LLMs: Where are we and how did we get here?

Lunch & Learn

Sacha Davis | June 19, 2024

Introduction

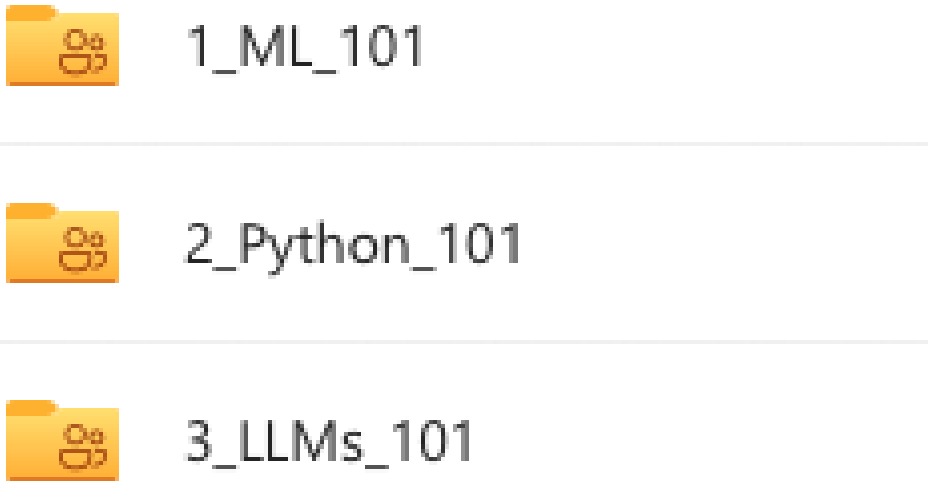
- University of Alberta:
  - 2020: B.Sc. Biological Sciences with minor in Computing Science
  - 2023: M.Sc. Computing Science (Thesis-Based)
- Co-Director of the Artificial Intelligence in Medical Systems Society (aimss.ca)
- Research, project management, ML engineering, community engagement, teaching, etc.



Sacha Davis
Co-op work experience Student
Sacha.Davis@albertahealthservices.ca

Next Lunch & Learn Sessions

- ~~April 17: Machine Learning 101~~
- ~~May 15: Python 101~~
- June 19: Large Language Models



Presentation Goal #1:

Explain, at a high level, how Large Language Models work

Presentation Goal #2:

Cover a history of the recent AI race and the key players

Presentation Goal #3:

Cover what LLMs can and cannot do well

Presentation Not Goal #1:
Cover tons of technical details

Presentation Not Goal #2:
Cover the history of AI pre-2015

Presentation Not Goal #3:
Get in trouble with AHS

AHS Generative AI Warning

Alberta Health Services

You are entering a public, generative AI site where the information you provide is retained by a third-party, which is not within the control of Alberta Health Services. This is a reminder not to share confidential AHS health, personal, or business information on this site.

For more information, please review the resources at please click here

URL https://chat.openai.com/

Proceed | Go Back

InfoCare

Best Practices:
Using third-party artificial intelligence (AI) at AHS

Artificial intelligence (AI) can refer to any attempt to simulate human thinking with software. This can include generative AI, large language models (such as OpenAI's ChatGPT), speech-to-text dictation and analyzing trends from data sets.

As Alberta Health Services (AHS) considers how to approach the use of third-party AI tools, our workforce is asked to follow these preliminary guidelines for the use of AI at AHS:

- Do not load any confidential AHS health, personal, or business information into unapproved AI programs or tools.
- Use caution when accessing AI websites with AHS devices.

Approval to use AI tools may only be granted upon the completion of a Privacy Impact Assessment (PIA), as outlined in policy 1145. A PIA is typically completed when a new application, information system or administrative process is introduced, or an existing system or process is changed.

When attempting to access a third-party AI tool online, users may be redirected to an alert message asking them to acknowledge their understanding of the appropriate use of AHS data:

> *"You are entering a public, generative AI site where the information you provide is retained by a third-party, which is not within the control of Alberta Health Services. This is a reminder not to share confidential AHS health, personal, or business information on this site [...] Click **here** to return to AHS.ca or click **here** to proceed."*

The absence of this alert does not mean that a third-party site is approved.

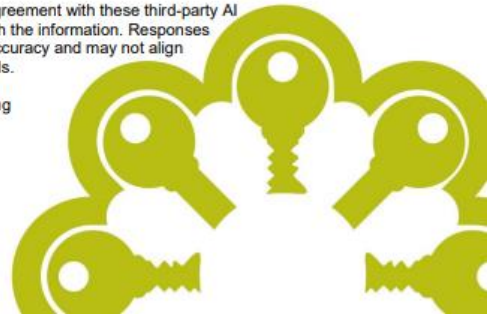
AI is a rapidly developing field with significant implications for organizations like AHS. As we navigate this evolving landscape, we are doing so with caution. While AI may be useful in certain applications, it also presents a number of risks. Sharing health, personal, or business information with generative AI sites put our organization at risk.

This is because we do not have a formal agreement with these third-party AI tools, and we can't control what they do with the information. Responses generated by AI applications also vary in accuracy and may not align with AHS' core values or business standards.

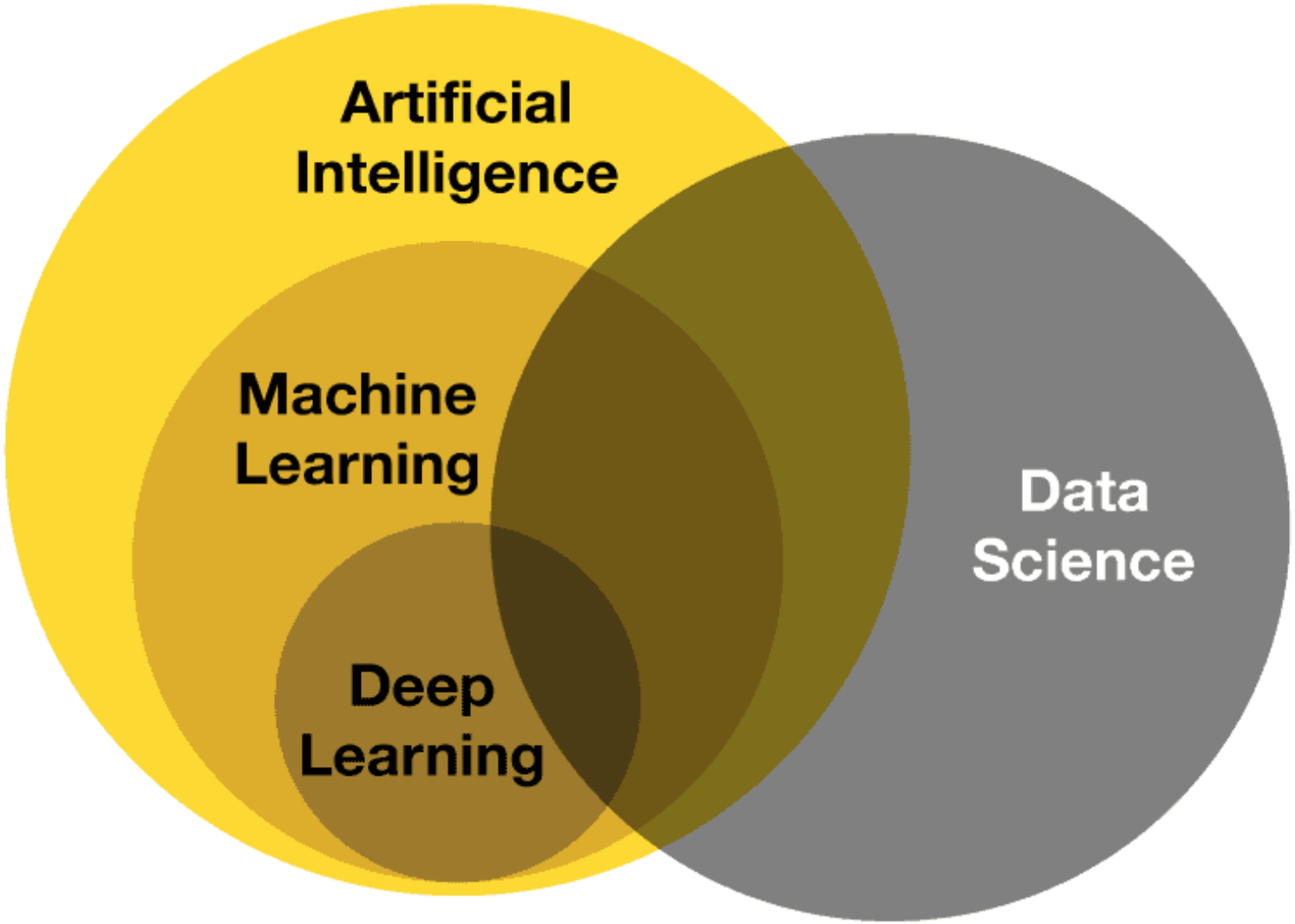
We continue to evaluate this quickly evolving topic and appreciate your patience.
If you have questions, email InfoCare@ahs.ca.

