Chloe Li

✉ *chloeli561@gmail.com*

Education

2024–25 **MSc Machine Learning**, *University College London*,
Graduated with Distinction.

2020–23 **BA Hons Natural Sciences (Neuroscience, Psychology, & Behavior)**,
University of Cambridge (Emmanuel College).
Double First Class (ranked 2nd in c. 100 for 2 years).
Thesis (90/100, rank 1st): Efficient Bayesian Observer Models. Supervised by Paul Bays.

Publications

Sep 2025 **Chloe Li**, Mary Phuong, Daniel Tan. "Spilling the Beans: Teaching LLMs to Self-Report Their Hidden Objectives". *In Proceedings of ICLR 2026*.

Jun 2025 **Chloe Li**, Mary Phuong, Noah Y. Siegel. "LLMs Can Covertly Sandbag on Capability Evaluations Against Chain-of-Thought Monitoring." *In Proceedings of IJCNLP-AACL 2025* (Oral). Also accepted as Oral at ICML 2025 Workshop on Technical AI Governance.

Professional Experience

Nov 2025 – Present **Anthropic Fellow**, Anthropic.
Research Fellow doing alignment midtraining research with Sam Marks, Jon Kutasov, and Sara Price.

Fall 2024 **Program Lead and ML Curriculum Designer**, ARENA.
Led a 5-week ML program for technical AI safety in London. Co-wrote an LLM evaluations curriculum, advised by Marius Hobbhahn (CEO, Apollo Research).

2023–2024 **Director**, Cambridge AI Safety Hub.
Led a non-profit student group aimed to reduce risks from advanced AI. Founded MARS, a research program with 150+ students mentored and publications at NeurIPS, ICML, and EMNLP.

Spring 2024 **Research Intern**, Center for AI Safety.
Supervisor: Andy Zou (CTO, Grey Swan). Evaluated robustness of representation engineering and steering methods for model honesty against prompt and gradient-based attacks (GCG).

Summer 2022 **Research Intern**, Harvard University.
Supervisors: Sam Gershman, Elizabeth Spelke. Bayesian models of object shape perception.

Awards

2023 **Herchel-Smith Scholarship (Declined)**, Emmanuel College, Harvard MSc tuition.

2022 **Harvard PRISE Fellowship**, Harvard University, $5000 for 10-week research.

2021, 2022 **Davies Senior Scholarship & College Prize**, University of Cambridge, First Class exam ranking.

Volunteering

Feb 2026 – Present **Research Mentor**, Supervised Program for Alignment Research.
Mentoring two students on a research project related to honesty fine-tuning and model self-incrimination.

Invited Talks

2026 Deep Learning: Classics and Trends (DLCT)