

SOFIIA NIKOLENKO

Munich, Germany | sofia.nikolenko@campus.lmu.de | github.com/ssophiee

M.Sc. candidate in Statistics and Data Science at LMU Munich specializing in Machine Learning with focus on building trustworthy AI systems. **Research interests:** AI Safety, Uncertainty Quantification, Mechanistic Interpretability. **Current research:** moral reasoning in refusal mechanisms (SODA Lab), and deception detection in LLMs (Günemann Lab, TUM).

EDUCATION

Ludwig Maximilian University of Munich

October 2024 – Present

M.Sc. in Statistics and Data Science, Machine Learning Track

Expected completion: 2026

- Selected for Konrad Zuse School of Excellence in Reliable AI (16 of 200 applicants)
- Seminar Project: Uncertainty Quantification techniques to detect Adversarial Attacks on LLMs → Developed into research collaboration with Dr. Shireen Kudukil Manchingal
- Applied Deep Learning: Benchmarking Bayesian vs. Deterministic Approaches to Text Generation

Plekhanov Russian University of Economics

September 2021 – July 2024

B.Sc. in Analytical Economics and Econometrics | GPA: 4.93/5.0 (Top 10% of graduates)

RESEARCH EXPERIENCE

Research Assistant, Data Analytics and Machine Learning Group, TUM

July 2026– Present

Supervisor: PhD student Martin Hu, Dr. Leo Schwinn, Prof. Dr. Stephan Günemann

Focus: AI Safety through LLM Alignment

- Detecting deception in black-box LLMs through interrogation

Research Assistant, SODA Lab, LMU Munich

December 2025– Present

Supervisor: Prof. Dr. Frauke Kreuter and PhD student Sarah Ball

Focus: AI Safety through LLM Alignment

- Investigating how moral foundations contribute structural depth to refusal mechanisms in LLMs, and whether that depth makes refusal more robust against adversarial and harmful prompts

Konrad Zuse School of Excellence in Reliable AI

October 2024 – Present

Supervisor: PhD student Shuo Chen (Tresp Lab)

Two research projects on AI safety

- Adversarial Attacks on Agentic Systems: Investigated how system topology influences vulnerability to attacks in agentic systems
 - Conducted literature review on memorization and training data leakage in LLMs and agentic system architectures
 - Designed and implemented multiple agent configurations (simple two-agent systems to complex multi-step architectures with quality control and guardrails), testing various attack methods (prompt injection, system information extraction)
- Multimodal LLM Robustness: Developing approach to ground reasoning of MLLMs into the visual content
 - Transferring RLM (Recursive Language Model) approach to MLLMs

SELECTED PUBLICATIONS & OPEN-SOURCE

- What Intermediate Layers Know: Jailbreak Signals in Entropy Traces. — ECML PKDD 2026. A short version accepted at EIML@ICML 2026
- Credal Large Language Models for Semantic Commitment under Uncertainty — [Under review]

- Collaborated with a team to extend **shapig** (Shapley-based explanation library, 762★) with a vision-language model explanation module; my role: implementing imputer/masker components, running benchmark validation experiments, and adding tests, examples, and fixes to the corresponding open PR to the main repository

PROFESSIONAL EXPERIENCE

Junior+ Data Scientist, Data & Solutions Group (Kept, ex-KPMG)

March 2024 – December 2025

Moscow, Russia (Part-time, remote after October 2024)

- Developed a RAG-based QA system for financial declarations: built retrieval and generation pipeline over internal documentation, enabling employees to query declarations in natural language.
- Built a semantic search engine for corporate Outlook mail: implemented embedding-based retrieval on LanceDB, improving the speed of email lookup for employees.
- Sole technical owner (ML, API, frontend, Docker) of a dynamic financial modeling tool for a top-5 Russian bank: fine-tuned a QA model to extract economic indicators from natural-language queries, built the API layer integrating it with the bank's BI platform for real-time forecast access, and containerized the full pipeline for deployment.
- Built a text classification system for asset standardization, achieving 90%+ F1-score and increasing appraiser efficiency.

Solution Engineer Intern, Toloka.AI (Yandex) October 2022 – March 2023 Moscow, Russia

- Designed fraud detection systems to identify malicious actors on data labeling platform - foundational experience that shaped research interest in adversarial robustness
- Created multilingual data collection projects (Japanese, German, English) for FAANG client with 90%+ accuracy
- Optimized quality control hyperparameters using Majority Vote method, balancing labeling quality with user retention
- Designed A/B testing framework to evaluate instruction impact on labeling quality

Teaching Assistant, Sberbank Corporate University

November 2023 – November 2024

Courses: Digital Teams (RL, Time Series, Clustering)

- Provided instruction on Reinforcement Learning applications for stock trading and price prediction with boosting algorithms

LEADERSHIP & SERVICE

Youth Representative, United Nations Internet Governance Forum

June 2023

Kyoto, Japan (fully-funded)

- Engaged with international community on AI regulation and cybersecurity policy
- Participated in discussions with representatives from Meta, GitHub, and various government organizations

Co-founder & Speaker, REU Data Science Club

November 2021 – November 2024

2000+ members, 5000+ course participants

- Built knowledge base with 4 tracks and 50+ courses (ds-club-roadmap.notion.site)
- Mentored three students in Data Science career preparation
- Delivered "Introduction to Machine Learning" course to 5000+ participants

HONORS & AWARDS

- Konrad Zuse School of Excellence in Reliable AI Scholarship (16 of 200 applicants) | 2024–Present
- United Nations Internet Governance Forum, Youth Representative (fully-funded) | Kyoto, Japan, 2023
- Kufstein Tirol University Student Exchange (Top 5 of 250 applicants) | 2023
- Changellenge Cup IT Russia High Quality Award (Top 15%) | 2022

SKILLS & LANGUAGES

AI Safety & Robustness: Adversarial attacks, uncertainty quantification, Bayesian inference, credal sets, RAG systems, LLM fine-tuning (LoRA, PEFT), influence functions, jailbreak attacks

Technical Skills: Python, R, SQL, PyTorch, TensorFlow, Transformers (HuggingFace), scikit-learn, Docker, Git

Languages: English (C1, IELTS 7.5), Russian (Native)