# Large Neural Networks

Rahul G. Krishnan  Steven Coyne

# Large multi-modal language models

- The last few years have seen an increasing rise in large neural networks such as ChatGPT, GPT4, Bard, Claude etc.
- Training regimen for these models:
  1. Trained to predict words conditioned on previous context on very large corpora of data (see RedPajama),
  2. Fine-tuned via instructions (human annotated examples of questions and responses),
  3. Further fine-tuned via Reinforcement Learning with Human Feedback (treat the output of a model as an action and use humans as the "simulator" to identify which is a good action and which is not)
  4. Further fine-tuning with vision, speech based models (e.g. GPT4Vision)

# Machine learning and the rise of large models

- Last year I said: "In the next 5-10 years, we will see a dramatic rise in the use of large language models and generative models of realistic images. These models are good enough that they produce outputs that are indistinguishable from human generated content."
- This year: I think we'll have a system that integrates speech, vision and language and interact with humans using all three in the next iteration of GPT.
- Next year: We'll figure out how to do that with 0.01x the number of parameters as we currently use.
- Five years: We'll have them running on smartwatches and earbuds.

# LLM ~~demo~~

- truth → verify information content
- impact on everyones lives
- privacy → what do they do w/ it?

[ co? dependency? → critical thinking.

- lack of knowledge
- (misinformation) depending on demographics
- too easy

controls → who do they answer to?

whose ethics should this follow

---

- architecture & data transparency
- interpretability → pause?
- mechanisms to not use private information.
- accountability.

# DALL-E

- An synthwave style ancient city in a lush rainforest with a backdrop of moonlight and lightning
- A halloween themed introduction to neural networks in the style of Studio Ghibli

# Ethical concerns with large generative models

- Trained on open source data from the web – often (via maximum likelihood estimation) on content that is sexist, racist and misogynist.
- The resulting predictive content can then be biased.
- Legal implication of training models on data from the internet.
- Safety and alignment – who watches the watchmen?

# Acknowledgements

This module was created as part of an Embedded Ethics Education Initiative (E3I), a joint project between the Department of Computer Science[1] and the Schwartz Reisman Institute for Technology and Society[2], University of Toronto.

**Instructional Team:**

Roger Grosse, Steven Coyne, Emma McClure, Rahul G. Krishnan

**Faculty Advisors:**

Diane Horton[1], David Liu[1], and Sheila McIlraith[1,2]

**Department of Computer Science**
**Schwartz Reisman Institute for Technology and Society**
**University of Toronto**