

Akbir Khan

akbir.dev

Education

2021-2024	Ph.D. in Computer Science, University College London — Advised by Tim Rocktäschel & Ed Grefenstette
2017-2018	MPhil in Advanced Computer Science, <i>with distinction</i> , University of Cambridge
2013-2017	MSci in Mathematics with Physics, <i>with 1st class honours</i> , University College London
2015-2016	Exchange student, <i>as Mathematics Specialist</i> , University of Toronto

Experience

2024-	Incoming member of the technical staff at Anthropic , in the alignment science team.
2023-2024	Research analyst at Cooperative AI Foundation , grant making and field building.
2021-2023	Senior applied researcher at Tractable AI , developing OCR pipelines for japanese fax.
2017-2020	Chief Research Officer at Spherical Defence , raised \$2M in seed funding.
2016	Four month internship as a software engineer at Deutsche Bank .
2015	Research assistant at the Quantum Optics and Laser Group , Imperial College London.

Selected Publications

2024	Debating with More Persuasive LLMs Leads to More Truthful Answers — A Khan, J Hughes, D Valentine, L Ruis, K Sachan, A Radhakrishnan, E Grefenstette, S Bowman, T Rocktäschel & E Perez. Best Paper Award at <i>International Conference on Machine Learning (ICML)</i> .
2024	Scaling Opponent Shaping to High Dimensional Games — A Khan, T Willi, N Kwan, A Tachetti, C Lu, T Rocktäschel, E Grefenstette & J Foerstor. Oral at <i>The Autonomous Agents and Multi-Agent Systems (AAMAS)</i> .
2023	The Goldilocks of Pragmatic Understanding: Fine-Tuning Strategy Matters for Implicature Resolution by LLMs — L Ruis, A Khan, S Biderman, S Hooker, T Rocktäschel, & E Grefenstette. Spotlight [top 3%] at <i>The Neural Information Processing Systems (NeurIPS)</i> .
2023	MAESTRO: Open-Ended Environment Design for Multi-Agent Reinforcement Learning — M Samvelyan, A Khan, M Dennis, M Jiang, J Parker-Holder, JN Foerster, R Raileanu, T Rocktäschel. Accepted at <i>International Conference on Learning Representations (ICLR)</i>

Awards

2024	Best Paper Award, ICML
2024	SuperAlignment Fellowship, OpenAI
2023	Astra Fellowship, Redwood Research
2018	1st Place, Cambridge Ternary Hackathon
2014	Finalist, European Debating Championship