Alberta Health Services

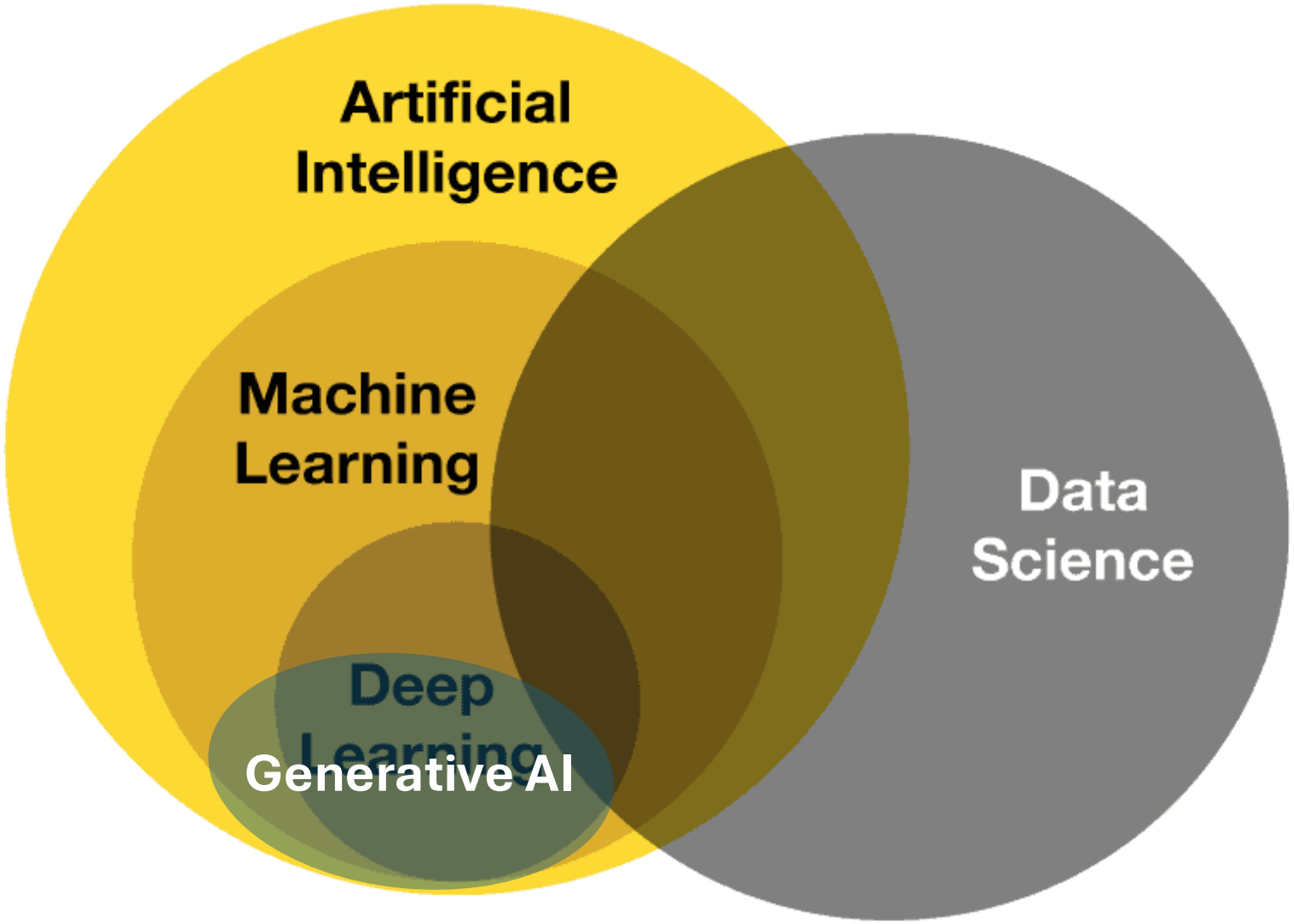
Updated: September 2023



What is a Large Language Model?


Artificial Intelligence
Machine Learning
Deep Learning
Data Science

[Source]


Artificial Intelligence
Machine Learning
Deep Learning
Generative AI
Data Science

[Source]

Predictive AI

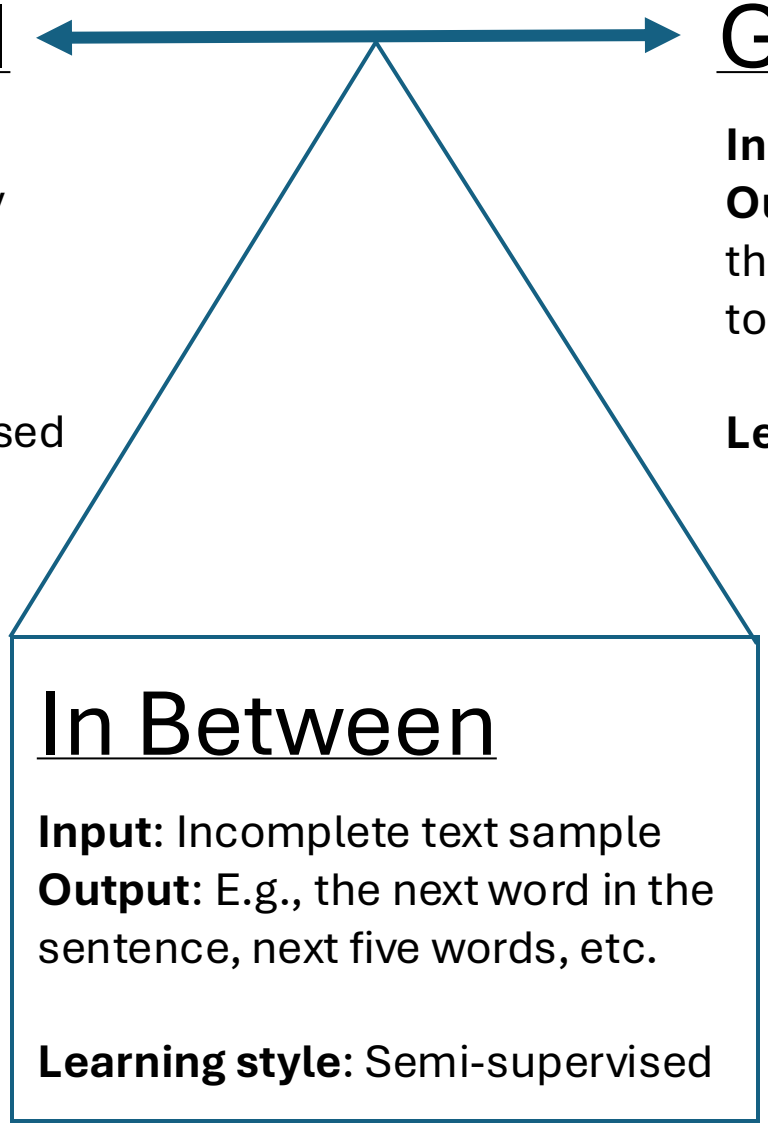
Input: Text sample
Output: E.g., most likely classification (positive, negative, neutral)

Learning style: Supervised

Generative AI

Input: Text sample
Output: A new text sample that is an appropriate response to the input

Learning style: Semi-supervised

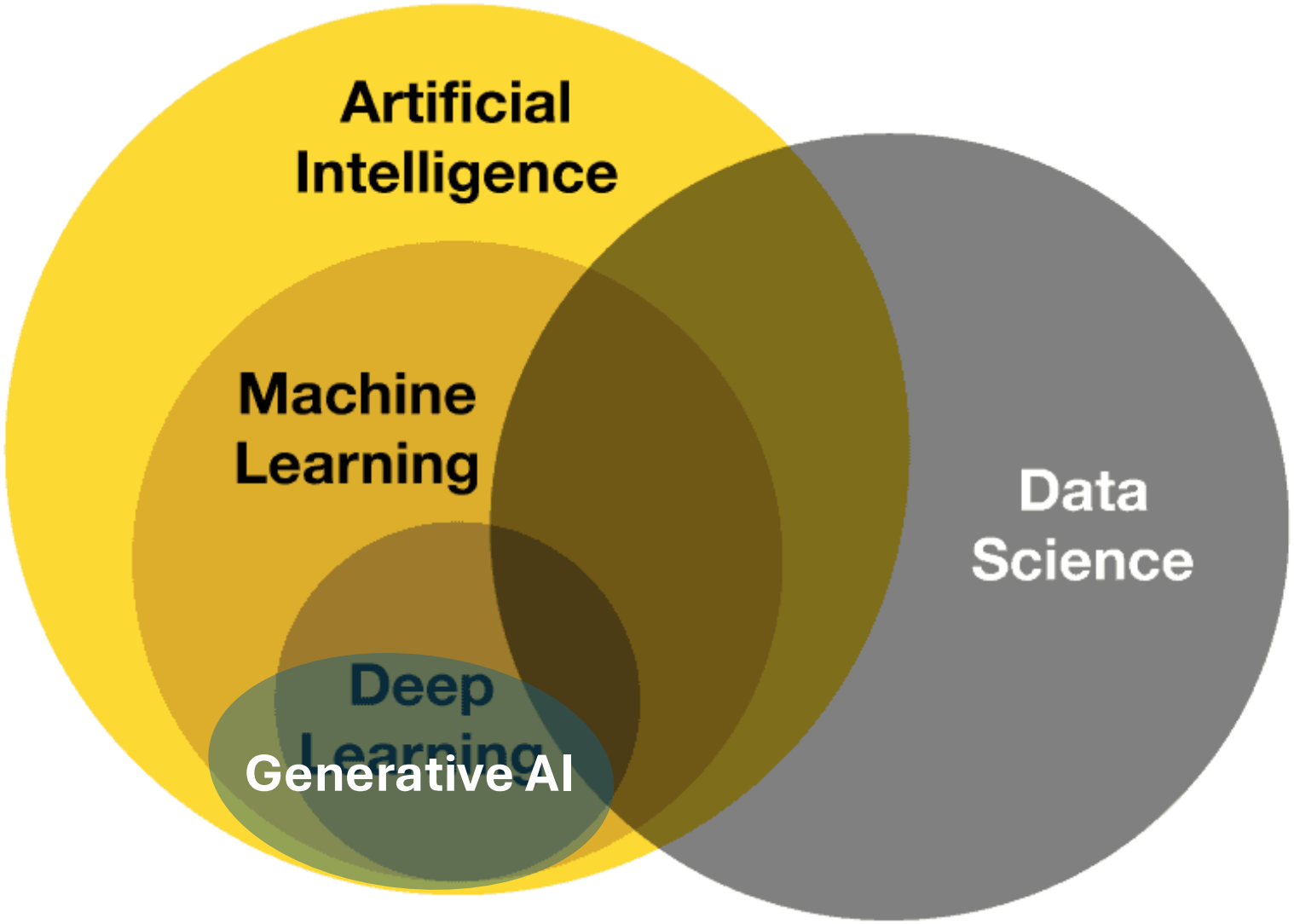


In Between

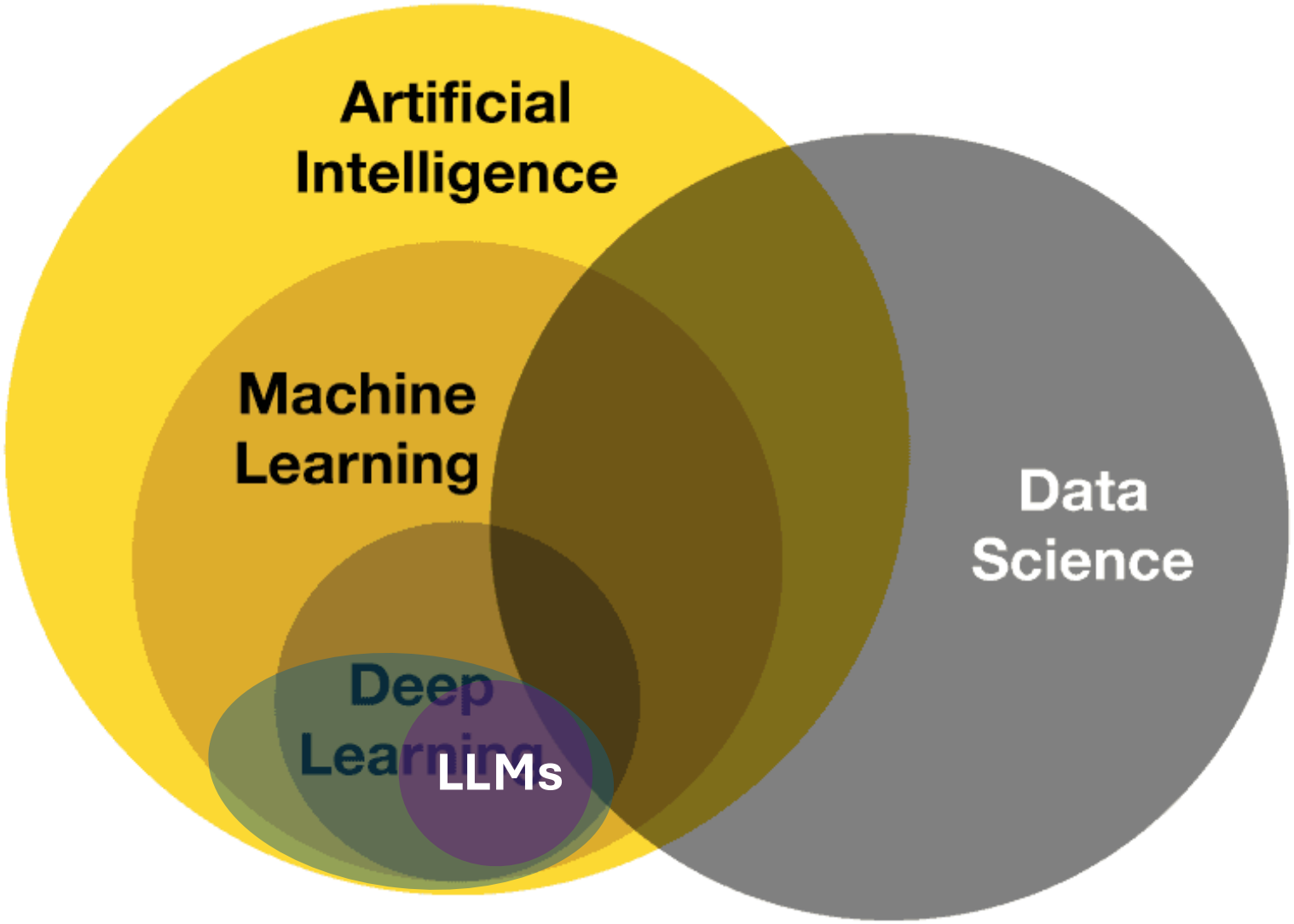
Input: Incomplete text sample
Output: E.g., the next word in the sentence, next five words, etc.

