Samuel Nellessen AI Safety Researcher

+49 1522 8982798
samuelgerrit.nellessen@gmail.com
snellessen.com
LinkedIn

Education

Radboud University Nijmegen, Netherlands
B.Sc. in Artificial Intelligence (Honours Programme) 2023 - 2026 *(expected)*

- Current GPA: 4.00/4.00 (Dutch system: 8.9/10). Top of class.

Neuromatch Academy Remote / Summer School
Computational Neuroscience & Deep Learning Jul 2023

- Completed intensive 3-week curriculum on biologically plausible learning rules and high-dimensional data analysis.

Otto-Von-Guericke-University / Stendal University Germany
B.A. Philosophy, Neuroscience & B.Sc. Psychology (Foundational Studies) 2020 - 2023

- Consistently ranked in top percentile (GPA ≈ 3.9/4.0). Transferred to Radboud to specialize in technical AI.

Experience

Student Researcher | KachmanLab @ Radboud University Sep 2025 - Present

- Developing novel jailbreaking techniques and adversarial attacks on LLMs.
- Orchestrating multi-GPU distributed training (Slurm/HPC) for Group Relative Policy Optimization (GRPO) to test alignment boundaries.

ARENA 5.0 Fellow | Alignment Research Engineer Accelerator Apr 2025 - Jun 2025

- Selected for competitive fellowship with 4% acceptance rate, training alongside PhDs and industry professionals.
- Implemented Transformers from scratch and conducted deep dives into RLHF and interpretability.

Research Assistant & Foresight Fellow | Donders Institute / Foresight Institute Feb 2024 - Present

- Architected a scalable model-fitting pipeline on Slurm HPC clusters to quantify controllability, translating theoretical frameworks into executable code.
- Selected as **Foresight Fellow** to lead computational modeling of agency, independently developing algorithms to extract latent parameters from behavioral data.

AI Safety Grantee & Scholar | Long-Term Future Fund / AIMM Jul 2022 - Sep 2023

- Awarded competitive grant (<5% acceptance) and mentorship to execute independent research under Jan Hendrik Kirchner (OpenAI/Anthropic).

Skills

AI Safety & Research: LLM Jailbreaking, Adversarial Attacks (GRPO), Red Teaming, Representation Engineering, Mechanistic Interpretability.

Engineering: LLM Fine-Tuning (RLHF, Verifiers), Multi-GPU Distributed Training, Model Optimization.

Stack: Python (PyTorch, Pandas, Transformers, AgentDojo), Slurm/HPC, JavaScript (React), MATLAB, Stan, Java, Scala.

Projects

ARENA 5.0 Capstone: Internal Representations in SONAR Autoencoders 2025
Executed intensive 5-day research sprint in Mechanistic Interpretability. Analyzed model representations to identify features correlated with correctness.

AI Safety Camp 2025: Reasoning Capabilities of LLMs 2025
Mechanistic Interpretability research in an international team. Aiming for publication at NeurIPS/ICLM workshops. Paper available on ArXiV.

Leadership & Awards

- **Technical Advisor**, Foresight AI Safety Grant Program 2024
- **User Experience Lead**, Fridays for Future International 2019-2020
- **Top of Year (Philosophy)**, A-levels 2019