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* Fall 2023 – Present
Developed and implemented adversarial testing methods for state-of-the-art large language models. Contributed to plans for producing postive safety cases and aligning ASL-4 level systems as part of the Alignment Science team.

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
Studied and developed causality and fine-tuning based interpretability methods, with applications to mechanistic anomaly detection. Supervised a month-long intern project aiming to automate interpretability techniques.

Research Intern | *Alignment Research Center* Spring 2023
Worked on formalizing heuristic arguments, finding efficient heurstic estimators for sparse covariance propagation.

Research Assistant | *Center for Human-Compatible Artificial Intelligence, UC Berkeley* Summer 2021
Built on the PAIRED algorithm for unsupervised environment design to incorporate human feedback with the aim of speeding up and simplifying the training process. 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 manged 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 Meeiting, 2023

Reviewing

JMLR, TMLR, JRSS-B, ICML 2023 Workshop Frontiers4LCD, Cooperative AI Foundation

Publications

Many-shot Jailbreaking. Cem Anil, Esin Durmus, Mrinank Sharma, **Joe Benton**, Sandipan Kundu, Joshua Batson, Nina Rimsky, Meg Tong, Jesse Mu, Daniel Ford, Francesco Mosconi et al. *Anthropic Blog*, 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. Kamélia 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.