Learning style: Semi-supervised

(In the natural language context…)


Artificial Intelligence
Machine Learning
Deep Learning
Generative AI
Data Science

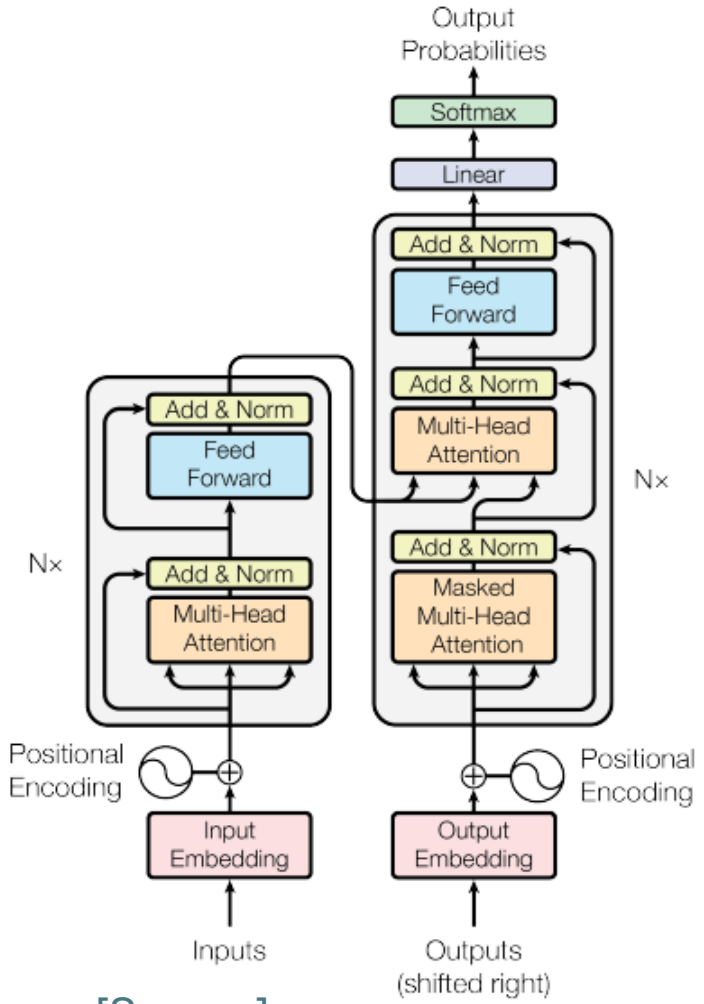
[Source]


Artificial Intelligence
Machine Learning
Deep Learning
LLMs
Data Science

[Source]

Large Language Models

- Huge generative models capable of understanding and generating human-like text
- Core components of an LLM:
  - Model architecture
  - Training data
  - Training process
- Scaling hypothesis?



[Source]

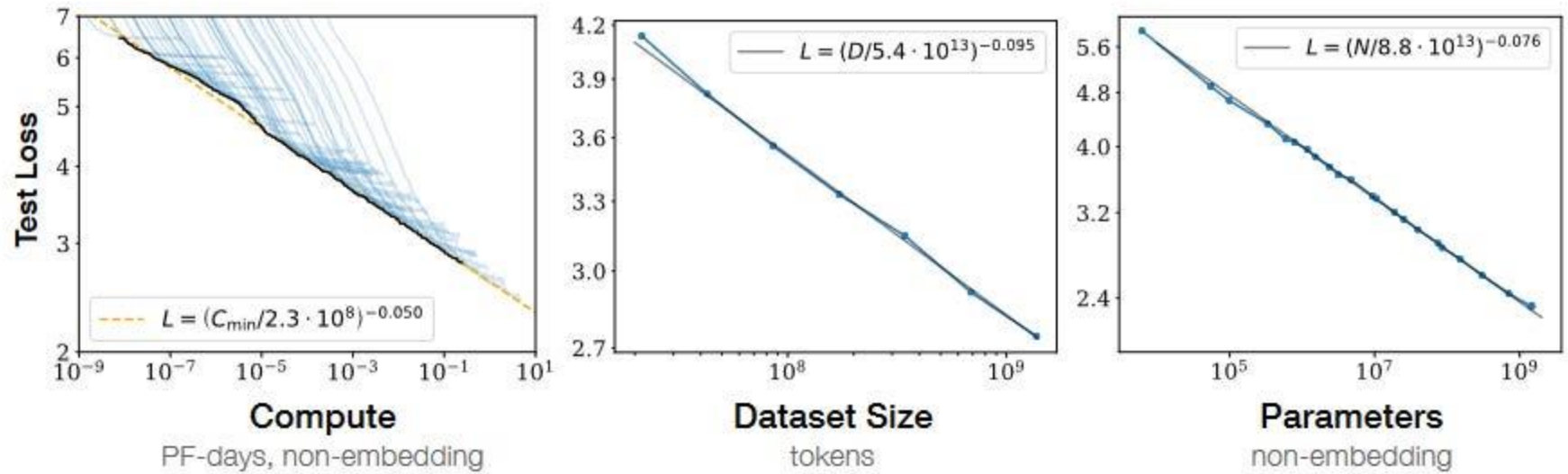


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

[Source]

An Abridged History of Our Current AI Spring

Major players:











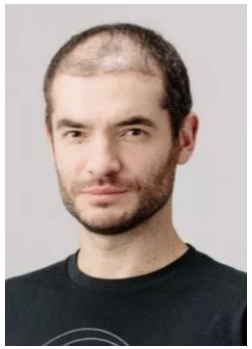




2015-2020


Sam Altman


Elon Musk

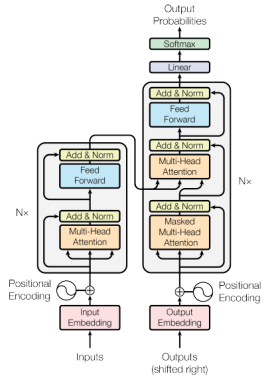

Ilya Sutskever

OpenAI

- 2015: Founded
- 2018: First Generative Pretrained Transformer (GPT), Elon leaves
- 2019: Our structure | OpenAI, GPT-2, Microsoft agreement
- 2020: GPT-3



Microsoft



Google Brain

- 2017: "Attention is all you need"

cohere

- 2019: Founded
- Enterprise LLM solutions

2021



[Source]

OpenAI

7 employees

ANTHROP\C

- Founded to "study safety properties <of AI> at the technological frontier"

GitHub Copilot

Alphabet Google

- LaMDA model

2022



OpenAI

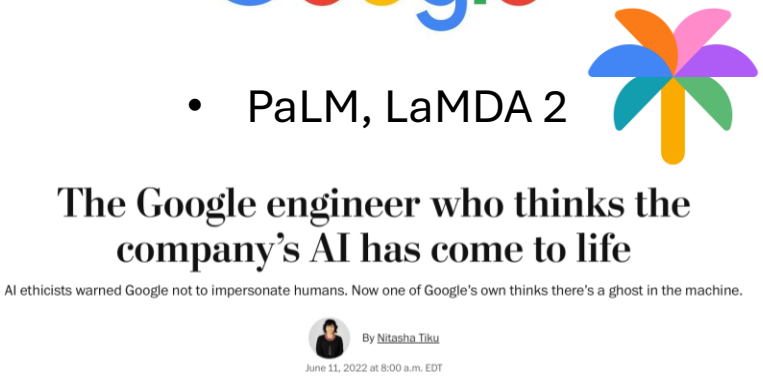
- November: ChatGPT released

cohere



Alphabet
Google

- PaLM, LaMDA 2



Meta

- OPT model

2023

Meta

Llama



- MTIA v1 accelerator

Claude

ANTHROP\C

- Claude, Claude Instant, Claude 2

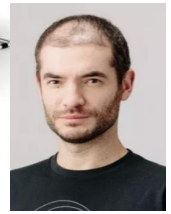
$2bn



Sam Altman







Ilya Sutskever

cohere

ORACLE





Alphabet

Google

aws

Bard

OpenAI



$10bn

Microsoft

Bing

Copilot

Jan-May 2024

Alphabet

Google

~~Bard~~ Gemini

GOOGLE / ARTIFICIAL INTELLIGENCE / TECH

Google CEO says Gemini AI diversity errors are ‘completely unacceptable’



Elon Musk





Sam Altman

GPT-4o



OpenAI

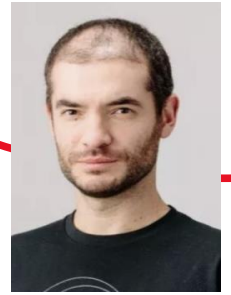
- GPT 4o demo
- GPT 4 vs. 4o

Microsoft



ANTHROP\C

Claude 3



Ilya Sutskever

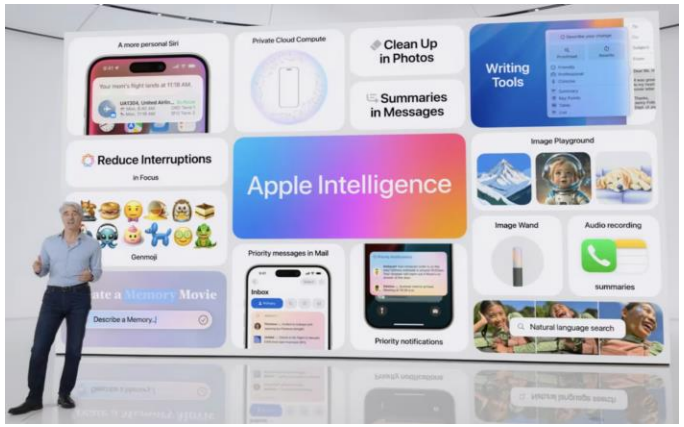


Jan Leike



June 2024





"We're doing AI too!"

