

# Scaling Large Language Models

Intro to HLT, December 2023

Daniel Khashabi



The

# The cat

# The cat sat

The cat sat on

The cat sat on \_\_\_\_?\_\_\_\_

The cat sat on the mat.

$P(\text{mat} \mid \text{The cat sat on the})$



next word



context or prefix



# Probability of Upcoming Word

$$\mathbf{P}(X_t | X_1, \dots, X_{t-1})$$

  
next word

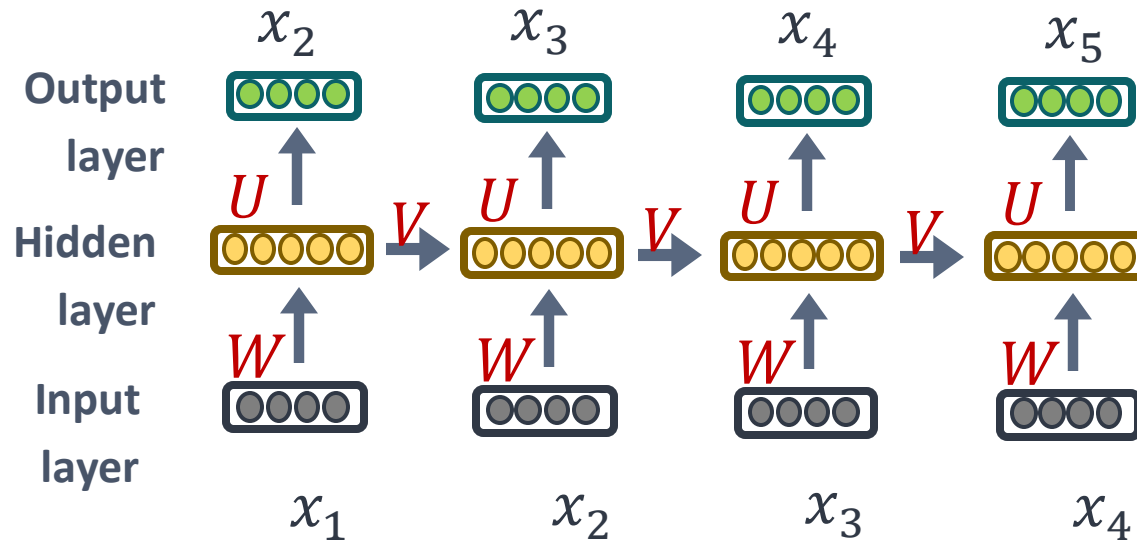
  
context or prefix

# Language Modeling: Motivation

- **Language Models** are **Self-Supervised** or, are trained to be **predictive models** of the world!
- Now the question is: how do you formulate and build them?

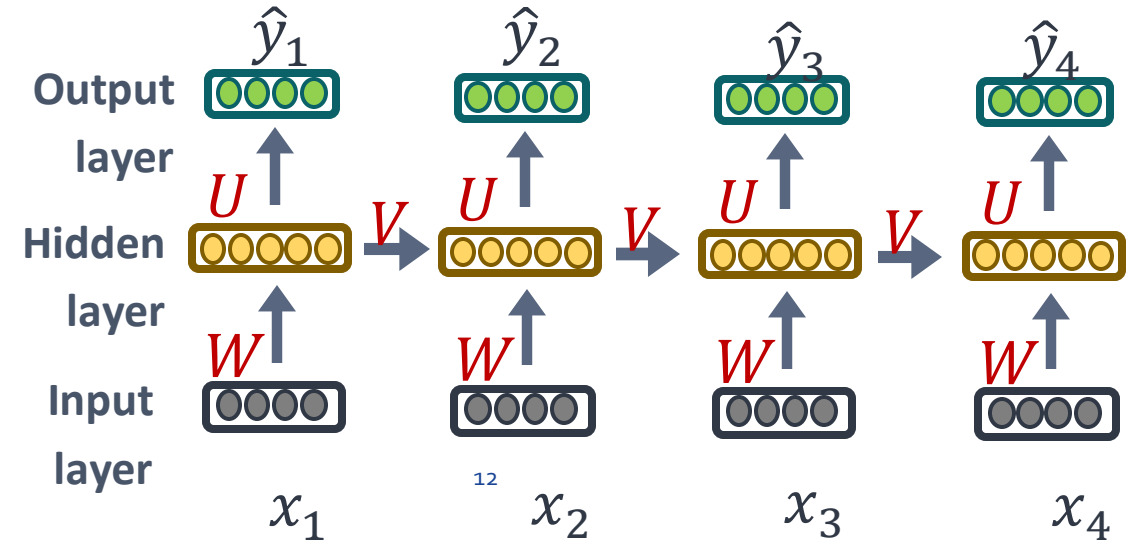
# Terminology: Causal or Auto-regressive Model

