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
Working on scalable oversight, thinking about how to provide reliable training rewards without human supervision. Previously, I've worked on model organisms of misalignment, chain-of-thought monitoring and control evaluations.

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 heurstic 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 | *ML Alignment & Theory Scholars* Winter 2025
Mentored three students on research projects related to AI control and chain-of-thought faithfulness.

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.

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

Service

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

Publications

Optimizing AI Agent Attacks With Synthetic Data. Chloe Loughridge, Paul Colognese, Avery Griffin, Tyler Tracy, Jon Kutasov, **Joe Benton**. *arXiv preprint arXiv:2511.02823*, 2025.

Evaluating Control Protocols for Untrusted AI Agents. Jon Kutasov, Chloe Loughridge, Yuqi Sun, Henry Sleight, Buck Shlegeris, Tyler Tracy, **Joe Benton**. *arXiv preprint arXiv:2511.02997*, 2025.

SHADE-Arena: Evaluating Sabotage and Monitoring in LLM Agents. Jonathan Kutasov, Yuqi Sun, Paul Colognese, Teun van der Weij, Linda Petrini, Chen Bo Calvin Zhang, John Hughes, Xiang Deng, Henry Sleight, Tyler Tracy, Buck Shlegeris, **Joe Benton**. *arXiv preprint arXiv:2506.15740*, 2025.

Reasoning Models Don't Always Say What They Think. Yanda Chen, **Joe Benton**, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, Ethan Perez. *arXiv preprint arXiv:2505.05410*, 2025.

Inverse Scaling in Test-Time Compute. Aryo Pradipta Gema, Alexander Hgele, Runjin Chen, Andy Arditi, Jacob Goldman-Wetzler, Kit Fraser-Taliente, Henry Sleight, Linda Petrini, Julian Michael, Beatrice Alex, Pasquale Minervini, Yanda Chen, **Joe Benton**, Ethan Perez. *arXiv preprint arXiv:2507.14417*, 2025.

Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety. Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, **Joe Benton**, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, Scott Emmons, Owain Evans, David Farhi, Dan Hendrycks, et al. *arXiv preprint arXiv:2507.11473*, 2025.

Teaching Models to Verbalize Reward Hacking in Chain-of-Thought Reasoning. Miles Turpin, Andy Arditi, Meihua Li, **Joe Benton**, Julian Michael. *arXiv preprint arXiv:2506.22777*, 2025.

Constitutional Classifiers: Defending against Universal Jailbreaks across Thousands of Hours of Red Teaming. Mrinank Sharma, Meg Tong, Jesse Mu, Jerry Wei, Jorrit Kruthoff, Scott Goodfriend, Euan Ong, Alwin Peng, Raj Agarwal, Cem Anil, Amanda Askell, Nathan Bailey, **Joe Benton**, Emma Bluemke, et al. *arXiv preprint arXiv:2501.18837*, 2025.

Sabotage Evaluations for Frontier Models. **Joe Benton**, Misha Wagner, Eric Christiansen, Cem Anil, Ethan Perez, Jai Srivastav, Esin Durmus, Deep Ganguli, Shauna Kravec, Buck Shlegeris et al. *arXiv preprint arXiv:2410.21514*, 2024.

Many-shot Jailbreaking. Cem Anil, Esin Durmus, Mrinank Sharma, **Joe Benton**, Sandipan Kundu, Joshua Batson, Nina Rimsky, 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. *NeurIPS 2024 Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models*, 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. Kamlia 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.