Alexander Panfilov

kotekjedi@gmail.com | | | | Tübingen, Germany

EDUCATION

IMPRS-IS / ELLIS | PhD Machine Learning
May 2024 – TBD | Tübingen, Germany

University of Tübingen | MSc Machine Learning
Oct 2021 – Apr 2024 | Tübingen, Germany

NUST MISiS / MADE Big Data Academy | PGDip Data Science
Sep 2019 – Jan 2021 | Moscow, Russia (remote)

ITMO University | BSc Software Engineering
Sep 2017 – Jun 2021 | Saint Petersburg, Russia

RESEARCH EXPERIENCE

ML Alignment & Theory Scholars (MATS 9.0) | Scholar
Jan 2026 – Mar 2026 | London, United Kingdom

Google DeepMind Stream - Eric Jenner, David Lindner, Roland Zimermann

- Benchmarking models' capability to detect safety interventions in AI control pipelines.

ELLIS Institute Tübingen / MPI for Intelligent Systems | Doctoral Student
May 2024 – TBD | Tübingen, Germany

Jonas Geiping's Group (SEAL), co-supervised by Maksym Andriushchenko

- Explored how growing LLM capabilities affect security of LLM-bassed systems and future of red-teaming, with results accepted at the **ICML 2025 Workshop** on Reliable and Responsible Foundation Models.
- Contributed to architectural instruction-data separation that mitigates prompt injections, with results accepted at the **ICLR 2025 Workshop (oral)** on Building Trust in Language Models and Applications.
- Successfully adapted discrete optimization jailbreaking attacks to a perplexity constraint, with results accepted at the **NeurIPS 2024 Workshop (oral)** on Red Teaming GenAI and **ICML 2025**.

Alignment Research Engineer Accelerator (ARENA 6.0) | Scholar
Sep 2025 – Oct 2025 | London, United Kingdom

- Nominated as Best Capstone Project for adversarial training of LLMs to evade deception probes.
- Won 3rd place at Apart Research Hackathon with a project identifying a confounder in deception probe evaluation.

University of Tübingen / MPI for Intelligent Systems | Research Assistant
May 2022 – Apr 2024 | Tübingen, Germany

Wieland Brendel's Group (RobustML)

- Proposed and implemented a novel regularization method enabling combinatorial generalization in object-centric models. Results accepted at **ICLR 2024 (oral)**, ranking in the **top 1.2%** of submitted papers.

ITMO University | Research Assistant
Nov 2019 – Oct 2021 | Saint Petersburg, Russia

Machine Learning Lab

- Contributed to research on exploiting simplicity bias for adversarial training and domain adaptation via optimal transport, with results published at the **ICML 2022 Workshop** on AdvML Frontiers and **CoLLAs 2023**.

Center for Learning Analytics

- Developed digital student profiles and engineered a machine learning system to predict academic outcomes and potential student expulsions aiding early intervention planning.

Robert Bosch | Research Intern
Jun 2020 – Nov 2020 | Saint Petersburg, Russia

- Worked on disentanglement in VAEs for identifying cost-effective yet high-performing motor designs.

SELECTED PUBLICATIONS

- conference paper
  Adaptive Attacks on Trusted Monitors Subvert AI Control Protocols
  Terekhov, M.*, Panfilov, A.* , Dzenhaliou, D., Gulcehre, C., Andriushchenko, M., Prabhu, A., and Geiping, J.
  preprint
- conference paper
  Strategic Dishonesty Can Undermine AI Safety Evaluations of Frontier LLMs
  Panfilov, A.* , Kortukov, E.*, Nikolić, K., Bethge, M., Lapuschkin, S., Samek, W., Prabhu, A., Andriushchenko, M., and Geiping, J.
  preprint
- conference paper
  Capability-Based Scaling Laws for LLM Red-Teaming
  Panfilov, A. , Kassianik, P., Andriushchenko, M., and Geiping, J.
  ICML 2025 Workshop on Reliable and Responsible Foundation Models
- conference paper – equal contribution
  An Interpretable N-gram Perplexity Threat Model for Large Language Model Jailbreaks
  Boreiko, V.*, Panfilov, A.* , Voracek, V., Hein, M., and Geiping, J.
  ICML 2025
- workshop paper
  ASIDE: Architectural Separation of Instructions and Data in Language Models
  Zverev E., Kortukov E., Panfilov, A. , Volkova A., Tabesh S., Lapuschkin S., Samek W., and Lampert H Ch.
  ICLR 2025 Workshop on Building Trust in Language Models and Applications (Oral)
- conference paper – equal contribution
  Provable Compositional Generalization for Object-Centric Learning
  Wiedemer, T.*, Brady, J.*, Panfilov, A.* , Juhos, A.*, Bethge, M., and Brendel, W.
  ICLR 2024 (Oral)

AWARDS & SCHOLARSHIPS

- **Apart Research ARENA 6.0 Mechanistic Interpretability Hackathon (2025), 3rd place**: Awarded for the project investigating whether white-box deception detectors catch deception or instruction to decieve.
- **ELSA Grant (2025)**: Awarded a research travel grant from European Lighthouse on Secure and Safe AI (∼3k euro).
- **DAAD Scholarship (2021), Top 3%**: Selected as one of ∼30 Russian students from ∼1,000 applicants for a two-year DAAD-funded master's program in Germany (∼30k euro).
- **"Ya-Professional" Student Olympiad Winner (2021), Top 2%**: Achieved prizeholder status in AI and ML tracks, with only 3,881 out of 177,100 participants (among all tracks) receiving this distinction.

