Max Kaufmann

maxkaufmann.com

Education

University of Cambridge, BA in Computer Science *October 2019 - June 2022*

Grade: 2.i. Notable achievements include:

- 88% avg. in Category Theory Module.
- 97% avg. in 1st Year Maths.
- 78% avg. in 3rd year research-focused dissertation.

Carmel RC College, A-level and GCSE qualifications *October 2012 - June 2019*

Ranked 1st in the high school for each subject. Spanish and German qualifications taken in spare time.

Experience

Founding Team UK AI Safety Institute *June 2023 - Present*

Third employee at the world's first AI Safety institute. As part of the founding team worked on headhunting, interviewing, strategy, managing external contractors, general operations and UK public policy. Now leading LLM agent scaffolding. Biosecurity work [4] presented at the inaugural International AI Safety Summit.

Researcher Owain Evans Research Group *Febuary 2023 - June 2023*

Supervised by Owain Evans, studied both the generalisation successes [2] and failures [3] of language models. Project involved both supervised and RL finetuning of LLMs up to 65 billion parameters.

Research Assistant Center for AI Safety *October 2022 - Febuary 2023*

Continued work begain at UC Berkely, worked with Dan Hendrycks to produce two works [5, 6] exploring out-of-distribution generalisation in the context of adversarial robustness.

Research Intern UC Berkeley *July 2022 - October 2022*

Supervised by Dan Hendrycks, worked on testing generalisation of adversarial robustness, and on building evaluations for the cooperative tendencies of LLMs.

ML Engineering Intern, Infosys *June 2021 - October 2021*

Post-hoc interpretability methods, certified adversarial robustness bounds and causal methods for fairness.

Research

[1] Alan Chan, Carson Ezell, **Max Kaufmann**, Kevin Wei, Lewis Hammond, H. Bradley, Emma Bluemke, Nitarshan Rajkumar, David Krueger, Noam Kolt, et al. *Visibility into AI Agents.* (Arxiv). 2024.

[2] L.* Berglund, A. Stickland*, M. Balesni*, **Max Kaufmann***, M. Tong*, T. Korbak, D. Kokotajlo, and Owain Evans. *Taken out of context: On measuring situational awareness in LLMs.* (Arxiv). 2023.

[3] Lukas Berglund, Meg Tong, **Max Kaufmann**, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. *The Reversal Curse: LLMs trained on "A is B" fail to learn "B is A"*. NeurIPS 2023 Workshop on Attributing Model Behavior at Scale. (Arxiv). 2023.

[4] **Max Kaufmann**. *Dual-use biology capabilities across model scale.* Presented at the 2023 International AI Safety Summit. Unpublished due to sensitivity. 2023.

[5] **Max Kaufmann***, Dron Hazra*, and Dan Hendrycks. *MatAttack: Differentiable materials for adversarial attacks [forthcoming].* 2023.

[6] **Max Kaufmann***, Daniel Kang*, Yi Sun*, Steven Basart, Xuwang Yin, Mantas Mazeika, Akul Arora, Adam Dziedzic, Franziska Boenisch, Tom Brown, Jacob Steinhardt, and Dan Hendrycks. *Testing Robustness Against Unforeseen Adversaries.* ICML 2024 Submission (Arxiv). 2023.

[7] **Max Kaufmann**, Yiren Zhao, Ilia Shumailov, Robert Mullins, and Nicolas Papernot. *Efficient Adversarial Training With Data Pruning.* (Arxiv). 2022.