+49 1522 8982798
samuelgerrit.nellessen@gmail.com
Website & Blog
LinkedIn

Samuel Nellessen

Education

Faculty of Social Sciences, Radboud University — Nijmegen, Netherlands
B.Sc. in Artificial Intelligence — 2023 - 2026 *(expected)*
- Current GPA: 4.00/4.00 (in Dutch system: 8.9/10), Best of Year.
- Honours Programme in Artificial Intelligence

Otto-Von-Guericke-University — Magdeburg, Germany
B.A. in Philosophy, Neuroscience and Cognition (not finished) — 2021 - 2023
- GPA: 3.87/4.00

Magdeburg-Stendal University of Applied Sciences — Stendal, Germany
B.Sc. in Psychology (not finished) — 2020 - 2021
- GPA: 3.95/4.00

Experience

Student Researcher | KachmanLab @ Radboud University — 2025.09 - present
- Conducting research on jailbreaking LLMs

ARENA 5.0 Participant | Alignment Research Engineer Accelerator — 2025.04 - 2025.06
- A 4-5 week in-person ML bootcamp with a focus on AI safety.

Research Assistant Computational Psychiatry | Motivational and Cognitive Control lab, Donders Institute — 2024.02 - present
- Computational Modelling for a decision-making behavioural task in MATLAB.

Neurotech Foresight Fellow | Foresight Institute — 2024 - 2025

AI Safety Scholar | AIMM Programme — 2022.08 - 2023.09
- The main goal of the "AI safety scouts scholars" program is to get promising people into AI safety earlier. The program comes with intense mentorship from experienced "scouts" in the field.

Long-Term Future Fund Grantee | EA Funds — 2022.07 - 2023.01
- Funding for self-studying ML and researching the possible applications of a Neuro/CogScience perspective for AGI Safety.

Knowledge Manager & Team Development Referee, Future Matters Project | Future Matters Project — 2021.02 - 2022.06
- Supported team operations, task management, and knowledge base creation.

Project Manager | Together For Future e.V. — 2019.11 - 2021.05

Skills

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

Machine Learning Engineering: LLM Fine-Tuning (incl. RL-based methods, verifiers), Multi-GPU Distributed Training (PyTorch), Model Optimization & Deployment.

Programming & Libraries: `Python` (PyTorch, Pandas, Transformers, AgentDojo), `JavaScript` (React), `MATLAB`, `Stan`, `Java`, `Scala`.

Languages: German (Native), English (Fluent).

Projects

ARENA 5.0 Capstone: Investigating Internal Representations of Correctness in SONAR Text Autoencoders

Executed an intensive 5-day research sprint in Mechanistic Interpretability, analyzing model representations to identify features correlated with correctness. Authored a technical write-up detailing the project's methodologies and findings.

2025

AI Safety Camp 2025: Understanding the Reasoning Capabilities of LLMs

Conducting Mechanistic Interpretability research in an international team of 4 people, aiming for publication at NeurIPS or ICLM MechInterp workshop. Paper on ArXiV.

2025

Reinforcement Learning Agent Development

Implemented and evaluated a Proximal Policy Optimization (PPO) agent within a custom Gymnasium environment as part of university coursework. Ongoing extension with variants of PPO/other SOTA RL algorithms

Ongoing

Personal data scraping project

Developed a data pipeline using Python (requests, pandas) to scrape, process, and analyze local supermarket product data, implementing custom algorithms to identify optimal food choices based on nutritional value and cost.

2025

Developing 'sAIm'

Fine-tuning a large language model (LLM) on personal text corpus to explore personalized agent capabilities, potential biases, and data privacy considerations.

Ongoing

Competing in Kaggle's BirdCLEF 2024 Challenge

Research Code Competition on Bird Sound Classification

2024

Assessing Artificial Sentience: Are we responsible?

Essay series about artificial sentience

2023

How do mice vary in the amount of past information they use to make decisions under uncertainty?

Neuromatch Academy Computational Neuroscience Course

2023

Research Sequence on "Hebbian Natural Abstractions"

Blog Project with Jan Hendrik Kirchner

2022

Grants and Awards

- **Research grant,** AI Safety Mentors & Mentees Programme 2022-2023
- **Research grant,** Long-Term Future Fund 2022-2023
- **Top of the year in Philosophy,** A-levels 2019

Courses and Summer Schools

- **Computational Neuroscience Summer School,** Neuromatch Academy 2023

Volunteering

- **User Experience Lead,** Fridays for Future International 2019-2020
- **Press Officer,** Fridays for Future Düsseldorf 2019

Advising

Advising for: *Foresight AI Safety Grant Program*