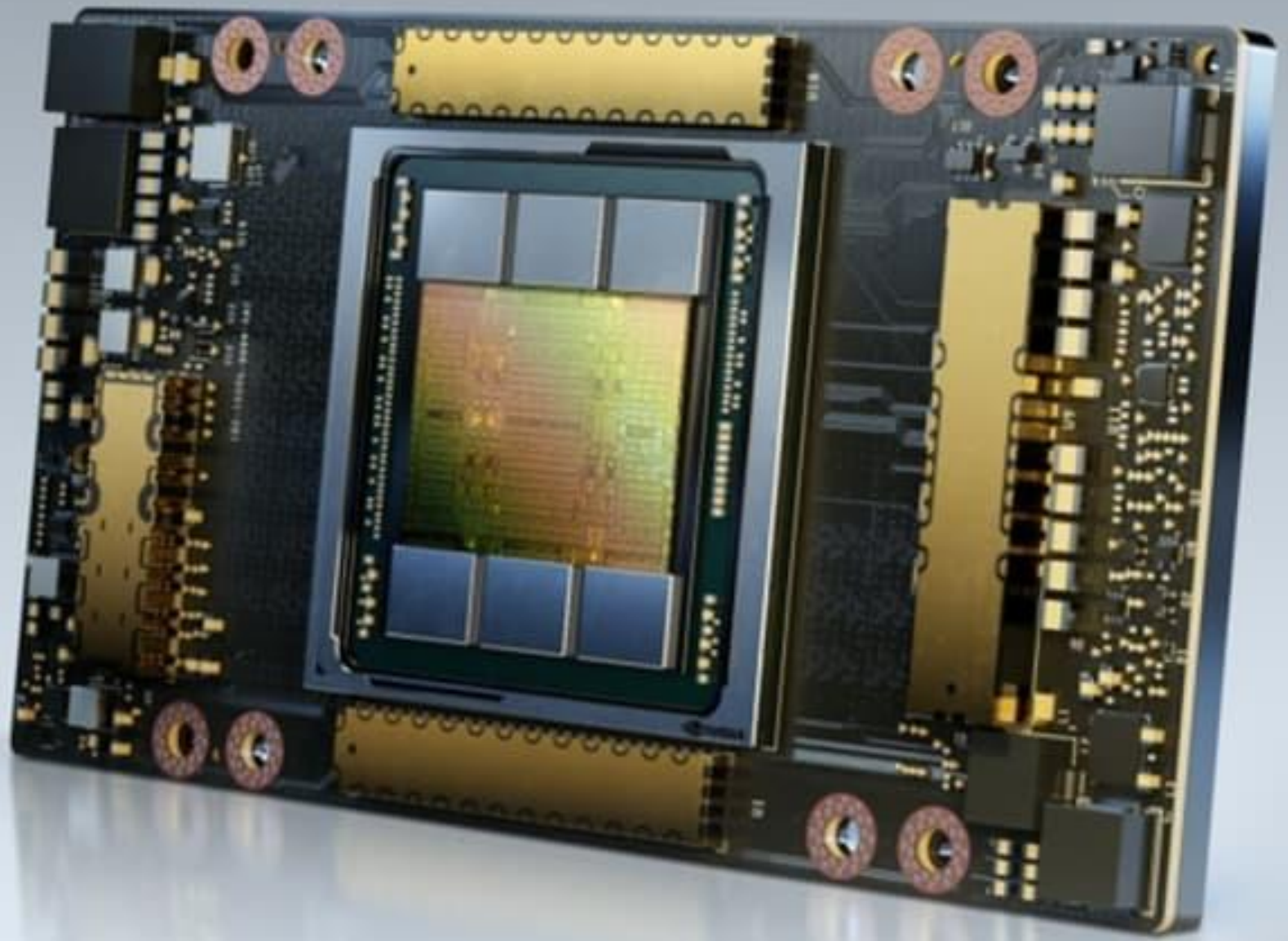




Elon Musk



OpenAI



[Source]

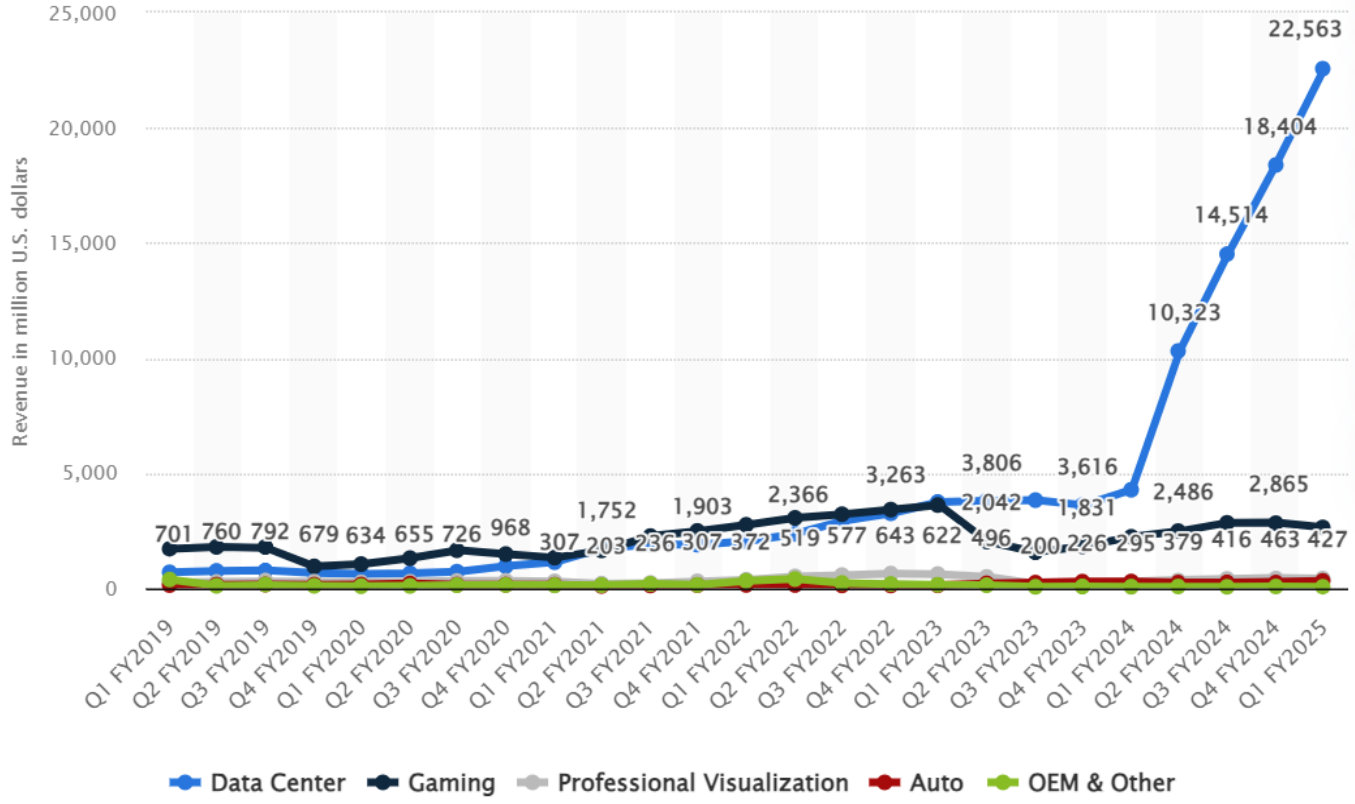
NVIDIA Takeover



(Now the ~~second~~ highest valued company in the world)

Nvidia specialized market revenue worldwide from fiscal year 2019 to 2025, by quarter

(in million U.S. dollars)



[Source]

Full Coverage

Nvidia surpasses Microsoft as world's most valuable company

Top news



Bloomberg

Nvidia's (NVDA) Rally to Most Valuable Stock: How It Happened

20 hours ago

Forbes

Nvidia's Jensen Huang Now World's 11th-Richest Person As Nvidia Soars Past…



22 hours ago

CNBC

Nvidia passes Microsoft in market cap to become most valuable public company



1 day ago

Yahoo Finance

Q&A: How did Nvidia become world's most valuable company?



3 hours ago

CTV News

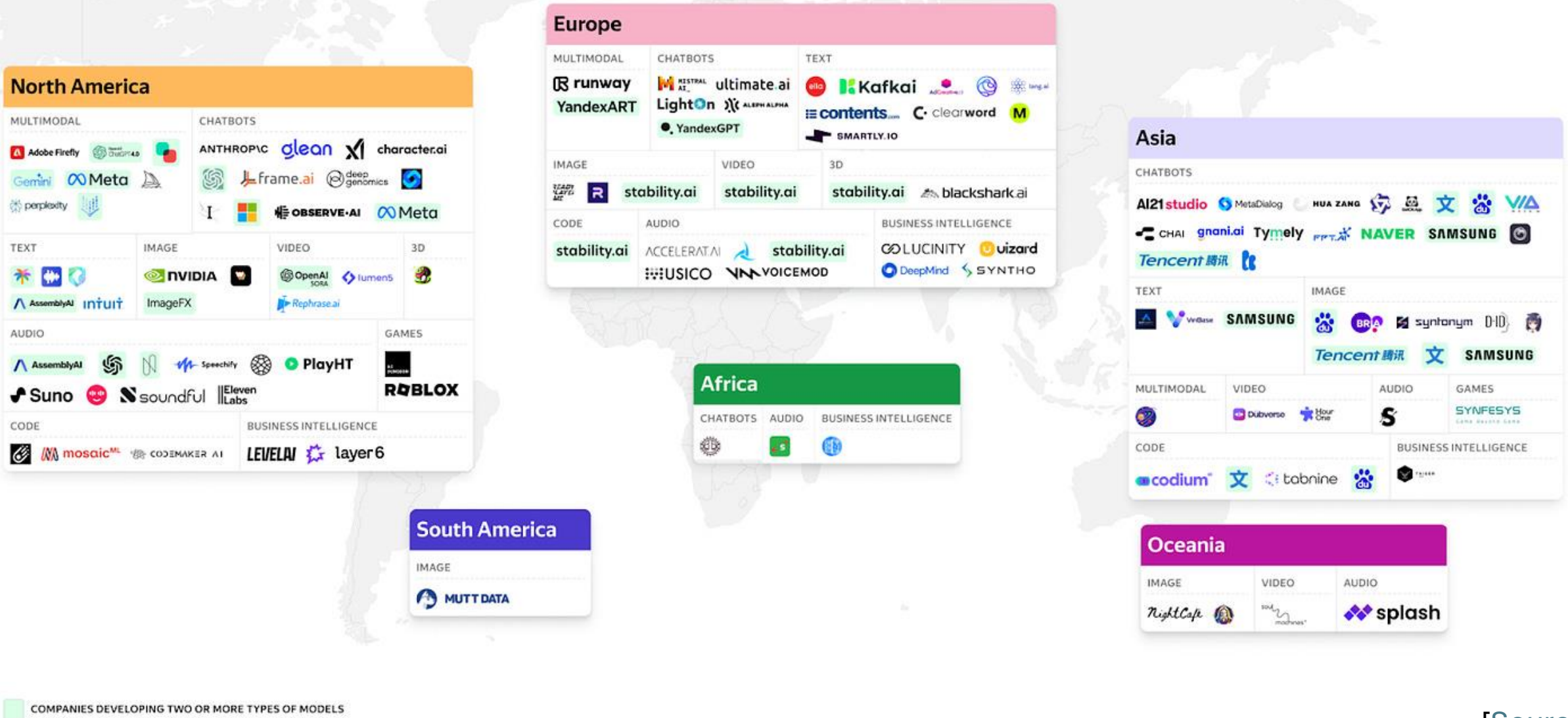
AI news: Nvidia becomes world's most valuable company



