Harry Mayne

www.harrymayne.com

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

I am a PhD candidate at the University of Oxford. My research evaluates the reliability of LLM self-explanations: natural language explanations LLMs give to justify their own decision-making. I measure the extent to which these explanations are faithful to the real internal reasoning and I develop new training incentives to improve self-explanation faithfulness. I'm motivated by explainability and AI safety. Further research details can be found on my website. I've also produced the LingOly (NeurIPS 2024 Oral) and LingOly-TOO benchmarks, as well as forthcoming science of evals papers.

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.
- Fully funded by the Grand Union DTP (Economic and Social Research Council).

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

2025 **The Validity–Minimality Trade-Off: The Limits 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. — *Under review at NeurIPS 2025*
Bean, A., et al. (incl. **Mayne, H.**)

LINGOLY-TOO: Disentangling Reasoning from Knowledge with Templatised Orthographic Obfuscation. *Under review at NeurIPS 2025*
Khouja, J., Korgul, K., Hellsten, S., Yang, L., Neacsu, V., **Mayne, H.**, Kearns, R., Bean, A., Mahdi, A.

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.

Ablation is Not Enough to Emulate DPO: A Mechanistic Analysis of Toxicity Reduction. *Interpretable AI, NeurIPS 2024*
Yang, Y., Sondej, F., **Mayne, H.**, Mahdi, A.

Large language models can help boost food production, but be mindful of their risks. *Frontiers in Artificial Intelligence*
De Clercq, D., Nehring, E., **Mayne, H.**, Mahdi, A.

Teaching

University of Oxford, Oxford Internet Institute *Sep. 2023 – Present*

- 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.
- Designed a new seminar syllabus to improve student understanding.

Stanford University, Stanford House *Nov. 2023 – Present*

- Employed by Stanford University to tutor 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).

King Saud University, Oxmedica/Mawhiba *Jun. 2024*

- Designed and delivered a three-week course in introductory machine learning to 20 talented Saudi teenagers. The full course is available on my website. Received excellent feedback.

Positions

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

- AI consultant in the Zambian Evidence Lab. Using AI to aid public service delivery. See my IGC page.

Bank of England *Summer 2020 & 2021*

- **2020:** First-year intern in Insurance Supervision. Selected as one of 9 interns from a pool of 3,000 applicants; offered a place on the penultimate-year internship.
- **2021:** Penultimate-year intern in the General Insurance Actuarial team (PRA). Offered a place on the actuarial graduate programme.

Additional Skills and Interests

Programming Languages: Python (PyTorch, NumPy, Matplotlib, vLLM, Sklearn...etc), LaTeX.
Conference Reviewing: ICML 2024 DCML Workshop, NeurIPS 2024 MINT Workshop
Soft Skills: Public speaking (Grade 8, Distinction), organisation, leadership, presentation.
Other Leadership Roles: Selwyn College Boat Club Captain (2022-2023), Selwyn College JCR Entertainment Officer (2021-2022), Selwyn College JCR Freshers' Representative (2020-2021).
Other Interests: Travel, surfing, and music.