

Joe Benton

EDUCATION

- D.Phil. in Statistics** 2021 – 2023
Department of Statistics, University of Oxford
Thesis title: *Generative Modelling: Theory and Applications*, supervised by Arnaud Doucet and George Deligiannidis
- B.A. with M.Math. in Mathematics** 2017 – 2021
Trinity College, University of Cambridge
Graduated with Distinction (6th out of c. 250 in year)
Thesis title: *Activated Random Walks*, supervised by Perla Sousi

PROFESSIONAL EXPERIENCE

- Member of Technical Staff** | *Anthropic* November 2023 – Present
Currently thinking about how to make safety cases for transformative AI systems. Lead contributor to design and production of ASL-4 misalignment evaluations. Previously worked on white-box adversarial testing of LLMs.
- Technical Staff** | *UK Frontier AI Taskforce* Fall 2023
Part of the founding technical team of the UK Frontier AI Taskforce (now UK AI Safety Institute). Produced technical demonstrations of AI risk for presentation in front of world leaders at the first AI Safety Summit.
- Machine Learning Researcher** | *Redwood Research* Winter 2022 – 2023
Developed and studied causality and fine-tuning based interpretability methods, with applications to mechanistic anomaly detection. Supervised an internship project aiming to automate interpretability techniques.
- Research Intern** | *Alignment Research Center* Spring 2023
Worked on formalizing heuristic arguments and finding efficient heuristic estimators for sparse covariance propagation.
- Research Assistant** | *Center for Human-Compatible Artificial Intelligence, UC Berkeley* Summer 2021
Extended PAIRED algorithm for unsupervised environment design to incorporate human feedback to speeding up and simplify model training. Supervised by Michael Dennis.

VOLUNTEERING

- Research Mentor** | *Supervised Program for Alignment Research* Spring – Fall 2022
Mentored a student research project on decoding sparse feature representations for neural network interpretability.
- Trustee, Cofounder** | *Raise: A Celebration of Giving* 2018 – Present
Trustee and co-founder of Raise, a student charity initiative raising over 460,000 for the [Against Malaria Foundation](#).
- Strategic Advisor** | *AI Safety Hub* 2022 – 2023
Advised an AI safety outreach and mentoring organization, and managed their [AI Safety Fundamentals](#) program.

AWARDS

- International Mathematics Olympiad (1 Gold – 7th out of 615, 3 Silver)** 2014 – 2017
- International Olympiad in Informatics (1 Gold – 6th out of 304, 1 Silver, 1 Bronze)** 2015 – 2017
- Romanian Masters in Mathematics (3 Gold – Best record of any competitor)** 2015 – 2017

INVITED TALKS

- University of Warwick Algorithms and Computationally Intensive Inference Seminars, 2023
- Royal Statistical Society International Conference, Probabilistic and Statistical Aspects of Machine Learning Discussion Meeting, 2023

SERVICE

- Program Chair, AAAI 2025 Alignment Track
- Reviewed for NeurIPS 2024, JMLR, TMLR, JRSS-B, NeurIPS 2023 SoLaR Workshop, NeurIPS 2023 Workshop on Diffusion Models, ICML 2023 Workshop Frontiers4LCD, Cooperative AI Foundation

PUBLICATIONS

Many-shot Jailbreaking. Cem Anil, Esin Durmus, Mrinank Sharma, **Joe Benton**, Sandipan Kundu, Joshua Batson, Nina Rinsky, Meg Tong, Jesse Mu, Daniel Ford et al. *Advances in Neural Information Processing Systems*, 2024.

When Do Universal Image Jailbreaks Transfer Between Vision-Language Models?. Rylan Schaeffer, Dan Valentine, Luke Bailey, James Chua, Cristóbal Eyzaguirre, Zane Durante, **Joe Benton**, Brando Miranda, Henry Sleight, John Hughes et al. *arXiv preprint arXiv:2407.15211*, 2024.

From Denoising Diffusions to Denoising Markov Models. **Joe Benton**, Yuyang Shi, Valentin De Bortoli, George Deligiannidis, Arnaud Doucet. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(2):286–301, 2024.

Nearly d -Linear Convergence Bounds for Diffusion Models via Stochastic Localization. **Joe Benton**, Arnaud Doucet, George Deligiannidis. *International Conference on Learning Representations*, 2024.

Error Bounds for Flow Matching Methods. **Joe Benton**, George Deligiannidis, Arnaud Doucet. *Transactions on Machine Learning Research*, February 2024.

Alpha-divergence Variational Inference Meets Importance Weighted Auto-Encoders: Methodology and Asymptotics. Kamliia Daudel, **Joe Benton***, Yuyang Shi*, Arnaud Doucet. *Journal of Machine Learning Research*, 24(243):1–83, 2023.

Measuring Feature Sparsity in Language Models. Mingyang Deng, Lucas Tao, **Joe Benton**. *NeurIPS 2023 Workshop on Socially Responsible Language Modelling Research*, 2023.

A Continuous Time Framework for Discrete Denoising Models. Andrew Campbell, **Joe Benton**, Valentin De Bortoli, Tom Rainforth, George Deligiannidis, Arnaud Doucet. *Advances in Neural Information Processing Systems*, 2022.

Polysemanticity and Capacity in Neural Networks. Adam Scherlis, Kshitij Sachan, Adam S. Jermyn, **Joe Benton**, Buck Shlegeris. *arXiv preprint, arXiv:2210.01892*, 2022.