23 hours ago

The Global Generative AI Landscape 2024

Alport



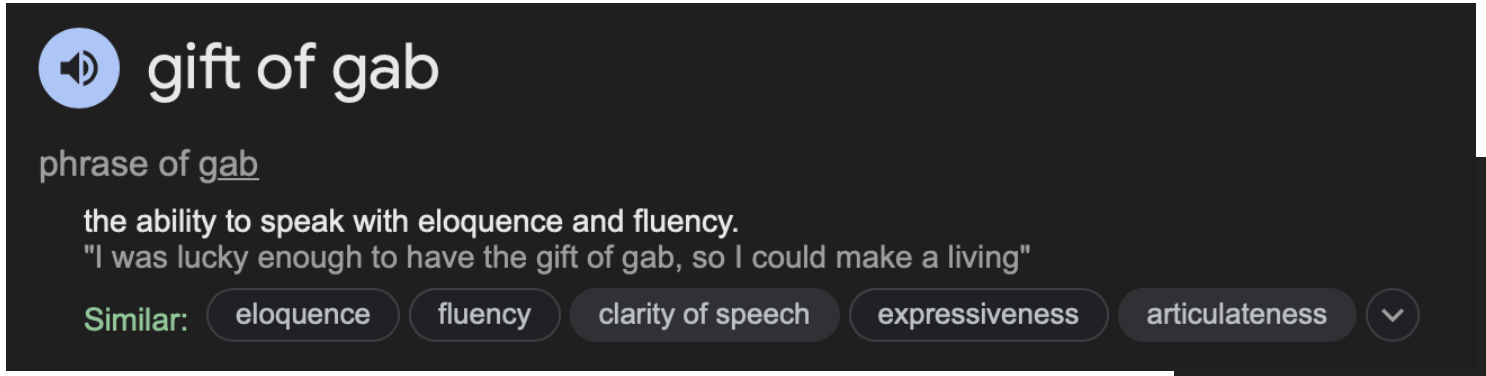
COMPANIES DEVELOPING TWO OR MORE TYPES OF MODELS

[Source]

Using LLM-Based Tools: Strengths and Weaknesses

LLM Talents

LLM Talent: Gift of Gab, Grammar & Spelling



Can you do the same for my work at the Alberta Machine Intelligence Institute as an AI in Genomics Consultant?

Developing a detailed curriculum outline for the upcoming AI+Genomics MOOC.

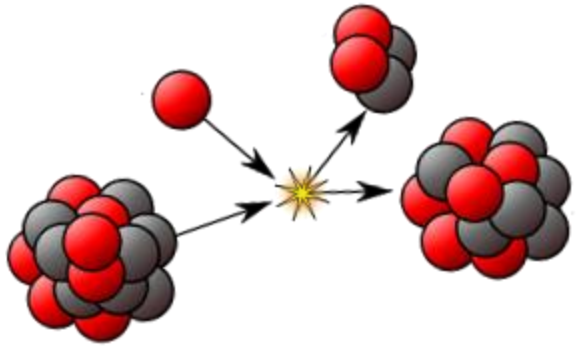
Certainly! Here's a reworded resume point for your work at the Alberta Machine Intelligence Institute:

- Crafted a comprehensive curriculum for the AI+Genomics Massive Open Online Course (MOOC), outlining key learning objectives, course materials, and instructional strategies to enhance genomic research through artificial intelligence.

LLM Talent: Information Transmutation



Starting with a seed idea and expanding on it (Zooming out)

