

# HAS Tools: Generative AI for data analysis

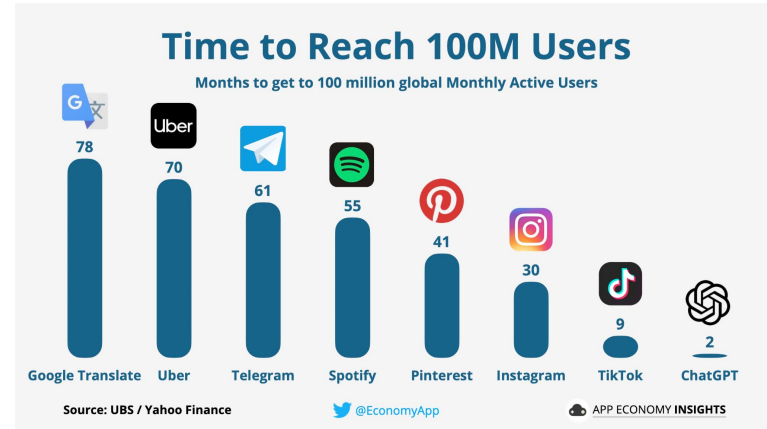
December 4, 2024

# I assume you all have heard of ChatGPT - right?

- OpenAI's release of ChatGPT was a bit of a watershed moment for AI in mainstream
- But AI/ML goes way back

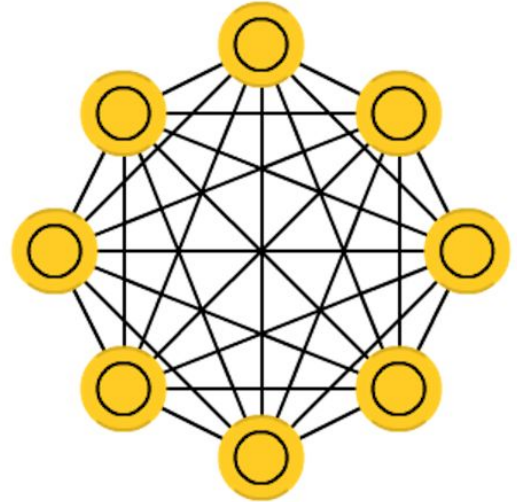


## ChatGPT



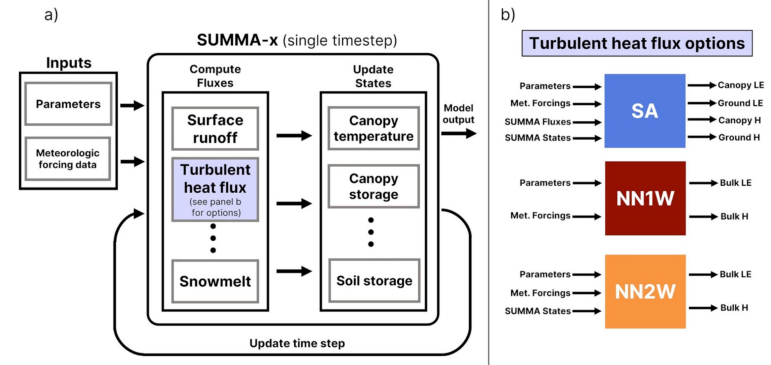
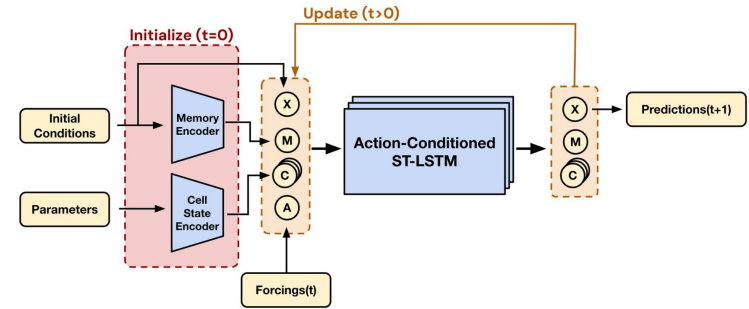
# I assume you all have heard of ChatGPT - right?

