
- Supervisors: Prof. Dr. Wojciech Samek, Dr. Sebastian Lapuschkin

Tübingen AI center — Tübingen, Germany
Student Research Assistant, ***Scalable Trustworthy AI group*** — *Jun 2023 – Mar 2024*
- Researched LLM knowledge updating behavior under knowledge conflicts. [3].
- Contributed to a qualitative study on ML practitioners' views of training data attribution methods. [4]
- Supervisors: Prof. Dr. Seong Joon Oh, Alexander Rubinstein and Elisa Nguyen

University of Tübingen — Tübingen, Germany
Student Research Assistant, ***Decision Making Group*** — *Oct 2021 – Jan 2023*
- Designed a numerical evaluation environment for applying multi-armed bandit algorithms to large-scale recommender systems [5].
- Worked on a project extending the contextual multi-armed bandit problem to the settings of distribution shifts and costly information acquisition [6].
- Supervisors: Prof. Dr.-Ing. Setareh Maghsudi and Dr. Saeed Ghoorchian

Engineering Experience

Yandex — Moscow, Russia
Software Developer, ***Storage Engineering Department*** — *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. [7].

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 — Tübingen, Germany
Teaching Assistant — *Oct 2023 – Apr 2024*

- Held tutorials and created homeworks for the Explainability part of the lecture.
- Course website

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] 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.

[2] 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.

[3] **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.

[4] 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.

[5] 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.

[6] 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.

[7] 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 *ICLR 2026 Workshop World Models*, 2026.