Aengus Lynch, Ph.D

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

EDUCATION

PhD in Artificial Intelligence 2021–2025
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

EXPERIENCE

Anthropic Fellow Jan 2026–present
Building alignment evaluations that identify compelling and concerning examples of misalignment in Claude Opus 4.5. The better we can probe a model's alignment, the faster we can build trust in deploying autonomous remote workers inside and outside of the labs.

Founder and CEO February 2025–December 2025
Watertight AI
Founded an AI safety organization to measure and prevent agentic misalignment through agentic evals and monitoring.

- Executed a technical research contract with Anthropic to develop datasets for 8 weeks.
- Secured 2 term sheets at a $15M valuation cap and $500k in angel commitments after 2 weeks fundraising.
- Ultimately decided to decline funding and return capital after I decided that it was more important to measure and prevent agentic misalignment within the AI labs than as a startup.

Anthropic contractor August 2024–April 2025
LLM jailbreaking and agentic evaluations

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

REMIX program, Redwood Research Jan 2023
Mechanistic interpretability research

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

PUBLICATIONS

2025: Agentic Misalignment: How LLMs Could be Insider Threats
Aengus Lynch, Benjamin Wright, Caleb Larson, Stuart J. Ritchie, Sören Mindermann, Evan Hubinger, Ethan Perez, Kevin K. Troy
arxiv:2510.05179

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