

# LOVKUSH AGARWAL

lovkush.com

## AI and AI Safety Experience

Published Mech Interp paper in NeurIPS

September 2024

- Studied how much information about the next paragraph is stored in ‘\n \n’ token.
- Link to [arxiv pre-print](#).

Project Lead for AI Safety Camp 10

November 2024 – Present

- Leading a project asking how we can adapt Sakana’s AI Scientist to AI Safety.

Teaching Fellow for BlueDot AI Alignment Course

October 2024 – Present

- Facilitating four cohorts. I am by far the most active facilitator on the Slack and have had more people join my cohorts. In particular, each week, I write an [opinionated guide to the readings](#).

Adviser for CEEALAR

September 2024 – Present

- Review applications for their AI Safety Winter program.

Teaching Lead for ML4Good bootcamp

September 2024

- Responsible for the delivery of ML4Good, a 10-day in-person AI Safety bootcamp.
- Significantly improved the camp in various ways: my transformer lecture achieved highest ever student feedback score for any session, improved focus on meta-skills via a daily nudge, creating ML4G newsletter to help improve alumni community, etc.

Writing about AI Safety

- Examples include a distillation of [Do Language Models Plan for Future Tokens](#), notes from a talk by [Singapore AISI](#) and [How I Keep Up With AI Safety](#).

Mentee for SPAR

June 2024 – September 2024

- Working with Nicky Pochinkov. Result was the NeurIPS paper above.

Participant in Apart Hackathons

- This weekend, seeing if we can use Goodfire’s API to recover their proprietary SAE feature vectors.
- Running deception evals using AISI’s Inspect open source framework. Code available on [GitHub](#).

Data Scientist (R&D), Shell

Apr 2021 – July 2024

- R&D for alignment of panel time series data. Includes lit reviews and designing new algorithms and metrics.
- Created a python package that allows geologists to align and measure uncertainty in well log correlations

## Education

PhD in Pure Mathematics, University of Leeds, [Link to thesis](#)

MMath in Mathematics (Distinction), University of Cambridge, [Link to thesis](#)

## Selected Position of Responsibility

President of The Cambridge University Mathematical Society

- Revived a stagnant society, by re-branding and obtaining sponsorship to fund weekly events for members
- Surpassed previous membership figures: increased from 50 per year to 200+ per year