Aengus Lynch

aenguslynch@gmail.com — @aengus_lynch1 — Google Scholar — GitHub

Research Interests

AI Safety, Agentic LLMs, Adversarial Robustness, Mechanistic Interpretability

Education

PhD in Artificial Intelligence 2021–Present
University College London
Advisor: Prof. Ricardo Silva

MSci Mathematics, First Class Honours 2017–2021
University College London
Master's Project: Computations and Analysis on Fluid Flow Through a Flexible Channel

External Research Programs

MATS Scholar Jan 2024–Mar 2024
LLM unlearning and adversarial robustness

REMIX program, Redwood Research Jan 2023
Mechanistic interpretability research

Publications

2024: Best-of-N Jailbreaking
John Hughes*, Sara Price*, **Aengus Lynch***, Rylan Schaeffer, Fazl Barez, Sanmi Koyejo, Henry Sleight, Erik Jones, Ethan Perez, Mrinank Sharma
arXiv:2412.03556

Latent Adversarial Training Improves Robustness to Persistent Harmful Behaviors in LLMs
Abhay Sheshadri*, Aidan Ewart*, Phillip Guo*, **Aengus Lynch***, Cindy Wu*, Vivek Hebbar*, Henry Sleight, Asa Cooper Stickland, Ethan Perez, Dylan Hadfield-Menell, Stephen Casper
arXiv:2407.15549

Analyzing the generalization and reliability of steering vectors
Daniel Tan, David Chanin, **Aengus Lynch**, Adrià Garriga-Alonso, Brooks Paige, Dimitrios Kanoulas, Robert Kirk
NeurIPS 2024
arXiv:2407.12404

Eight methods to evaluate robust unlearning in LLMs
Aengus Lynch*, Phillip Guo*, Aidan Ewart*, Stephen Casper, Dylan Hadfield-Menell
arXiv:2402.16835

2023: Towards automated circuit discovery for mechanistic interpretability
Arthur Conmy, Augustine N. Mavor-Parker, **Aengus Lynch**, Stefan Heimersheim, Adrià Garriga-Alonso
NeurIPS 2023 (Spotlight)
arXiv:2304.14997

Spawrious: A benchmark for fine control of spurious correlation biases
Aengus Lynch*, Gbètondji J-S Dovonon*, Jean Kaddour*, Ricardo Silva
arXiv:2303.05470

2022: Causal machine learning: A survey and open problems
Jean Kaddour*, **Aengus Lynch***, Qi Liu, Matt J. Kusner, Ricardo Silva
arXiv:2206.15475

INDUSTRY EXPERIENCE

AI Consultant 2023–2024
Focal Therapy Clinic
Building clinical chatbot systems

AI Consultant Jan 2022–Apr 2022
ChangeBlock
Advised on data science strategy

Rates Trading Summer Analyst Jul 2020–Aug 2020
JP Morgan, London
UK rates trading desk for index-linked gilts

TEACHING & SERVICE

Maths Tutor 2017–2022

Sponsorship Director, UCL AI Society 2018–2019