Sergei Smirnov

Research Engineer in AI Safety | AI Alignment and Evals | 5+ years of experience

mg1sergey@gmail.com ◇ +358 40 475 9635 ◇ linkedin.com/in/seriozh1 ◇ github.com/seriozh1

ABOUT

AI Safety Research Engineer with **5+ years** in production ML and current research focused on **evaluating and aligning AI character, values and behavior**. Owned end-to-end ML development as a **founding engineer**, from requirements and GPU infrastructure to model development and analytics; contributed to open-source distributed ML pipeline library. Built the public AI Values demo for CivAI. Replicated Anthropic's agentic misalignment experiments and developed a multi-turn extension investigating how prior deception affects subsequent agent behavior during **ARENA** (FAEB). Research Fellow at **MARS** and **SPAR**.

Founding Research Scientist, Cedar Research — Dec 2026 – Present

- Research on AI character evaluation.

Doctoral Researcher, AI Alignment and Evals, University of Helsinki — Sep 2024 – Present

- Research on evaluating AI character, values and behavior, including whether a model's stated character predicts its actual choices in agentic environments.

Founding AI Engineer, SMEY.AI — Jun 2024 – Nov 2026

- Developed interpretable transformer-based DNA sequence models for yeast strain selection, owning the work from GPU infrastructure to model training and analytics; cut R&D cycles from 2 years to 30 days and guided gene edits that increased oil production by 30%.

AI Safety Research Fellow, SPAR — Sep 2025 – Dec 2025

- Studied whether longer reasoning entrenches LLM beliefs, measured with Martingale Score in test-time inverse scaling settings; built the evaluation pipeline and found inverse scaling was task conditional and sensitive to belief elicitation. Mentored by Tianyi Alex Qiu and Zhonghao He.

AI Safety Research Fellow, MARS — Jul 2025 – Sep 2025

- Conceived and built from scratch an interactive AI Values demo for CivAI, comparing human and LLM positions on contested questions; reached 20k+ page views in the first 3 months and the top-5 on Hacker News.

Machine Learning Engineer, Epoch8 — Nov 2022 – Jul 2023

- Parallelized the open source datapipe engine with Ray (5x throughput) for multi-threaded ML data processing graph execution, built a Rasa chatbot MVP that cut support response time by 60%, and engineered a search and recommendation system that boosted engagement by 14% at Looky.

Machine Learning Engineer, Somin.ai — Nov 2021 – Jul 2022

EDUCATION

FAEB, Technical AI Safety Bootcamp based on ARENA curriculum. LessWrong post. — Jun 2025 – Jul 2025

PhD in Computer Science, University of Helsinki, Finland; Supervisor: Prof. Giulio Jacucci — 2024 – Present

MSc in Technology, LUT University, Finland; with honors, GPA 3.79/4.0; Supervisor: Prof. Leonid Chechurin — 2023 – 2024

BSc in Computer Science, ITMO University, Russia; with honors, GPA 3.90/4.0 — 2019 – 2023

SKILLS

AI safety and evaluations: LLM evaluations, benchmark design, capability and safety evaluations, agent evaluations, evaluation harnesses, automated graders, LLM judges, model behavior, reasoning reliability, calibration and uncertainty, adversarial robustness, red teaming, jailbreak and prompt injection defense, interpretability; Inspect AI, TransformerLens.

ML and research engineering: Python, PyTorch, Hugging Face, LLMs, transformers, NLP, agents and tool use, RL, RLHF, synthetic data, fine tuning, post training, model training and inference, experimental design, statistics; distributed systems and data pipelines, Git, Linux, SQL, Bash, C++, Docker, Kubernetes, Ray, AWS, CI/CD, W&B, MLflow, CUDA, JAX.