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, *with distinction*, University of Cambridge
2013-2017 MSci in Mathematics with Physics, *with 1st class honours*, University College London
2015-2016 Exchange student, *as Mathematics Specialist*, 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, built seq2seq models for anomaly detection.
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 *International Conference on Machine Learning (ICML)*.

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 *The Autonomous Agents and Multi-Agent Systems (AAMAS)*.

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 *The Neural Information Processing Systems (NeurIPS)*.

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 *International Conference on Learning Representations (ICLR)*

Awards

2024 Best Paper Award, International Conference on Machine Learning
2024 SuperAlignment Fellowship, OpenAI
2023 Astra Fellowship, Redwood Research
2020 Foundational Artificial Intelligence Scholarship, ESPRC