Ole Jorgensen

London, United Kingdom
ojorgensen1417@gmail.com
Research Scientist - AISI and Oxford
Website: ojorgensen.github.io
LinkedIn: ole-jorgensen

Professional and Research Experience

Research Scientist, UK AI Security Institute October 2025 - Present
Working with the Science of Evaluations team to develop new methodologies for LLM evaluation.

Research Engineer, UK AI Security Institute Jan 2024 - Oct 2025
Building a world-leading suite to assess the chemical and biological risks arising from frontier language models. Responsible for:

- Leading multiple pre-deployment testing exercises for new models, coordinating a team to produce reports to tight deadlines.
- Eliciting the maximal capabilities of language models, and developing novel evaluations for agentic systems.
- Optimising testing pipelines, halving the number of engineers needed to complete rigorous testing.
- Communicating results and methodologies to AI labs and international partners.

Independent Researcher Oct 2023 - Dec 2023
Continued improving a new method of activation steering developed during my MSc. Reduced the toxicity of language models and improved the effectiveness of function vectors over the previous SOTA. Led to a best paper award at a AAAI 2024 workshop.

Researcher, AI Safety Camp Mar 2023 — June 2023
Designed and implemented a pipeline to investigate LLM behaviour under ambiguity. This work was published at BlackboxNLP 2023.

Education

University of Oxford, DPhil in Statistics Oct 2025 - present

- Working to develop new methods for LLM evaluation, supervised by Tom Rainforth.
- Funded by a scholarship from the Laboratory for AI Security Research (LASR).

Imperial College London, MSc in Artificial Intelligence 2022 - 2023

- Graduated with a **distinction** overall. My module average was 85%, and my dissertation received a grade of 77%.
- My thesis, supervised by Prof. Murray Shanahan, developed new methods for extracting feature representations from language models, enabling more general activation steering than previous SOTA methods.

University of St Andrews, MMath (Hons) Mathematics 2018 – 2022

- Graduated with a **first class** degree. Awarded the Principal's Scholarship for Academic Excellence, granted to the **top 50 students among approximately 2,000** in the Class of 2022.
- My thesis, supervised by Dr. Thomas Coleman, successfully generalised several results regarding homogeneous graphs in the context of homogeneous k-hypergraphs. It was **highly commended** by assessors.

Ole Jorgensen

London, United Kingdom
ojorgensen1417@gmail.com
Research Scientist - AISI and Oxford
Website: ojorgensen.github.io
LinkedIn: ole-jorgensen

Publications and Public Communication

Friederike Grosse-Holz and **Ole Jorgensen**, "Early Insights from Developing Question-Answer Evaluations for Frontier AI" 2024

Ole Jorgensen*, Dylan Cope, Nandi Schoots, Murray Shanahan. "Improving Activation Steering in Language Models with Mean-Centring", arXiv:2312.03813 [cs.CL]. **Best paper** at the *Human-Centric Representation Learning Workshop @ AAAI* 2024

Henning Bartsch*, **Ole Jorgensen***, Domenic Rosatti*, Jason Hoelscher-Obermaier and Jacob Pfau. "Self-Consistency of Large Language Models under Ambiguity", arXiv:2310.13439 [cs.CL]. Accepted at *BlackboxNLP @ EMNLP* 2023

Invited Talks

London Initiative for Safe AI, "Evaluating LLMs as Scientific Assistants" 2025

Responsible Language Models Workshop @ AAAI 2024, "Improving Activation Steering in Language Models with Mean-Centring" 2024

Imperial College AI Safety Reading Group, "Finding and Using Features in Language Models" 2023

Awards and Achievements

MSc Group Project Prize, Imperial College London 2023
Awarded to the best MSc Artificial Intelligence software engineering group project.

The Principal's Scholarship for Academic Excellence, University of St Andrews 2022
Award presented to each of the top 50 academically performing students in their final year of study at the University of St Andrews.

Senior Honours Project Commendation, University of St Andrews 2022
My dissertation "Generalising Homogeneous Structures" was highly commended by assessors.

Laidlaw Scholarship, University of St Andrews 2020 – 2021
Received the Laidlaw Research and Leadership Scholarship, a competitive programme from the Laidlaw Foundation.

Duncan Prize (Applied Mathematics), University of St Andrews 2019
Awarded to the best student of Applied Mathematics at Second Level.

Voluntary and Academic Services

Reviewer 2024-2023
Served as a reviewer for the NeurIPS SoLaR workshop in 2024 and 2023; and at the ICML TiFA workshop in 2024.

Open-source software 2024-2023
Contributor to Inspect, a framework developed by AISI for language model evaluations with over 1000 stars on Github.
Contributed a new environment to Jumanji, a Reinforcement Learning framework written in JAX with over 600 stars on Github.

Co-President, Effective Altruism St Andrews May 2020 — May 2022
Coordinated a committee to run successful Speaker Events, Career Workshops, and Discussion Groups.
Designed and ran an introductory mentorship programme for new members, more than doubling applications year on year.