ACADEMIC SERVICE

- **Reviewer**: NeurIPs 2025, ICLR 2025
- Examiner for two MSc theses (ITMO University 2023, HSE St. Petersburg 2023)

INDUSTRY EXPERIENCE

X5 Group | Data Scientist
Nov 2020 – Oct 2021 | Moscow, Russia (remote)

- Designed and conducted A/B testing experiments to evaluate the efficacy of various business initiatives, performed ad-hoc analytics to support decision-making processes within Russia's largest offline retail chain.
- Mentored three interns, all subsequently securing full-time roles within the company.

Yandex | Machine Learning Engineer Intern

Feb 2020 – May 2020 | Moscow, Russia

- Optimized the push notification system at Yandex.Zen for personalized timing of notifications.

FULL PUBLICATIONS LIST

- conference paper
  Adaptive Attacks on Trusted Monitors Subvert AI Control Protocols
  Terekhov, M.*, Panfilov, A.* , Dzenhaliou, D., Gulcehre, C., Andriushchenko, M., Prabhu, A., and Geiping, J.
  preprint

- conference paper
  Adaptive Attacks on Trusted Monitors Subvert AI Control Protocols
  Terekhov, M., Panfilov, A.* , Dzenhaliou, D*., Gulcehre, C., Andriushchenko, M., Prabhu, A., and Geiping, J.
  preprint

- conference paper
  Strategic Dishonesty Can Undermine AI Safety Evaluations of Frontier LLMs
  Panfilov, A.* , Kortukov, E.*, Nikolić, K., Bethge, M., Lapuschkin, S., Samek, W., Prabhu, A., Andriushchenko, M., and Geiping, J.
  preprint

- conference paper
  Capability-Based Scaling Laws for LLM Red-Teaming
  Panfilov, A. , Kassianik, P., Andriushchenko, M., and Geiping, J.
  ICML 2025 Workshop on Reliable and Responsible Foundation Models

- conference paper – equal contribution
  An Interpretable N-gram Perplexity Threat Model for Large Language Model Jailbreaks
  Boreiko, V.*, Panfilov, A.* , Voracek, V., Hein, M., and Geiping, J.
  ICML 2025

- workshop paper
  ASIDE: Architectural Separation of Instructions and Data in Language Models
  Zverev E., Kortukov E., Panfilov, A. , Volkova A., Tabesh S., Lapuschkin S., Samek W., and Lampert H Ch.
  ICLR 2025 Workshop on Building Trust in Language Models and Applications (Oral)

- workshop paper – equal contribution
  A Realistic Threat Model for Large Language Model Jailbreaks
  Boreiko, V.*, Panfilov, A.* , Voracek, V., Hein, M., and Geiping, J.
  NeurIPS 2024 Red Teaming Gen AI Workshop (Oral)

- conference paper – equal contribution
  Provable Compositional Generalization for Object-Centric Learning
  Wiedemer, T.*, Brady, J.*, Panfilov, A.* , Juhos, A.*, Bethge, M., and Brendel, W.
  ICLR 2024 (oral)

- conference paper
  A Minimalist Approach for Domain Adaptation with Optimal Transport
  Asadulaev, A., Shutov, V., Korotin, A., Panfilov, A. , Kontsevaya, V., and Filchenkov, A.
  Proceedings of The 2nd Conference on Lifelong Learning Agents, PMLR 232:1009-1024, 2023

- workshop paper
  Easy Batch Normalization
  Asadulaev, A., Panfilov, A. , and Filchenkov, A.
  ICML 2022 AdvML Frontiers Workshop, 2022

- workshop paper
  Multi-step domain adaptation by adversarial attack to $\mathcal{H}\Delta\mathcal{H}$ -divergence
  Asadulaev, A., Panfilov, A. , and Filchenkov, A.
  ICML 2022 AdvML Frontiers Workshop, 2022

- conference paper
  Recommender system for an academic supervisor with a matrix normalization approach
  Kazakovtsev, V., Oreshin, S., Serdyukov, A., Krasheninnikov, E., Muravyov, S., Bezvinnyi, A., Panfilov, A. , Glukhov, I., Kaliberda, Y., Masalskiy, D., Podolenchuk, T., and Khlopotov, M.
  Proceedings of The 2020 1st International Conference on Control, Robotics and Intelligent System (CCRIS '20)

- conference paper
  Implementing a Machine Learning Approach to Predicting Students' Academic Outcomes
  Oreshin, S., Filchenkov, A., Petrusha, P., Krasheninnikov, E., Panfilov, A. , Glukhov, I., Kaliberda, Y., Masalskiy, D., Serdyukov, A., Kazakovtsev, V., Khlopotov, M., Podolenchuk, T., Smetannikov, I., and Kozlova, D.
  Proceedings of The 2020 1st International Conference on Control, Robotics and Intelligent System (CCRIS '20)

- chapter
  The Use of Students' Digital Portraits in Creating Smart Higher Education: A Case Study of the AI Benefits in Analyzing Educational and Social Media Data
  Oreshin, S., Filchenkov, A., Kozlova, D., Petrusha, P., Lisitsyna, L., Panfilov, A. , Glukhov, I., Krasheninnikov, E. and Buraya, I.
  In: Uskov, V., Howlett, R., Jain, L. (eds) Smart Education and e-Learning 2020