James Mann

✉ james.mann.24@ucl.ac.uk ☎ 07458 394899 in jrhmann broverfitter

Education

University College London **Computer Science BSc** **Sept 2024 – May 2027**

- First Year Representative of the AI Society

Carmel College **A Levels** **Sept 2017 – May 2024**

- A*A*A*AA (Maths, Further Maths, EPQ, Physics, Computer Science)
- Elected Head Student

Experience

Arcadia Impact **Summer Research Internship** **May 2025 -**

- Interning over the summer with an organisation contracted by the UK AI Security Institute to develop the open source Inspect Evals software.
- Responsible for identifying key stakeholders in the AI Evals ecosystem, interviewing them about their work, then aggregating the results into actionable technical directions with technical corroborations.

Impact Research Groups **Technical AI Safety Research Sprint** **Feb 2025 - April 2025**

- First author of a paper produced as part of an 8 week competitive research sprint, coming third.
- Using mechanistic interpretability I identified a harmfulness direction, separate to refusal, that could steer between refusal and non-refusal, introducing the novel concept of a "tipping point" for refusal behaviour.
- Proposed a theoretical framework for the success of jailbreaks as suppressors of harmfulness perception.

Bruker **Machine Learning Experience Week** **March 2023**

- Integrated a Convolutional Neural Network with X-Ray Diffractometry, yielding a 100x speedup in deployment.
- Developed a Streamlit application to interface between the data, the model, and the underlying technology.

Projects / Hackathons

IdeaTracer at ICHack

- Developed IdeaTracer, a Claude powered agentic search engine aiming to build knowledge dependency graphs that trace the network of ideas over time.

DigiCapsule at UCLHack

- Aimed to solve the problem of trustless digital time capsules by creating unparallelisable computational proofs of work that guarantee access is only possible after a set time.

GPT from Scratch

- Created a transformer from scratch using NumPy with modular attention and multi-layer perceptron blocks.
- With no additional dependencies designed an implementation of gradient descent and backpropagation to train the model for next token prediction.

Conditional Diffusion Model for Denoising Images

- Implemented a denoising diffusion probabilistic model in Pytorch, trained on CIFAR100 to generate novel images from the distribution, iteratively optimising the hyperparameters to improve the plausibility of samples.

Interests

Rowing: Began rowing when I came to UCL, I now compete for the university in the novice A boat.

French: Self teaching the language with an interest to study abroad.