## Language Modelling



Auto-regressive

## 1-to-1 tagging/classification



Non-Auto-regressive

# GPT

Generative Pre-trained Transformer

GPT-2: A Big Language Model (2019)

---

**Language Models are Unsupervised Multitask Learners**

---

Alec Radford \*<sup>1</sup> Jeffrey Wu \*<sup>1</sup> Rewon Child<sup>1</sup> David Luan<sup>1</sup> Dario Amodei \*\*<sup>1</sup> Ilya Sutskever \*\*<sup>1</sup>

GPT: An Auto-Regressive LM (2018)

---

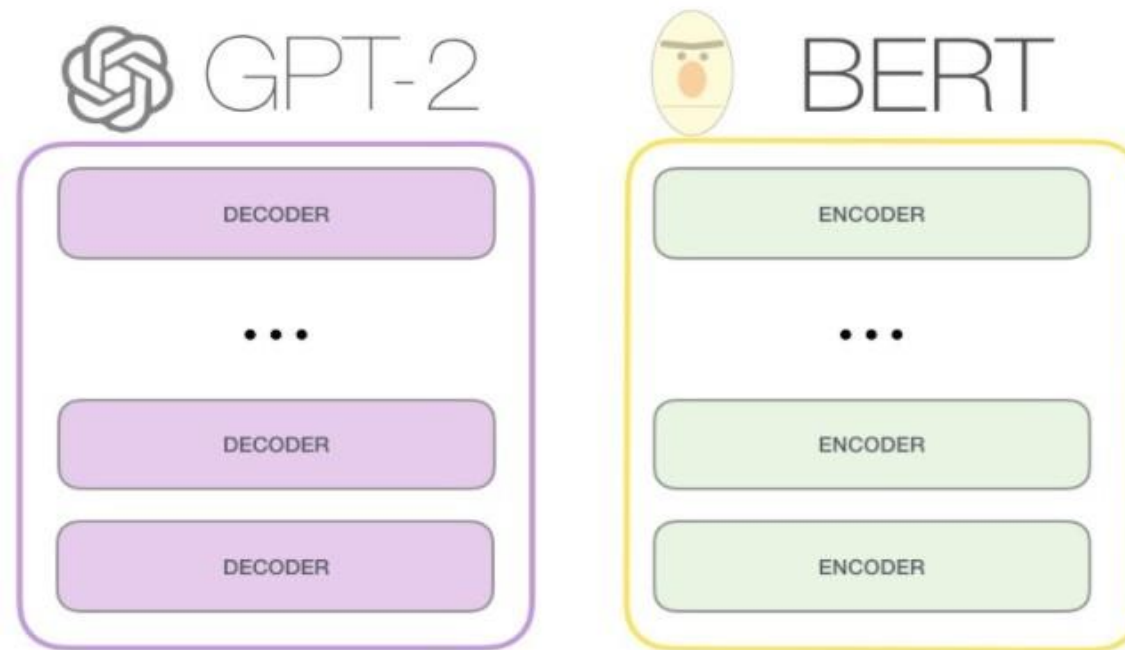
**Improving Language Understanding  
by Generative Pre-Training**