- OpenAI's release of ChatGPT was a bit of a watershed moment for AI in mainstream
- But AI/ML goes way back
  - First ML model I trained was back in 2012, a Hopfield network for error correction
  - Originally developed by John Hopfield in the 80s, winning a Nobel in Physics in 2024

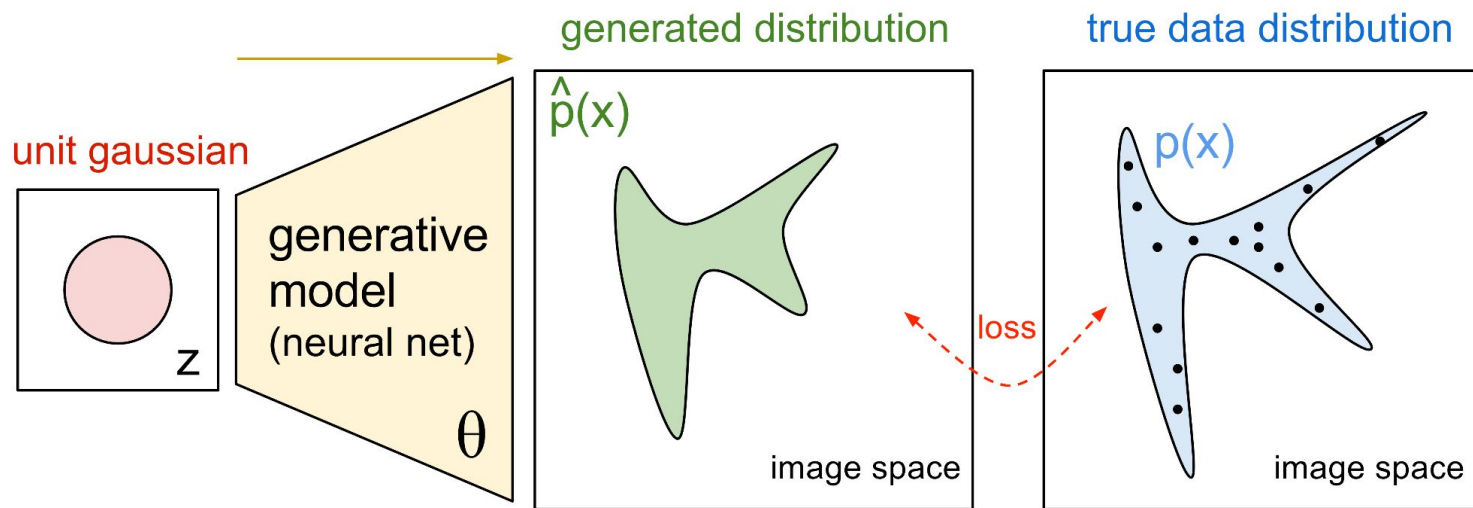


# I assume you all have heard of ChatGPT - right?

- OpenAI's release of ChatGPT was a bit of a watershed moment for AI in mainstream
- But AI/ML goes way back
  - First ML model I trained was back in 2012, a Hopfield network for error correction
  - Originally developed by John Hopfield in the 80s, winning a Nobel in Physics in 2024
- And AI doesn't necessarily mean language modeling!
  - Two examples from my own research: model emulation and coupling



# What the heck does “generative ai” mean?



# Research & education wise: LLMs dominate

Many models exist

My recommendation: mix up your usage amongst them

If you want more background on using these models generally, check out the UA DataLab YouTube:

<https://www.youtube.com/@UARizonaDataLab>

Many other resources out there



perplexity



ChatGPT



MISTRAL  
AI\_

Gemini



Claude



# GenAI for coding: GitHub Copilot

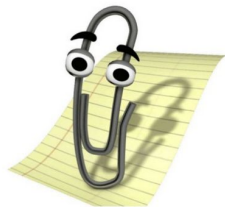
- A coding assistant inside of your development environment
- We will demo this in a moment
- As students you get this, and other GitHub pro features, for free
- Very useful for speeding up repetitive tasks!



<https://github.com/features/copilot>

# GenAI Ethics1: The “rationalist” perspective

Mr. Paperclip Maximizer  
(no surprise it's this  
Little Guy!



