Andrew Bean

andrew.bean@oii.ox.ac.uk | am-bean.github.io | andrew-m-bean

I am a doctoral student at the University of Oxford researching large language models, with a particular focus on evaluation of LLMs for abilities such as reasoning and collaboration with humans. At NeurIPS 2024, my papers were the **two highest reviewed** in the Datasets and Benchmarks track, including the **Best Paper**, and both were accepted for **Oral** presentations.

Education

University of Oxford — Oxford, UK
DPhil in Social Data Science — Oct 2022 - Sept 2025

Advisors: Dr Adam Mahdi, Dr Luc Rocher

University of Oxford — Oxford, UK
MSc in Social Data Science with Distinction — Oct 2021 - Sept 2022

Thesis Grade: Distinction / Cumulative Mark: Distinction

MSc Thesis: *Teaching AI to Cooperate with Human Partners through Rule-based Play*

Yale University — New Haven, CT
B.S. in Applied Mathematics *cum laude* — Sept 2013 - May 2017

Major GPA: 3.94/Overall GPA: 3.87 with Distinction in Applied Mathematics

Thesis: *nullSolve: An iterative kernel-based method for solving directed Laplacian systems*

Selected Publications

(Full publications on Google Scholar)

NeurIPS '24 Oral
LINGOLY: A Benchmark of Olympiad-Level Linguistic Reasoning Puzzles in Low-Resource and Extinct Languages.
Bean, A.M., Hellsten, S., Mayne, H., Magomere, J., Chi, E.A., Chi, R., Hale, S.A., Kirk, H.R.

NeurIPS '24 Best Paper
The PRISM Alignment Project: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models.
Kirk, H.R., Whitefield, A., Röttger, P., **Bean, A.M.**, Margatina, K., Ciro, J., Mosquera, R., Bartolo, M., Williams, A., He, H., Vidgen, B., and Hale, S.A.

NeurIPS '24 Workshop
Do Large Language Models have Shared Weaknesses in Medical Question Answering?
Bean, A.M., Korgul, K., Krones, F., McCraith, R., and Mahdi, A.

EMNLP '23
The Past, Present and Better Future of Feedback Learning in Large Language Models for Subjective Human Preferences and Values
Kirk, H.R., **Bean, A.M.**, Vidgen, B., Röttger, P., and Hale, S.A.

Nature Medicine Under Review
Clinical knowledge in LLMs does not translate to human interactions
Bean, A.M., Payne, R., Parsons, G., Kirk, H.R., Ciro, J., Mosquera, R., Monsalve, S.H., Ekanayaka, A.S., Tarassenko, L., Rocher, L., and Mahdi, A.

Industry Experience

Five years of professional experience in data science and research science roles, starting in quantitative finance at the largest hedge fund in the world, and currently developing LLMs for legal applications.

Thomson Reuters May 2025
Research Scientist Intern, Foundational Research London, UK

- Trained and benchmarked a state-of-the-art legal LLM
- Developed a new benchmark for agentic AI research systems
- Improved benchmark evaluation efficiency via technical innovations in sub-sampling techniques

HealthSherpa Mar 2021 – Oct 2021
Data Science Consultant Sacramento, CA

- Standardized data best-practices through templatisation and authoring of a data dictionary
- Renovated data pipelines by replacing manual data ingestion, cleaning, and matching with reusable scripts

Bridgewater Associates June 2016 – Jan 2021
Investment Associate Westport, CT

- Modelled and projected return scenarios to support risk allocation decisions for a $100bn flagship portfolio
- Developed proprietary algorithms to trade financial markets from quantitative understanding of market drivers
- Managed, trained, and assessed new hires and interns to become effective individual contributors
- Co-created and owned the product vision for an in-house visualisation platform improving diagnostics and accelerating exploratory research

Presentations/Invited Talks

Clinical knowledge in LLMs does not translate to human interactions 2025
UK Ofcom Tech Policy Team

LingOly: A Benchmark of Olympiad-Level Linguistic Reasoning Puzzles 2024
Meta Open Innovation AI Research Community Annual Research Workshop

Human-Centric Evaluation of LLMs 2024
MIT Media Lab - Center for Constructive Communication

HELP-Med: Human-Evaluation of LLM Partnership in Medical Self-Diagnosis 2024
Heidelberg University Faculty of Medicine

The Past, Present and Better Future of Feedback Learning in Large Language Models 2023
Oxford Department of Computer Science

Teaching Experience

Lecturer: Practical Ethics in Artificial Intelligence Hilary Term 2024
Stanford University (Bing Overseas Study Program)

Tutorial Instructor: Introduction to Statistics for Political Data Science Trinity Term 2023
Stanford University (Bing Overseas Study Program)

Teaching Assistant: Computational Methods for the Social Sciences Hilary Term 2023
University of Oxford

Teaching Assistant: Applied Analytical Statistics Michaelmas Term 2022
University of Oxford