# Ole Jorgensen

London, United Kingdom
ojorgensen1417@gmail.com
Website: ojorgensen.github.io
LinkedIn: ole-jorgensen

## 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, *Understanding and Controlling the Activations of Language Models*, was supervised by Prof. Murray Shanahan, Nandi Schoots and Dylan Cope. I presented novel investigations into the representation of features in language models, before developing new methods for extracting feature representations using datasets of text. This enabled new methods for engineering the activations of language models that are more general than previous methods.

**University of St Andrews**, MMath (Hons) Mathematics 2018 – 2022
Graduated with a **first class** degree. Awarded Principal's Scholarship for Academic Excellence as one of the top 50 of approximately 2000 students in their final year of study at the University, with a module average of 19.1/20.
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

**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

**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.
Funded my work with Prof. Peter Cameron and Open Cages.

**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.

# Ole Jorgensen

London, United Kingdom
ojorgensen1417@gmail.com
Website: ojorgensen.github.io
LinkedIn: ole-jorgensen

## Professional and Research Experience

**Research Engineer**, UK AI Safety Institute January 2024 - July 2024
Worked to develop a comprehensive suite of model evaluations as part of the UK Government's efforts to evaluate the bio-chem capabilities of the next generation of frontier language models. Joined on a fixed term contract.

**Independent Researcher** October 2023 - December 2023
Continuing research on steering large language models alongside Dylan Cope, Nandi Schoots and Murray Shanahan. We applied our new method of activation steering in a range of contexts to evaluate its effectiveness, successfully reducing the toxicity of language models and improving the effectiveness of function vectors. This work formed the basis of the publication at the Human-Centric Representation Learning Workshop 2024. Funded by a grant from Open Philanthropy.

**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.

**Assistant to CEO**, Open Cages May 2021 — Jul 2021
Conducted independent research on the international broiler market, informing the charity's big-picture strategy on which markets to target in order to maximise impact. Directly collaborated with the CEO to run a successful fundraising campaign, responsible for communications to tens of thousands of supporters.

**Researcher**, University of St Andrews Jun 2020 — Aug 2020
Completed original research in Pure Mathematics on the Rado Complex under the supervision of Prof. Peter Cameron, with funding from the Laidlaw Foundation. Explanatory paper and poster published on the Laidlaw Scholar's Network.

## Voluntary and Academic Services

**Reviewer**, SoLaR Workshop (NeurIPS) 2023
Served as a reviewer for the Socially Responsible Language Modelling Research Workshop at NeurIPS 2023.

**Open-source software** 2023
Contributed a new environment to Jumanji, a RL framework written in JAX with over $400$ 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 (SPEAK), 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. Adapted all sessions to adhere to COVID-19 restrictions. Was awarded Colours by the university for my services to the club.