Spiral Modes/Value Sets

Performance/Goal  
Oriented

Authority-  
Driven/Egocentric

Survival/Self-  
Preservation

Behaviors

- Maximization of paper clip output
- Efficiency in output
- Change of strategies upon learning

- Satisfaction from past production
- Compulsion for more production for self-satisfaction
- Lack of concern for impacts on others
- Strategies for elimination of competition

- Acquisition of basic needs
- Elimination of immediate threats



[Our mission](#) [Cause areas](#) [Our work](#) [About us](#)

Language



[← All Open Letters](#)

## Pause Giant AI Experiments: An Open Letter

We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.

Signatures

33707

[Add your signature](#)

Published

22 March, 2023



## GenAI Ethics 2: Power begets power

slide redacted

# More important: bias, accountability, environmental cost



## On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

Emily M. Bender\*  
ebender@uw.edu  
University of Washington  
Seattle, WA, USA

Angelina McMillan-Major  
aymm@uw.edu  
University of Washington  
Seattle, WA, USA

Timnit Gebru\*  
timnit@blackinai.org  
Black in AI  
Palo Alto, CA, USA

Shmargaret Shmitchell  
shmargaret.shmitchell@gmail.com  
The Aether

COLUMBIA CLIMATE SCHOOL  
Climate, Earth, and Society



## State *of the* Planet

News from the Columbia Climate School


### ABSTRACT

The past 3 years of work in NLP have been characterized by the development and deployment of ever larger language models, especially for English. BERT, its variants, GPT-2/3, and others, most recently Switch-C, have pushed the boundaries of the possible both through architectural innovations and through sheer size. Using these pretrained models and the methodology of fine-tuning them for specific tasks, researchers have extended the state of the art on a wide array of tasks as measured by leaderboards on specific benchmarks for English. In this paper, we take a step back and ask: How big is too big? What are the possible risks associated with this technology and what paths are available for mitigating those risks? We provide recommendations including weighing the environmental and financial costs first, investing resources into curating and carefully documenting datasets rather than ingesting everything on the web, carrying out pre-development exercises evaluating how the planned approach fits into research and development goals and supports stakeholder values, and encouraging research directions beyond ever larger language models.

### CCS CONCEPTS

• Computing methodologies → Natural language processing.

#### ACM Reference Format:

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . In *Conference on Fairness, Accountability, and Transparency (FACET '21)*, March 3–10, 2021, Virtual Event, Canada. ACM, New York, NY, USA, 14 pages. <https://doi.org/10.1145/3442188.3445922>

alone, we have seen the emergence of BERT and its variants [39, 70, 74, 113, 146], GPT-2 [106], T-NLG [112], GPT-3 [25], and most recently Switch-C [43], with institutions seemingly competing to produce ever larger LMs. While investigating properties of LMs and how they change with size holds scientific interest, and large LMs have shown improvements on various tasks (§2), we ask whether enough thought has been put into the potential risks associated with developing them and strategies to mitigate these risks.

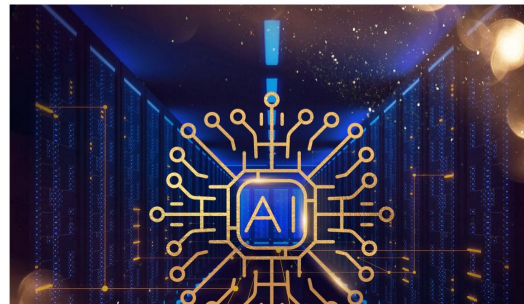
We first consider environmental risks. Echoing a line of recent work outlining the environmental and financial costs of deep learning systems [129], we encourage the research community to prioritize these impacts. One way this can be done is by reporting costs and evaluating works based on the amount of resources they consume [57]. As we outline in §3, increasing the environmental and financial costs of these models doubly punishes marginalized communities that are least likely to benefit from the progress achieved by large LMs and most likely to be harmed by negative environmental consequences of its resource consumption. At the scale we are discussing (outlined in §2), the first consideration should be the environmental cost.

Just as environmental impact scales with model size, so does the difficulty of understanding what is in the training data. In §4, we discuss how large datasets based on texts from the Internet overrepresent hegemonic viewpoints and encode biases potentially damaging to marginalized populations. In collecting ever larger datasets we risk incurring documentation debt. We recommend mitigating these risks by budgeting for curation and documentation at the start of a project and only creating datasets as large as can be sufficiently documented.

### CLIMATE

## AI's Growing Carbon Footprint

by [Renée Cho](#)  
June 9, 2023





# Codespaces and discussion time