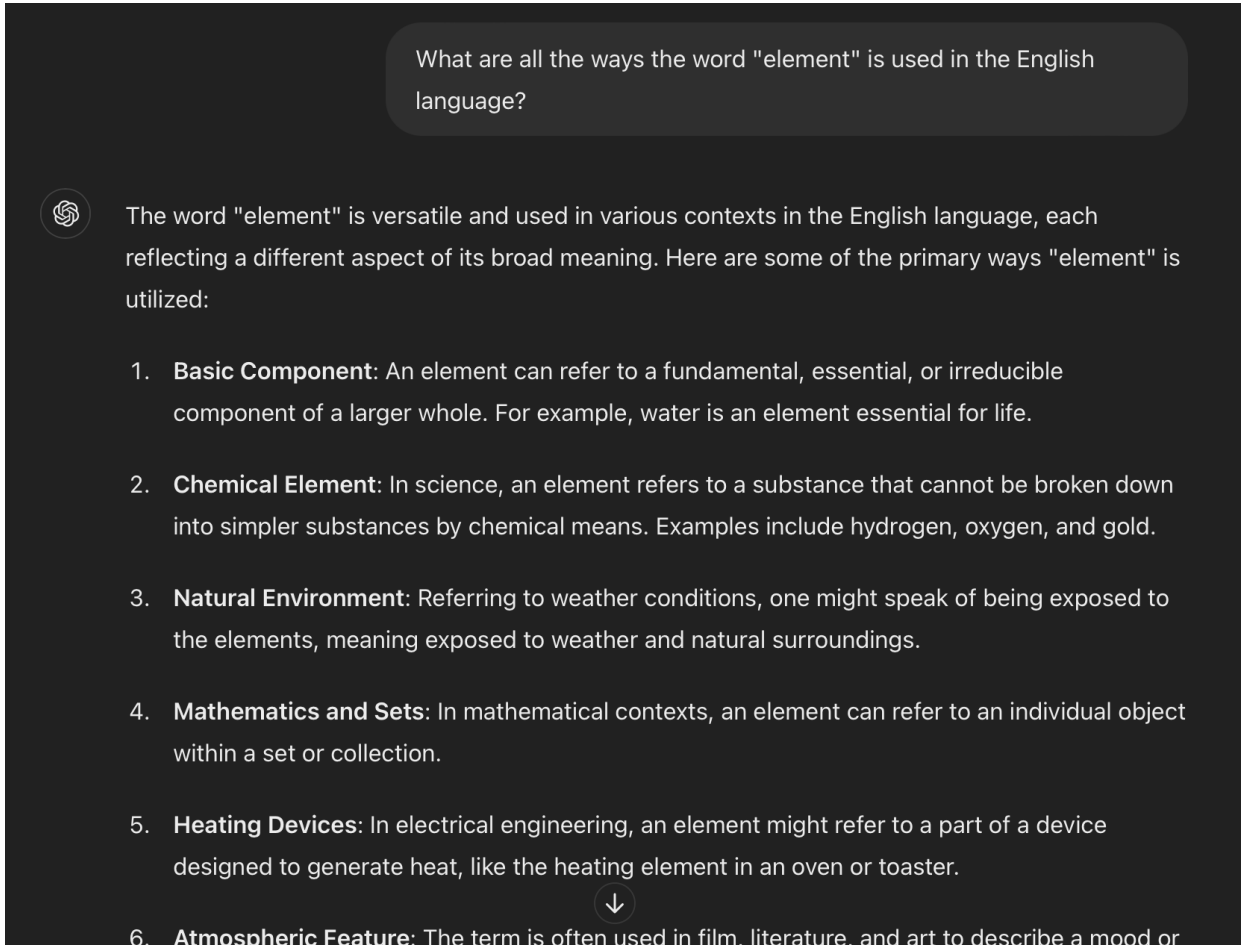


Converting information into different forms

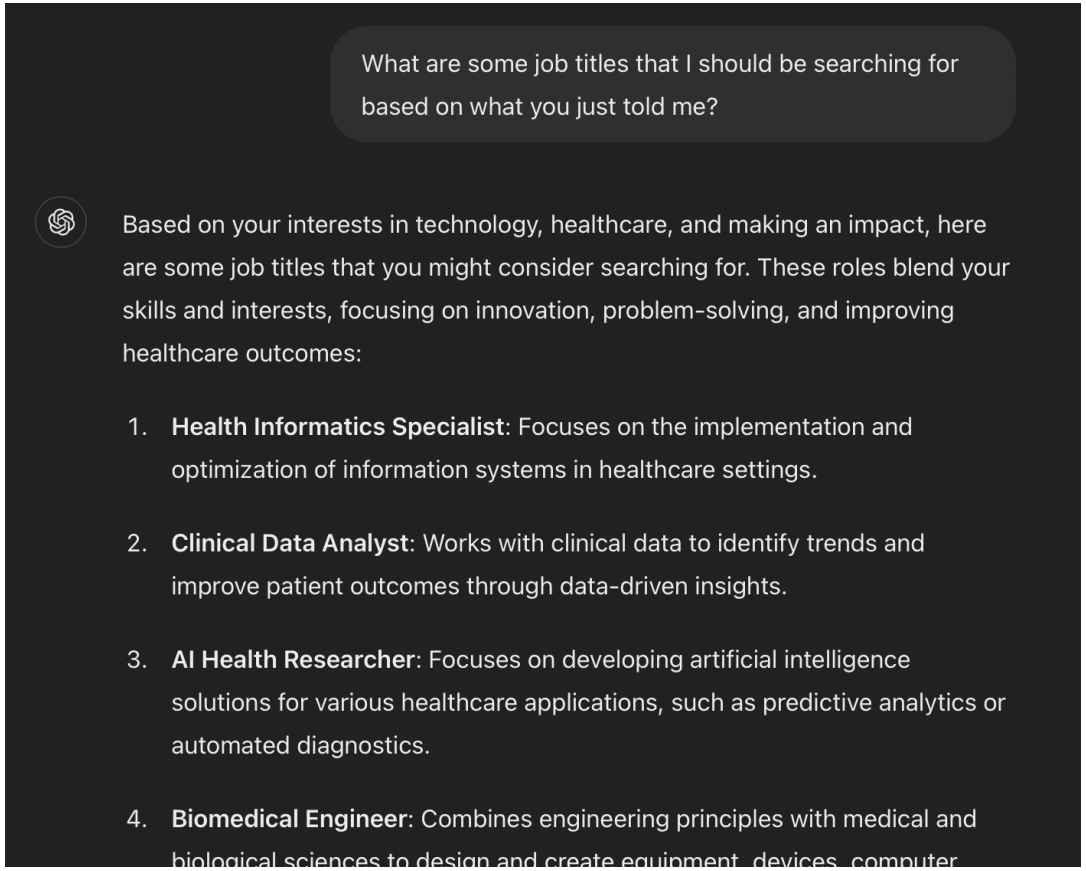


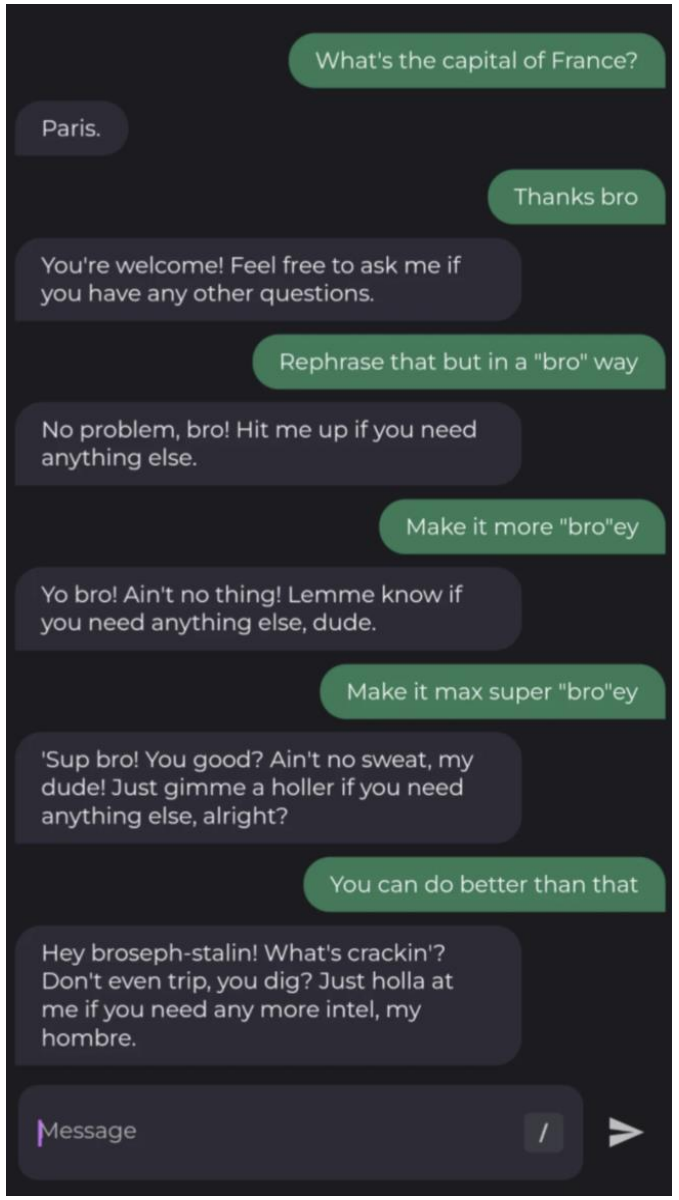
Finding important information in large heaps of unimportant information (Zooming in)

LLM Talent: Highly Knowledgeable



LLM Talent: Infinite Patience

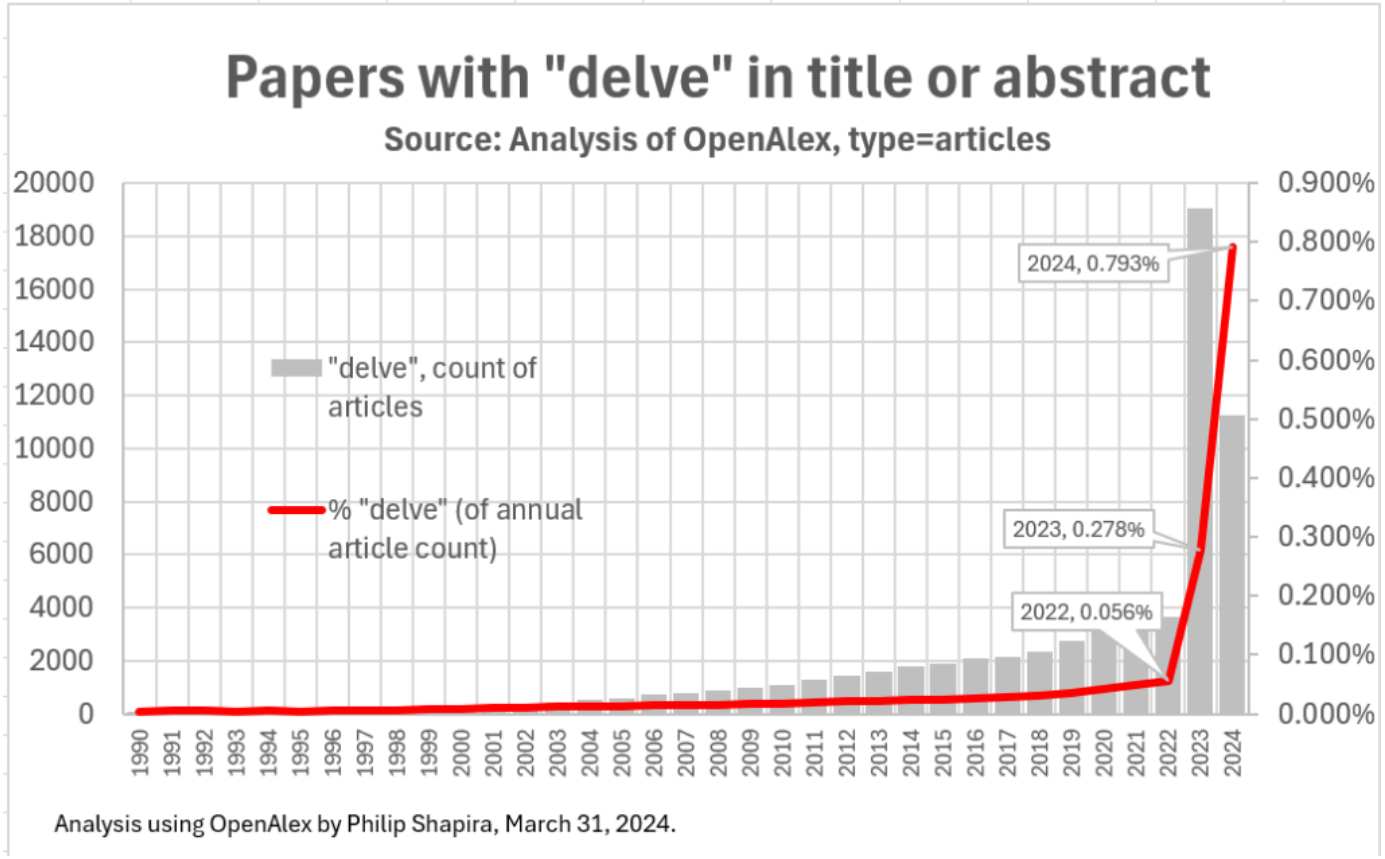




[Source]

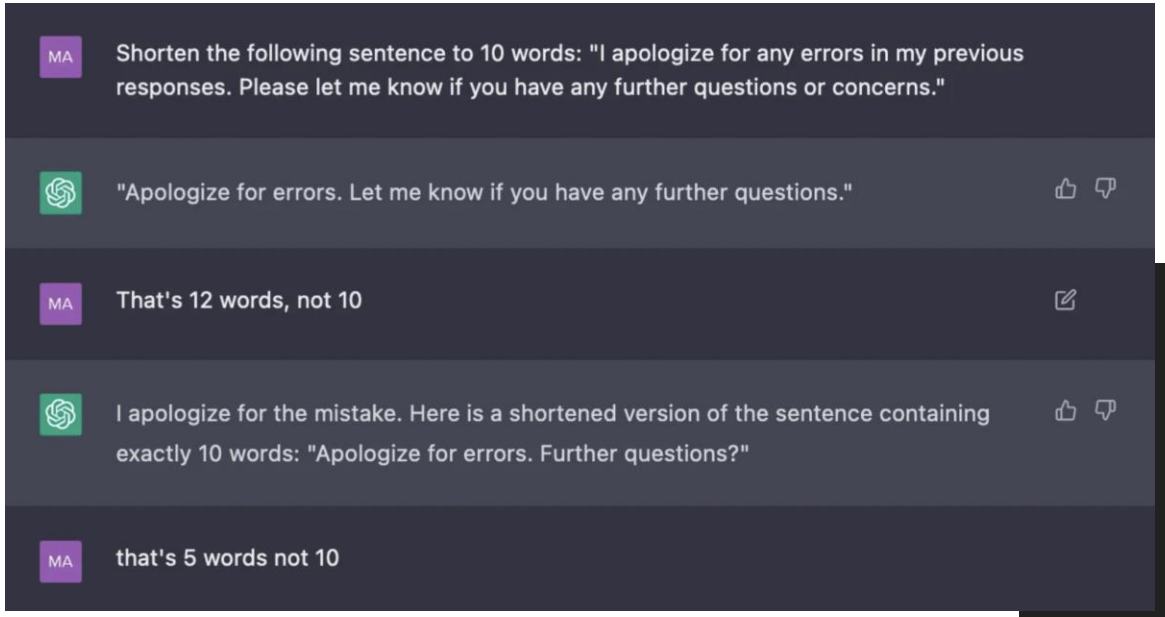
LLM Weaknesses

LLM Weakness: Creativity and Originality

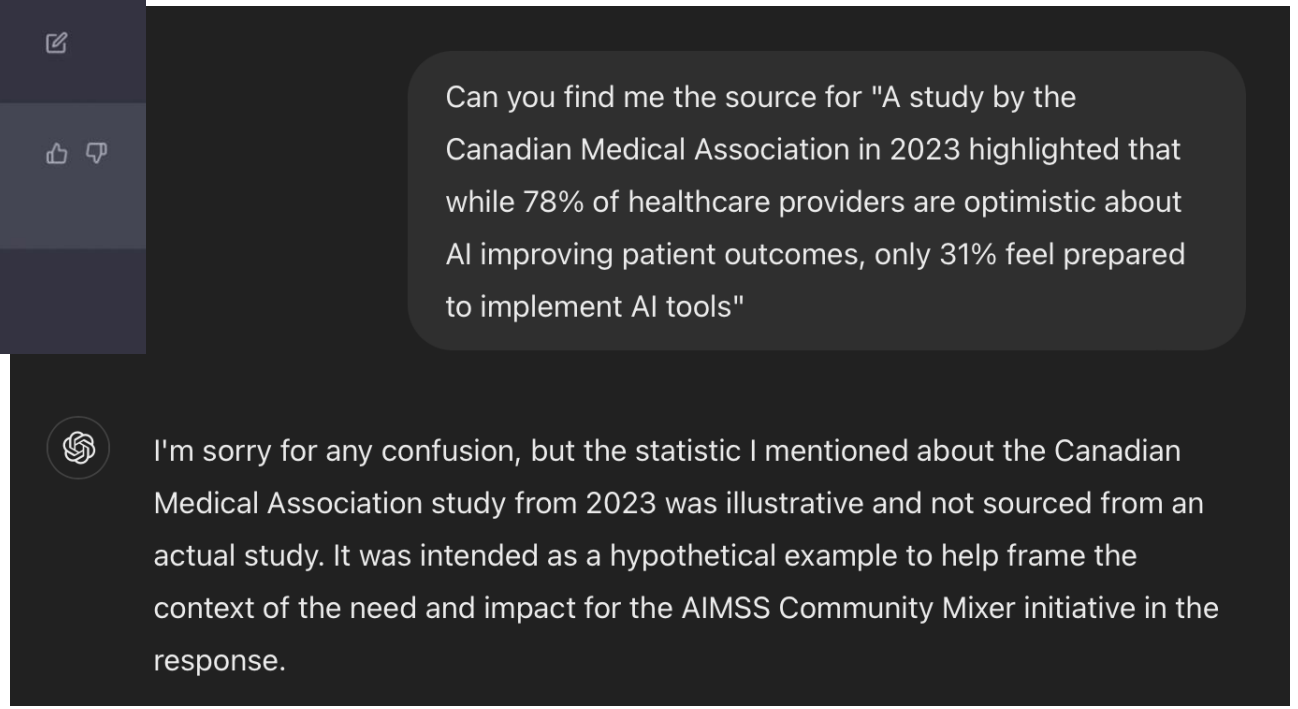


[Source]

