

Ole Jorgensen

London, United Kingdom
ojorgensen1417@gmail.com

Technical Staff - UK AISI

Website: ojorgensen.github.io
LinkedIn: [ole-jorgensen](#)

PROFESSIONAL AND RESEARCH EXPERIENCE

Research Engineer, UK AI Security Institute January 2024 - Present
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 novel models, coordinating the 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 to a broad range of partners including Nobel Prize winners.

Independent Researcher October 2023 - December 2023
Improved on the 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 formed the basis of the publication at BlackboxNLP 2023.

Researcher, University of St Andrews Jul 2021 — Sep 2021
Undertook original research in Pure Mathematics alongside Prof. Colin Bleak, wherein we investigated rotation and conjugacy in Thompson's Group T. We were able to identify novel conditions for describing rotation as a commutator.

EDUCATION

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.

PUBLICATIONS AND PUBLIC COMMUNICATION

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

UK AISI Technical Staff (Contributing Author), "Advanced AI evaluations at AISI: May update" 2024

Ole Jorgensen

London, United Kingdom
ojorgensen1417@gmail.com

Technical Staff - UK AISI

Website: ojorgensen.github.io
LinkedIn: [ole-jorgensen](#)

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.

Dean's List, University of St Andrews 2019, 2020, 2021, 2022
Awarded for obtaining an average of 16.5 or higher (a 1st) for the academic year.

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.

Club Captain, St Andrews Shinty Club Mar 2020 — Mar 2022
Coordinated 3x training sessions a week throughout the academic year, managed the team during games and tournaments. Was awarded Colours by the university for my services to the club.