---

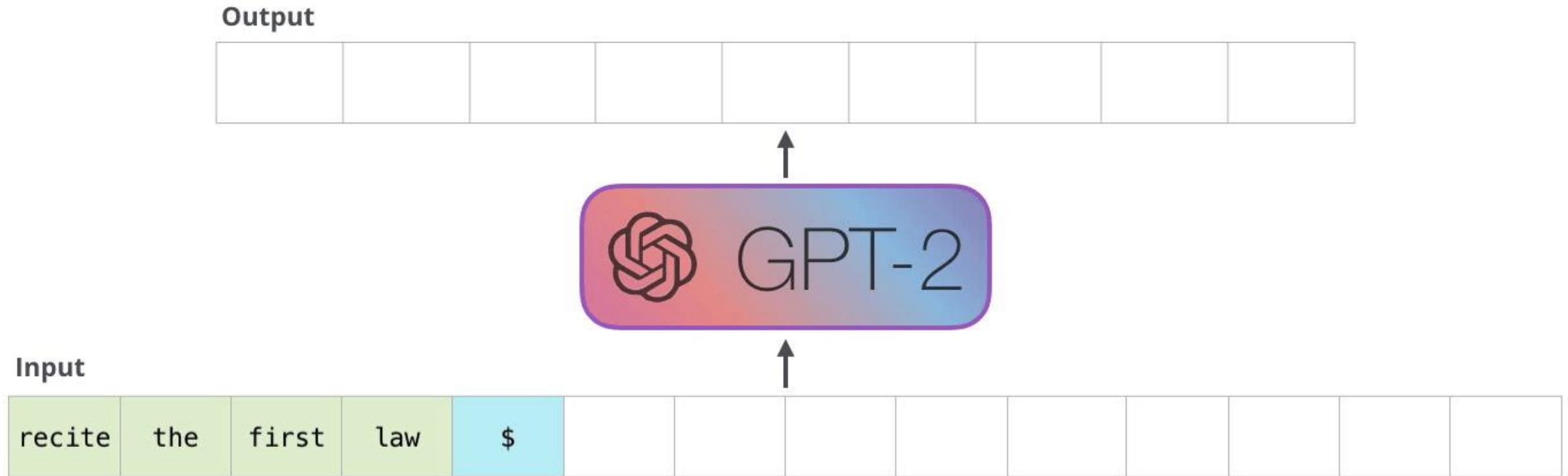
Alec Radford	Karthik Narasimhan	Tim Salimans	Ilya Sutskever
OpenAI	OpenAI	OpenAI	OpenAI
alec@openai.com	karthikn@openai.com	tim@openai.com	ilyasu@openai.com

# GPT-2

- GPT-2 uses only **Transformer Decoders** (no Encoders) to generate new sequences from scratch or from a starting sequence

