Harry Mayne

www.harrymayne.com

harrymayne@gmail.com LinkedIn GitHub +44(0)7425 889951 @HarryMayne5

I am an Astra Fellow working with Owain Evans (Truthful AI) and a PhD researcher at the University of Oxford with Adam Mahdi at OxRML and Jakob Foerster at FLAIR. My research evaluates LLM *self-explanations*: natural language explanations LLMs give to justify their own decision-making. I measure whether these explanations are faithful to models' true internal reasoning and I develop new training incentives to improve self-explanation faithfulness. I'm motivated by AI safety. My work has been published at leading venues including NeurIPS (including one Oral), ICLR, and EMNLP. Further research details can be found on my website.

Education

University of Oxford

DPhil LLM Explainability and Interpretability *Oct. 2023 – Present*

- Supervised by Prof. Adam Mahdi (Oxford Internet Institute) and Prof. Jakob Foerster (Engineering).
- Thesis on LLM explainability and interpretability, focusing on *self-explanations*.
- Fully funded by the Grand Union DTP (Economic and Social Research Council).
- Research agenda sponsored by the Dieter Schwarz Foundation.

University of Oxford **Distinction (77%)**

MSc Social Data Science *Oct. 2022 – Aug. 2023*

- Ranked 1st in cohort for overall exam performance and achieved the highest thesis mark (88%).
- Courses included Applied Machine Learning, Natural Language Processing, Data Analytics at Scale, Frontiers of Data Science and Applied Analytical Statistics.
- Fully funded by the Grand Union DTP (Economic and Social Research Council).
- **Thesis:** Unsupervised Learning Approaches to Intensive Care Reform: Opportunities and Challenges.

University of Cambridge **Double First Class**

BA Economics *Oct. 2019 – Jun. 2022*

- Ranked 1st/155 and 2nd/155 in econometrics and microeconomics examinations, respectively.
- Awarded the Patrick Cross Prize, two Tripos Prizes and the Corfield Scholarship for academic excellence.
- Courses included Time Series Econometrics, Microeconometrics, Labour Economics and Industrial Economics.
- **Thesis:** A Feast in the Time of Plague: A Theoretical Model of the UK Housing Market During COVID-19.

A Levels and GCSEs *Sep. 2011 – Jul. 2019*

- 4 A*s at A Level and 11 A*s (or equivalent) at GCSE.

Selected Publications

2026 **A Positive Case for Faithfulness: LLM Self-Explanations Help Predict Model Behavior.** *Under Review at ICML 2026*
Mayne, H.*, Kang, J.*, Gould, D., Ramchandran, K., Mahdi, A., Siegel, N.

LINGOLY-TOO: Disentangling Reasoning from Knowledge with Templatised Orthographic Obfuscation. *ICLR 2026*
Khouja, J., Korgul, K., Hellsten, S., Yang, L., Neacsu, V., **Mayne, H.**, Kearns, R., Bean, A., Mahdi, A.

2025 🔗 **LLMs Don't Know Their Own Decision Boundaries: The Unreliability of Self-Generated Counterfactual Explanations.** *EMNLP 2025*
Mayne, H., Kearns, R. O., Yang, Y., Delaney, E., Russell, C., Mahdi, A.

🔗 **Ablation is Not Enough to Emulate DPO: A Mechanistic Analysis of Toxicity Reduction.** *EMNLP 2025*
Yang, Y., Sondej, F., **Mayne, H.**, Lee, A., Mahdi, A.

🔗 **Measuring What Matters: Construct Validity in Large Language Model Benchmarks.** *NeurIPS 2025*
Bean, A., et al. (incl. **Mayne, H.**)

2024 🔗 **LingOly: A Benchmark of Olympiad-Level Linguistic Reasoning Puzzles in Low-Resource and Extinct Languages.** *NeurIPS 2024, Oral, Top 0.5%*
Bean, A., Hellsten, S., **Mayne, H.**, Magomere, J., Chi, E., Chi, R., Hale, S., Kirk, H.

🔗 **Can Sparse Autoencoders be used to Decompose and Interpret Steering Vectors?** *Interpretable AI, NeurIPS 2024*
Mayne, H., Yang, Y., Mahdi, A.

Positions and Programs

Astra Fellowship: Owain Evans, Truthful AI *Jan. 2026 – Present*

- Understanding how LLMs generalise during finetuning.
- Based at Constellation, Berkeley.

SPAR: Noah Siegel, Google DeepMind *Sept. 2025 – Present*

- Developing a new explanatory faithfulness metric based on counterfactual simulatability.
- First-author publication under review at ICML 2026.

International Growth Centre (IGC) *Jun. 2025 – Present*

- AI Engineer (Jun. 2025 – Dec. 2025). Advisor (Jan. 2026 – Present).
- Using AI to aid public service delivery. See my IGC page.

Bank of England *Summer 2020 & 2021*

- Internships in Insurance Supervision (PRA). Worked in *US Firms* and *Actuarial Science* teams.

Teaching

University of Oxford, Oxford Internet Institute *Sep. 2023 – Dec. 2024*

- Teaching Assistant for Applied Analytical Statistics (Social Data Science MSc).
- Taught over 50 Master's and PhD students. Voted the department's top TA of the year in 2024.

Stanford University, Stanford House *Nov. 2023 – Dec. 2025*

- Tutor Stanford University computer science undergraduate students completing semesters abroad at the University of Oxford.
- Designed and taught 8-week personalised courses in the students' areas of interest (focusing on machine learning, AI and data science).

Additional Skills and Interests

Programming Languages: Python (PyTorch, vLLM, verl, Sklearn, Matplotlib, etc), LaTeX.
Conference Reviewing: Workshops at ICML 2024, NeurIPS 2024 & 2025. ICLR 2026.
Soft Skills: Public speaking (Grade 8, Distinction), organisation, leadership, presentation.
Leadership Roles: Reasoning with Machines Lab Chair (Sept. 2025 – Dec. 2025).
Other Interests: Travel, surfing, and music.