James Mann

james.mann.24@ucl.ac.uk 07458 394899 broverfitter.github.io jrhmann broverfitter

Education

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

- First Year Representative of the AI Society
- Relevant Classes: Principles of Programming, Theory of Computation, Algorithms, Software Engineering, Computer Architecture and Concurrency, Logic, Object-Oriented Programming, Security

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 **Impact Research Groups** **Feb 2025 - April 2025**

- First author of a paper produced as part of an 8 week competitive research sprint.
- 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.

ENAIS **AI Safety Collab** **March 2025 - May 2025**

- Accepted to a reading group program following the BlueDot AI Alignment curriculum.
- Actively engaged in discussions on the key concepts in AI Safety and Control.

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

- Integrated a Convolutional Neural Network with X-Ray Diffractometry, yielding a 100x speedup in deployment.
- Independently developed a Streamlit web interface for users to interact with the model and visualise data.

Projects / Hackathons

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.

.

IdeaTracer at ICHack

- Competed in the largest student hackathon in Europe to develop an AI-powered app to traverse chains of knowledge. Leveraged Claude to build up interwoven stories of ideas.

Interests

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

Running: Enjoy running, currently training to run a marathon in the summer.

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