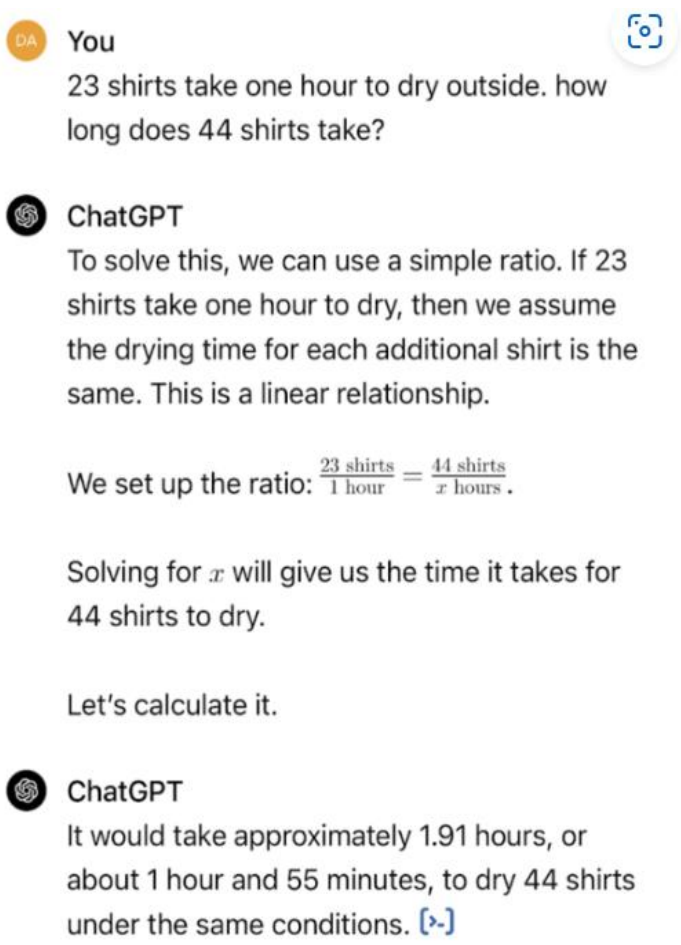
LLM Weakness: Precision and Correctness



"Hallucinations"

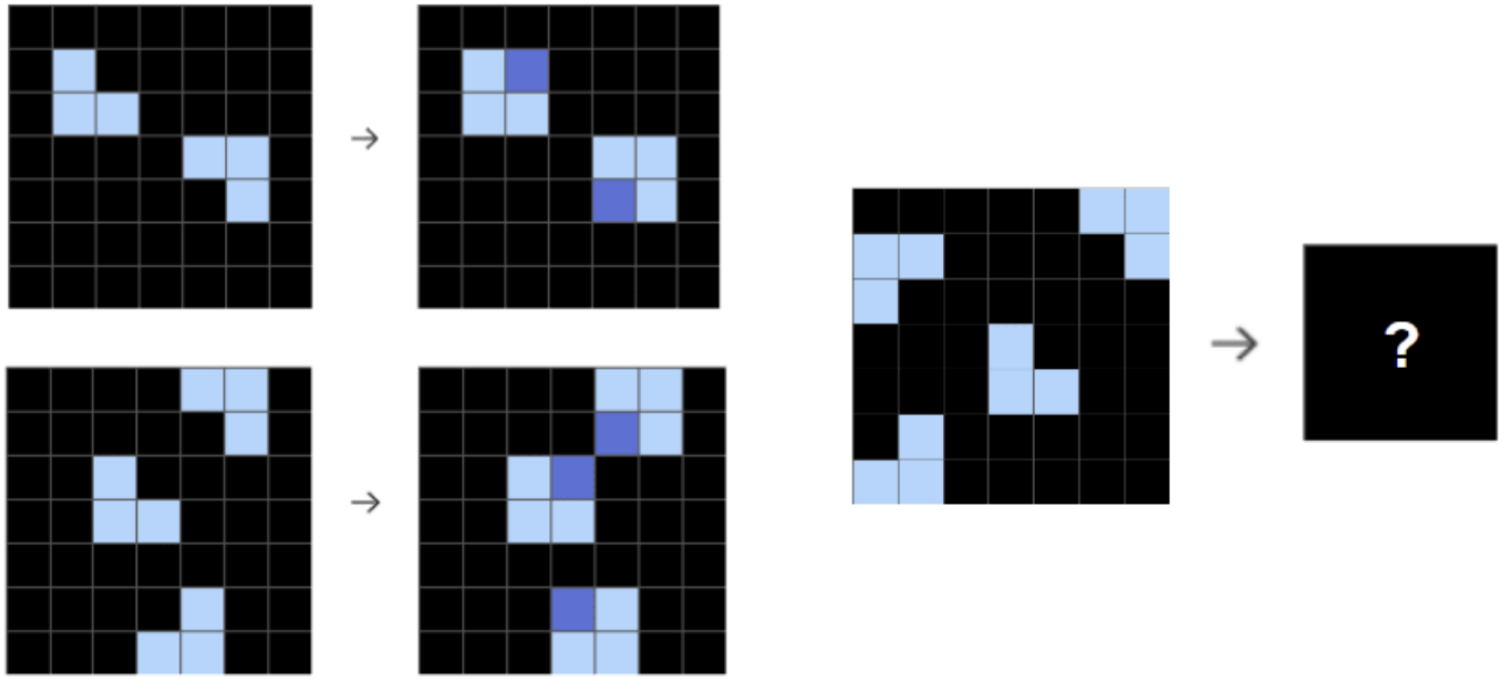


LLM Weakness: Causality and Logic



[Source]

ARC-AGI

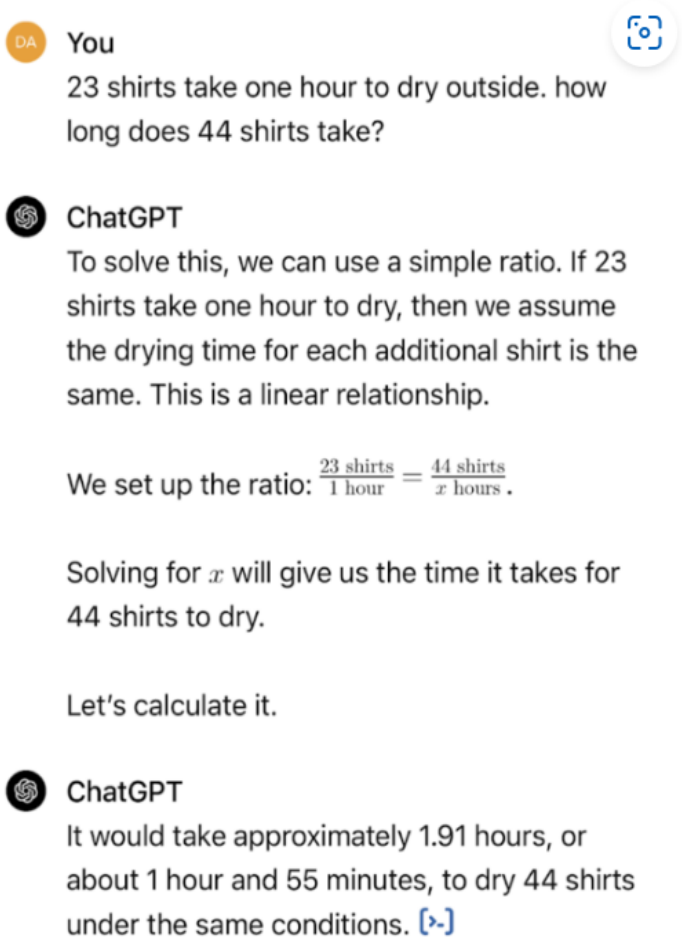




Francois Chollet

https://github.com/fchollet/ARC-AGI

LLM Weakness: Prompt Sensitivity



DA **You**

23 shirts take one hour to dry outside. how long does 44 shirts take?

ChatGPT

To solve this, we can use a simple ratio. If 23 shirts take one hour to dry, then we assume the drying time for each additional shirt is the same. This is a linear relationship.

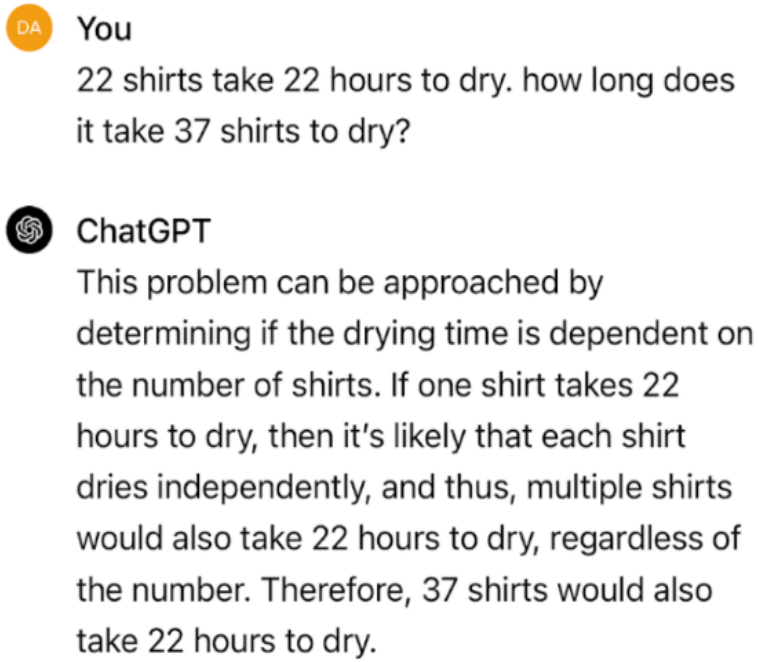
We set up the ratio: $\frac{23 \text{ shirts}}{1 \text{ hour}} = \frac{44 \text{ shirts}}{x \text{ hours}}$.

