# SHARAN MAIYA

sm2783@cam.ac.uk

sharanmaiya.com ◇ github.com/maiush ◇ linkedin.com/in/sharanmaiya

## EDUCATION

**University of Cambridge** *Oct 2023 - 2027 (expected)*
PhD Applied Linguistics
*Addressing Bottlenecks in Climate Decision-Making using AI Assistants*
*Supervised by Anna Korhonen, Ramit Debnath*
AI for science; monitoring/control; character training

MRes Environmental Data Science — Distinction
*Thesis: Aligning Language Model Evaluators with Human Judgement*
*Supervised by Ramit Debnath, Laura Cimoli, Anna Korhonen*

**Imperial College London** *Oct 2020 - Sep 2021*
MSc Statistics — Merit
*Thesis: A Novel Method of Tuning and Comparing Causal Discovery Algorithms on Real Data*
*Supervised by Ioanna Papatsouma, D.K. Arvind*

**The University of Edinburgh** *Sep 2016 - Jun 2020*
BSc Computer Science and Mathematics — First Class
*Thesis: Investigating the Respiratory Rate Response to $PM_{2.5}$ Exposure in Asthmatic Adolescents*
*Supervised by D.K. Arvind*

## WORK + RESEARCH EXPERIENCE

**ML Alignment and Theory Scholars** — Jan 2025 - Mar 2026
*Scholar*

Developed the first open implementation of character training for AI assistants.
- MATS 7.0: advised by Evan Hubinger (Anthropic).
- MATS 7.1: advised by Nathan Lambert (Allen Institute for AI).

**Cadenza Labs** — Aug 2024 - Present
*Co-Founding Research Scientist*
- Non-profit research group focused on studying honesty in AI assistants.
- Built a benchmark for AI lie-detectors—used in follow-up work at both Anthropic and Google DeepMind.

**Supervised Program for Alignment Research** — Oct 2023 - Jun 2024
*Student Researcher*
- Project 1: building evaluations for sycophancy in LLMs - advised by Gabriel Recchia.
- Project 2: developing unsupervised probing methods for latent knowledge in LLMs - advised by Walter Laurito.

**The University of Edinburgh** — Sep 2021 - Jun 2023
*Research Assistant*
- Statistical methods and machine learning for a range of problems in air pollution epidemiology.
- Causal discovery algorithms and causal effect estimation.
- Debiased (targeted) machine learning for semi/non-parametric models.
- Advising undergraduates and masters students on a weekly basis.

**TradingHub** — Jun 2020 - Aug 2020
*Software Engineer Intern*

**DataGrasp** Jan 2020 - Apr 2020
*Freelance Data Scientist*

**Royal Bank of Scotland** Jun 2019 - Aug 2019
*Summer Intern*

**The University of Edinburgh** Sep 2018 - Dec 2018
*Undergraduate Researcher*

## PREPRINTS + PUBLICATIONS

**Sharan Maiya**, Henning Bartsch, Nathan Lambert, and Evan Hubinger. "Open Character Training: Shaping the Persona of AI Assistants through Constitutional AI". *arXiv preprint arXiv:2511.01689*, 2025, https://arxiv.org/abs/2511.01689.

Kieron Kretschmar*, Walter Laurito*, **Sharan Maiya**, and Samuel Marks. "Liars' Bench: Evaluating Lie Detectors for Language Models". *arXiv preprint arXiv:2511.16035*, 2025, https://arxiv.org/abs/2511.16035.

Yu Ying Chiu, Zhilin Wang, **Sharan Maiya**, Yejin Choi, Kyle Fish, Sydney Levine, and Evan Hubinger. "Will AI Tell Lies to Save Sick Children? Litmus-Testing AI Values Prioritization with AIRiskDilemmas". *arXiv preprint arXiv:2505.14633*, 2025, https://arxiv.org/abs/2505.14633.

**Sharan Maiya**, Yinhong Liu, Ramit Debnath, and Anna Korhonen. "Improving Preference Extraction In LLMs By Identifying Latent Knowledge Through Classifying Probes". *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, https://aclanthology.org/2025.acl-long.444/.

Walter Laurito*, **Sharan Maiya***, Grégoire Dhimoïla*, Owen Ho Wan Yeung, and Kaarel Hänni. "Cluster-Norm for Unsupervised Probing of Knowledge". *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, https://aclanthology.org/2024.emnlp-main.780/.