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
Supervised by Anna Korhonen, Ramit Debnath
technical AI alignment: alignment evaluations, eliciting latent knowledge, LLM-as-a-Judge, personas

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

Constitutional AI for character training of AI assistants
- MATS 7.0: advised by Evan Hubinger.
- MATS 7.1: advised by Nathan Lambert.

Cadenza Labs — Aug 2024 - Present
Co-Founding Research Scientist

- Non-profit research group focused on technical AI alignment.
- Building state-of-the-art LLM deception-detection methods and establishing industry-standard evaluations for deception-detection.

Supervised Program for Alignment Research — Oct 2023 - Jun 2024
Student Researcher

- Project 1: investigating sycophancy in LLMs - advised by Gabriel Recchia.
- Project 2: developing unsupervised probing methods for latent knowledge in LLMs - advised by Walter Laurito.

Cambridge AI Safety Hub — Winter 2023 and Winter 2024
Fellowship Facilitator

Teaching / guiding reading groups on literature in technical AI Safety.

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

PUBLICATIONS

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/.