Last updated in January 2026

Jackson Kaunismaa

London, UK   jackson-kaunismaa   JacksonKaunismaa

Education

University of Toronto *Sept 2019 – June 2024*

BASc in Engineering Science, Machine Intelligence

- Major GPA: 4.0/4.0
- Cumulative GPA: 3.7/4.0
- **Coursework:** Artificial Intelligence; Machine Intelligence, Software and Neural Networks; Probabilistic Reasoning; Decision Support Systems; Matrix Algebra; Nonlinear Optimization; Mathematical Programming

Experience

Research Fellow *Berkeley / London*

ML Alignment & Theory Scholars *Jan 2025 – present*

- Developed method to elicit harmful capabilities in open-source models by fine-tuning on frontier model outputs from adjacent, non-refused domains. Recovered ∼40% of the capability gap on hazardous chemical synthesis tasks.
- Investigated whether prompt injection and jailbreaking documents improve frontier model performance on monitor-bypassing sabotage tasks. Found no significant effect, suggesting these documents can be filtered from pretraining data without capability loss.
- Developing realistic scheming evaluations and investigating effects of character training on scheming propensities and auditability. *In progress.*

Research Volunteer *Remote*

Supervised Program for Alignment Research *June 2024 – Feb 2025*

- Extended circuit discovery methods from Marks et al. to automatically identify sparse, interpretable circuits in GPT-2 using SAE features.
- Built infrastructure for testing scalable oversight protocols based on Kenton et al., including agentic tool-use pipelines to create reasoning gaps.

ML Researcher *Toronto, ON*

Gene2Lead *Oct 2023 – Jan 2024*

- Built feature extraction pipeline from RCSB Protein Data Bank; trained XGBoost and CNN models to predict covalent reactivity of amino acid active sites, achieving 93% AUC

Undergraduate Researcher *Toronto, ON*

Vector Institute (Supervised by Sanja Fidler) *May 2023 – Aug 2023*

- Implemented Transformers from scratch with multi-GPU data parallelism; benchmarked position embedding schemes across compute scales

ML Researcher *Toronto, ON*

University of Toronto (Supervised by Michael Guerzhoy) *Nov 2022 – Aug 2023*

- Developed novel saliency method for CNNs on intensity-sensitive classification tasks; published at ICLR 2023

Compiler Developer Intern *Markham, ON*

IBM Canada *May 2022 – May 2023*

- Solved bugs and integrated changes to the CPython interpreter for IBM z/OS; integrated BLAS/LAPACK into z/OS builds of NumPy and SciPy

ML Research Intern *Toronto, ON*

RBC Borealis AI (Supervised by Marcus Brubaker) *July 2019 – Sept 2019*

- Developed novel convolution method for generative normalizing flows, extending emerging convolutions for texture synthesis

Publications

Eliciting Harmful Capabilities by Fine-Tuning on Safeguarded Outputs *2025*

J. Kaunismaa, J. Hughes, C. Q. Knight, A. Griffin, M. Sharma, E. Jones

Under review [PDF]

A Benchmark for Scalable Oversight Mechanisms *2025*

A. Pallavi Sudhir, J. Kaunismaa, A. Panickssery

ICLR 2025 Bi-Align Workshop arXiv:2504.03731

An Investigation into Energy Minimization Properties of MLP Features in LLMs *May 2024*

J. Kaunismaa, V. Papyan

Undergraduate Thesis [PDF]

How do ConvNets Understand Image Intensity? *June 2023*

J. Kaunismaa, M. Guerzhoy

Tiny Papers @ ICLR 2023 arXiv:2306.00360

Projects

Sparse Circuit Discovery *Code* *Writeup*

- Automatic circuit discovery in GPT-2 using SAE features; extended methods from Marks et al. to find sparse, interpretable circuits

CUDA Evolution Simulator *Code*

- High-performance evolution engine using custom CUDA kernels, OpenGL visualization, and PyTorch

Transformer Benchmarking *Code*

- Transformer implementation from scratch with multi-GPU data parallelism; comparison of position embeddings

AlphaZero for Connect Four *Code*

- Implementation of AlphaZero for 8x8 Connect Four with comparison to alpha-beta search

Skills

Languages & Frameworks: Python, PyTorch, Inspect, C/C++, CUDA, Bash, Git, LaTeX

ML: Transformers, mechanistic interpretability, SAEs, evaluations, RL, normalizing flows

References

Erik Jones – Research Scientist, Anthropic

Erik Jenner – Research Scientist, Google DeepMind

Michael Guerzhoy – Assistant Professor, University of Toronto