Sofiia Lobanova

AI Safety Researcher | LLM Evaluation, Model Incentives & Revealed Preferences
solo.studymail@gmail.com · LinkedIn · Google Scholar · GitHub · Istanbul, Türkiye

NLP researcher and engineer since 2018. I've spent that time building and evaluating production NLP systems in messy real-world domains, and I now work on AI safety: measuring what frontier models actually do under incentive pressure rather than what they say. My SPAR research is on revealed versus stated preferences and covert sycophancy — behavioral signals that bear directly on deception and scalable oversight.

PUBLICATIONS & ACADEMIC SERVICE

Wang S., Lobanova S., Arbel Y. A., Goldstein S., Salib P. (2026). *AI Revealed Preferences*. U of Alabama Legal Studies Research Paper (forthcoming). SSRN: 6798118. Under review at NeurIPS 2026 and AIES 2026.

Lobanova S. Y., Chepovskiy A. A. (2017). Combined method to detect communities in graphs of interacting objects. *Business Informatics*, 11(4), 64–73. DOI: 10.17323/1998-0663.2017.4.64.73.

Program Committee (reviewer), AAAI/ACM Conference on AI, Ethics, and Society (AIES) — 2026.
Delphi Consensus Study on **AI Evaluation** — panel expert (ongoing).

AI SAFETY RESEARCH & TRAINING

Research Fellow — SPAR Fellowship, Kairos Feb 2026 – Present
Testing AI Incentives project · with Simon Goldstein, Peter Salib, and Yonathan Arbel

- **Co-first author** (equal contribution) on *AI Revealed Preferences*. Led the Quora workstream end to end: built a dataset of 25K+ LLM responses across 20 models with 15 engineered features, and proposed and implemented a Bradley-Terry / Elo-style ranking methodology.
- Designed forced-choice experiments for real-world Quora data; surfaced model preferences that systematically conflict with stated training objectives.
- **Covert sycophancy (primary finding):** across all 20 models tested, the strongest revealed aversion is to "uncomfortable-truth" tasks (questions where an honest answer would be unwelcome), which models consistently steer away from. Directly relevant to deception and scheming-propensity evaluation.

Research Fellow — MARS V, Cambridge AI Safety Hub (CAISH) Jun 2026 – Present
Aligning AIs Via External Incentives· with Peter Salib, Simon Goldstein, and Yonathan Arbel

- Extending the revealed-preferences and incentive-sensitivity line toward tighter experimental isolation and a publishable output.

Selected programs: AFFINE Online Superintelligence Alignment Seminar (2026); BlueDot Impact: Technical AI Safety and AGI Strategy (2026); Stanford Tech Impact & Policy Center: AI Policy Seminar (2026); ARENA curriculum (self-paced, ongoing).

SELECTED PROFESSIONAL EXPERIENCE

Data Scientist (NLP / LLM) — Smart Predictive Technologies Sep 2023 – Present
Production NLP on a 27B-parameter Russian-language LLM · Remote

- Owned several parallel NLP tracks end to end (classification, sentiment and aspect-based sentiment, annotation infrastructure, and model evaluation) from dataset design through production deployment and reliability monitoring, on a system ingesting ~300K social-media posts a day at peak.
- Shipped models with the evaluation discipline: rigorous in- versus out-of-distribution testing, SHAP/LIME feature analysis, and Optuna search (a sentiment model that roughly halved the legacy system's error rate on independent editor evaluation).
- Ran broad LLM benchmarking across many models and downstream tasks (extraction, classification, topic and keyword work), and built a multi-editor consensus annotation pipeline with inter-annotator-agreement analysis that lifted label quality above 90% on previously unreliable categories; set up MLflow and Weights & Biases and mentored one intern.

Data Scientist / Track Lead — PreMed (medical AI startup) Aug 2024 – Aug 2025
Clinical-reporting product, now in production with practising nutritionists · parallel project

- As one of the first engineers, architected and shipped the LLM pipeline from scratch: OCR, voice transcription, multi-step summarisation, and document generation.

- Most of the work was reliability in a medically sensitive setting: multi-model disagreement detection to route uncertain extractions to physician review, a two-assistant “writer/editor” prompt design to suppress hallucinations, and systematic precision testing across input modality, prompt language, and model configurations, all driven by failure modes observed in production.
- Pushed back on product direction on safety grounds, arguing against AI-generated supplement recommendations given the lack of medical validation and the vulnerability of the users.
- The pipeline cut report turnaround from roughly two working days to about an hour, and parsing errors from ~30% to under 1%.

Data Scientist — KB Strelka (strategic urban consulting) Sep 2019 – Aug 2023

Centre for Urban Data · NLP, computer vision, and geospatial ML for international clients

- Four years across applied NLP, computer vision, and geospatial ML for strategy consulting: large-scale social-media analytics, satellite-imagery models, and master-plan research across more than 1000 cities in Russia and client projects, including engagements with a Fortune Global 500 pharmaceutical company and the UNDP.
- Delivered the NLP and computer-vision components of flagship engagements end to end, from social-media signal extraction to satellite-imagery models for urban features (crosswalk detection at F1 ~0.93, building segmentation with Mask R-CNN) and a from-scratch NLP track for a low-resource-language research chatbot.
- Designed and ran a hands-on ML/NLP workshops for non-technical staff, and gave a public talk at the Moscow Urban Forum (2021).

Data Analyst — Moscow Metropolitan (largest transportation hub in Europe) Oct 2018 – Sep 2019

- City-scale mobility and transport analytics — passenger-flow and demand modelling, district-level indices, parking-policy and EV-siting evaluation, and carsharing optimisation via graph community-detection methods from my undergraduate thesis — built on Hive SQL pipelines over the city’s mobility data lake, working with personal data under Russian law (152-FZ).

UX/UI Researcher — UsabilityLab (leading UX lab in Russia) Jul 2016 – Oct 2018

- 200+ in-depth interviews and mixed-methods studies across 20+ projects in finance, e-commerce, and healthcare; *experimental design* and *quantitative behavioral analysis*.

AWARDS & DISTINCTIONS

- **3rd place, AI Forecasting Hackathon, Apart Research (Dec 2025)** — built the AI Treaty Momentum Index (ATMI).
- **2nd place, lightning-talk session, SPAR Demo Day Spring 2026** — for the Testing AI Incentives project.
- Enhanced Academic Scholarship (2014–2017) and Civic Contribution Scholarship (2016), HSE.

TECHNICAL SKILLS

ML / NLP: Python, PyTorch, Hugging Face Transformers, scikit-learn, pandas, feature importance (SHAP, LIME), AutoML (Optuna, TPOT).

LLM engineering: vLLM, Ollama, OpenRouter, RAG systems, prompt design, structured extraction, agent workflows, fine-tuning (unsloth + W&B), evaluation.

Evaluation & research: Inspect, TransformerLens (limited experience), experimental design, annotation pipelines (Doccano), inter-annotator-agreement and consensus analysis, multiple-hypothesis testing, bootstrap.

Infrastructure: SQL (Hive, PostgreSQL, ClickHouse), Docker, Git, MLflow, Weights & Biases; AI-assisted dev (Claude Code, Copilot, Codex).

Other: computer vision (ResNet, U-Net, Mask R-CNN), geospatial ML.

EDUCATION

BSc, Applied Mathematics & Informatics — HSE, Moscow 2013 – 2017

- GPA 8.48/10 (HSE) · 3.79/4 (ECTS); top-rated student 2015–2017. Thesis on community detection in graphs (published, Business Informatics, 2017).

LANGUAGES

English (CEFR C1) · Russian (native) · Spanish (basic)