Solving for x will give us the time it takes for 44 shirts to dry.

Let's calculate it.

ChatGPT

It would take approximately 1.91 hours, or about 1 hour and 55 minutes, to dry 44 shirts under the same conditions.



DA **You**

22 shirts take 22 hours to dry. how long does it take 37 shirts to dry?

ChatGPT

This problem can be approached by determining if the drying time is dependent on the number of shirts. If one shirt takes 22 hours to dry, then it's likely that each shirt dries independently, and thus, multiple shirts would also take 22 hours to dry, regardless of the number. Therefore, 37 shirts would also take 22 hours to dry.

[Source]

LLM Weakness: Conciseness



r/ChatGPT • 2 yr. ago
mvfsullivan

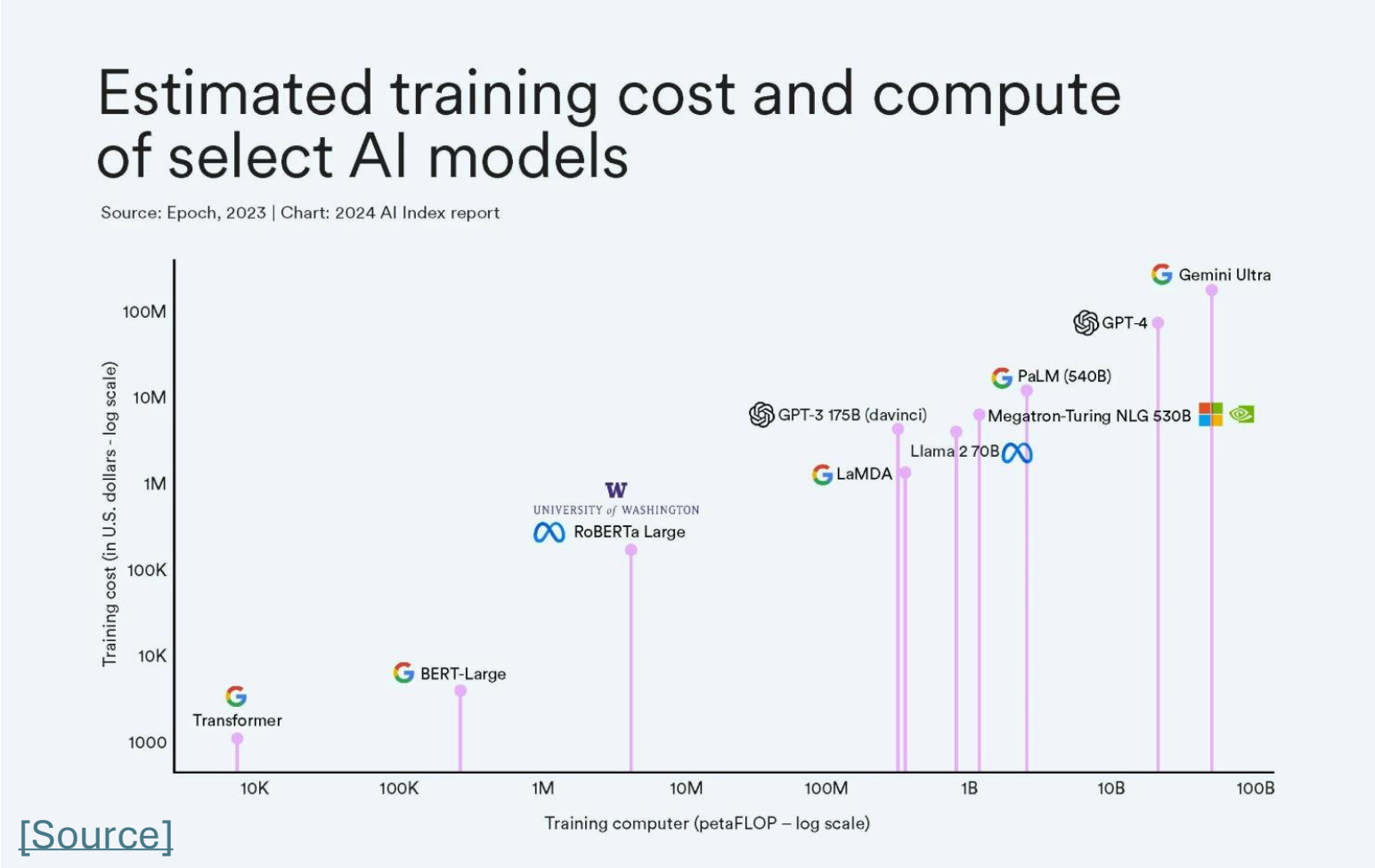
Why is chatgpt so excessively wordy?

Other

Its so infuriating having to sit and wait for it for 5 mins just to ask it something else. Is there a way to speed up the animation or skip it entirely?

Even if you say things like "short answer" or "reply in one sentence or less" or "reply but do not exceed 20 words"

LLM Weakness: Computationally Intensive



[Source]

You

- Experience
- Perspective
- Creativity
- Ability to count
- Discernment
- Intuitive understanding of cause and effect
- Desire for excellence and truth

LLM

- Gift of gab
- Ability to reformulate, tailor, and personalize messages
- "Zooming in" and "zooming out"
- Information from the entire English written internet
- Infinite patience

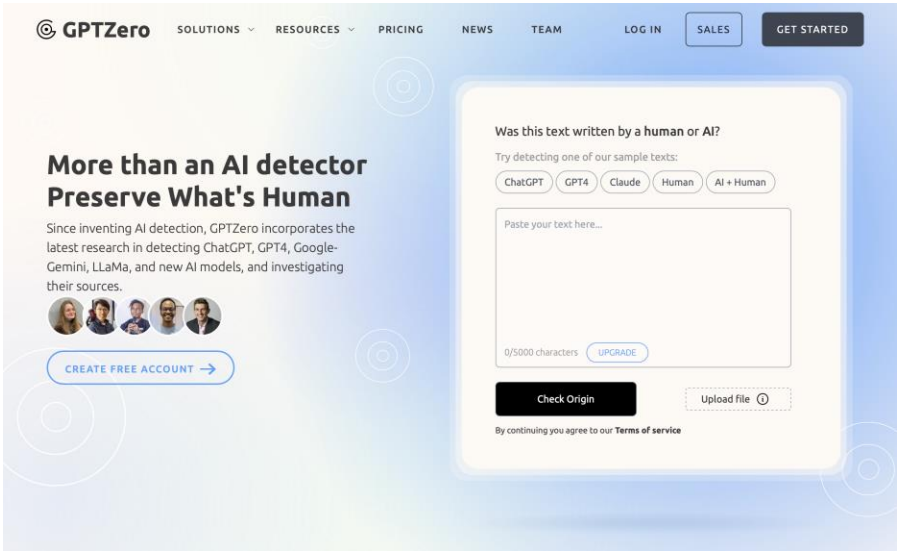
"How can we best work together?"

Writing an Effective Prompt

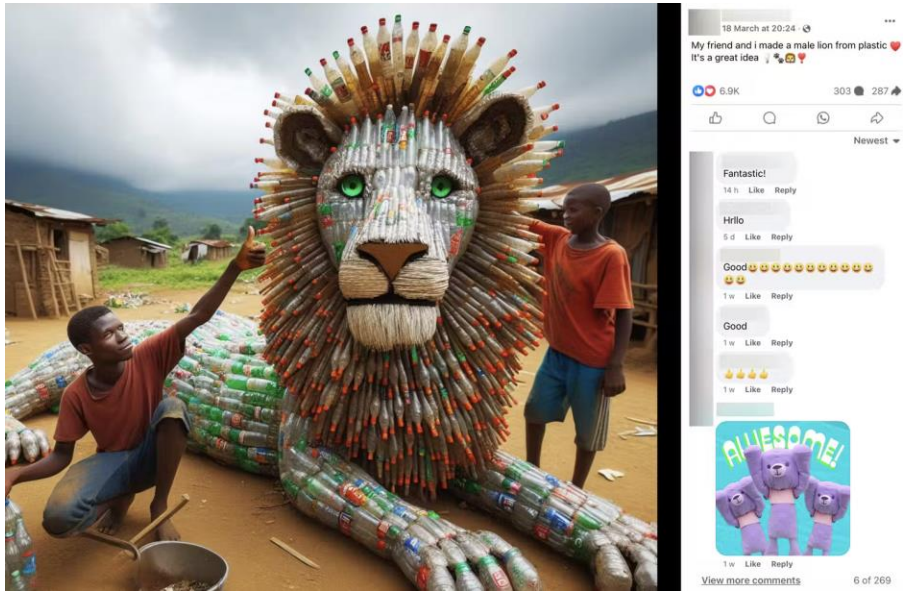


[Source]

Telling LLM from Reality



https://gptzero.me/





[Source]

Concluding thoughts

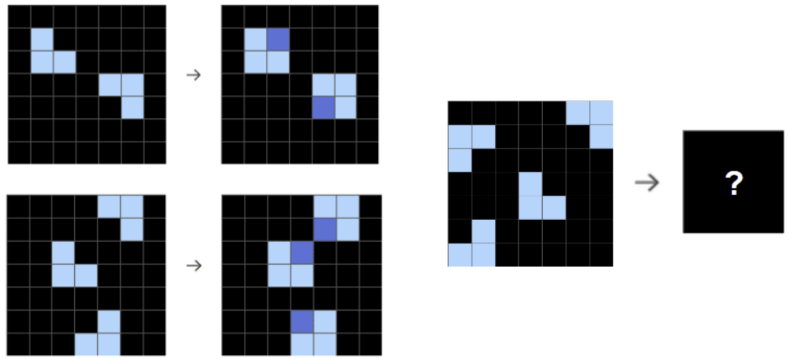
Artificial General Intelligence?



[Source]



[Source]



https://github.com/fchollet/ARC-AGI

What do we always do with new technology?



Thank you!

Questions?



Sacha Davis
Co-op work experience Student
Sacha.Davis@albertahealthservices.ca

Prompt Engineering:

Definition: The practice of designing and optimizing input prompts to elicit the desired responses from an LLM.

Example: Crafting specific questions or statements to get accurate and relevant answers from GPT-3.

Context Window:

The maximum length of text that an LLM can consider at one time while generating or processing language. It determines how much prior text (context) the model can use to inform its next output.

Example: In GPT-3, the context window is 2048 tokens, meaning it can consider up to 2048 tokens of previous text when generating the next part of the text.

Scaling Hypothesis:

The idea that increasing the size of neural networks (more parameters), the amount of training data, and computational resources will continue to improve the performance of AI models.

Example: The success of GPT-3 with 175 billion parameters supports the scaling hypothesis, showing significant improvements over smaller models.

Zero-shot learning:

The ability of an AI model to perform a task without having been explicitly trained on it, by leveraging its general understanding from pre-training.

Example: Asking GPT-3 to generate a poem in a specific style without any prior examples of that style provided during fine-tuning.