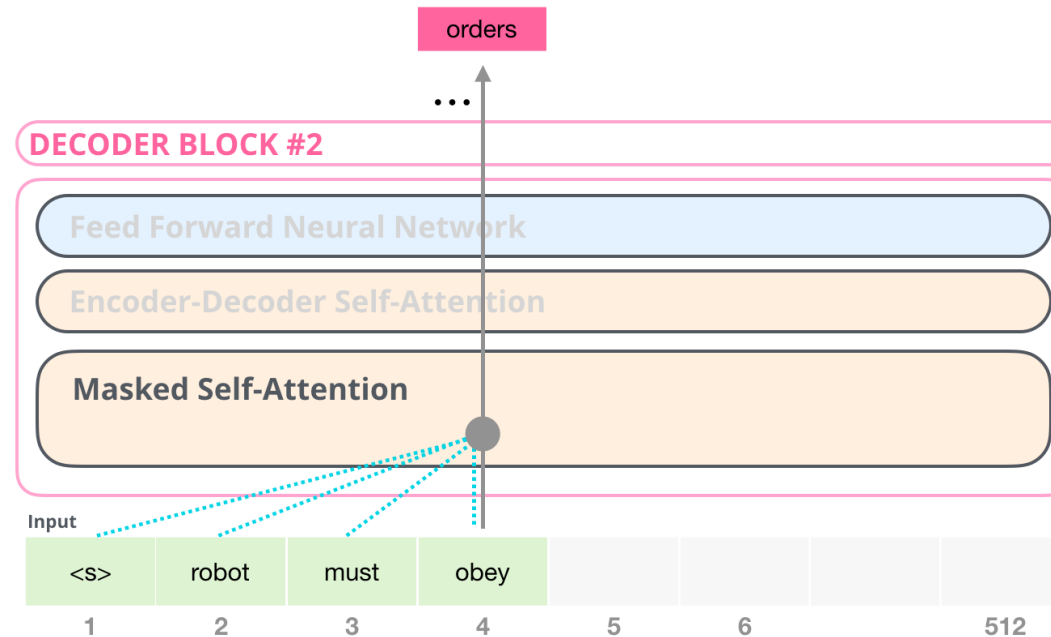


# GPT-2: Next Word Prediction



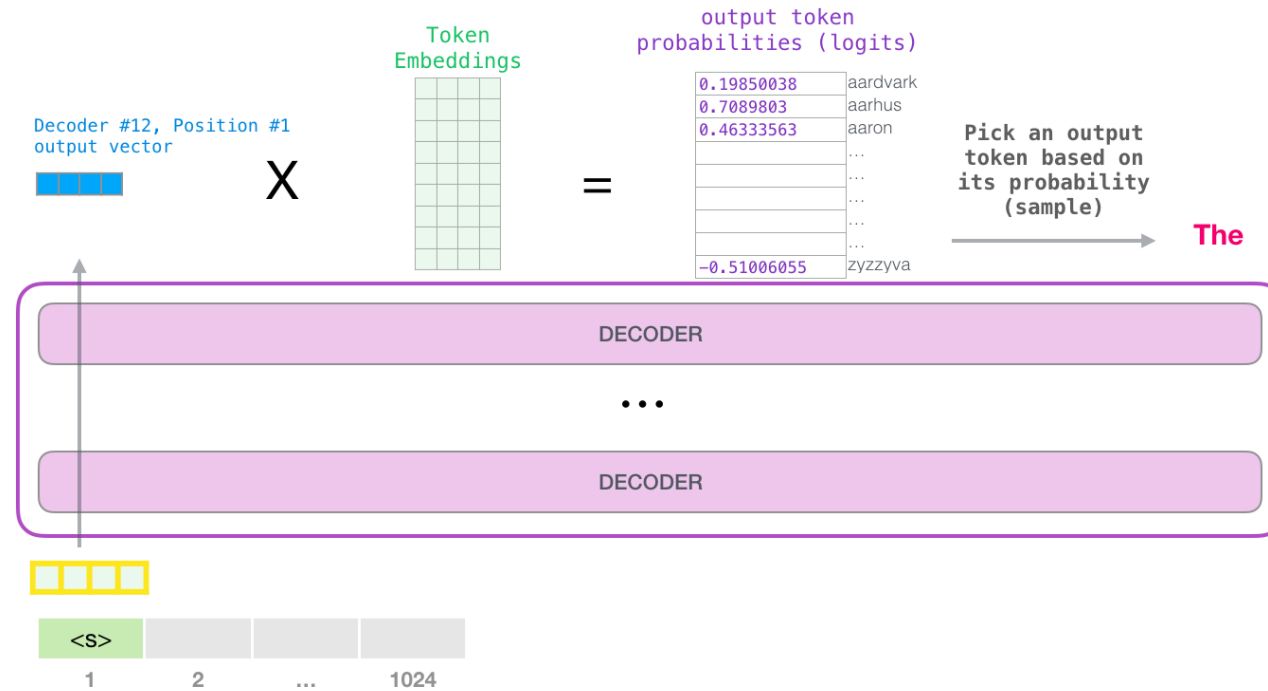
# GPT-2

- As it processes each subword, it masks the “future” words and conditions on and attends to the previous words



# GPT-2

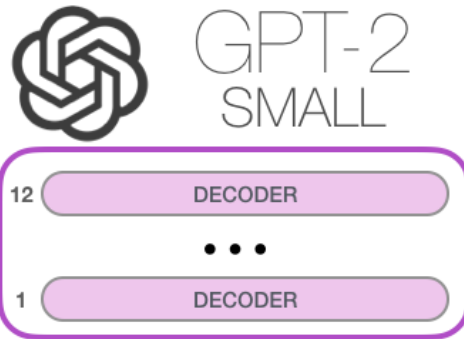
- As it processes each subword, it masks the “future” words and conditions on and attends to the previous words





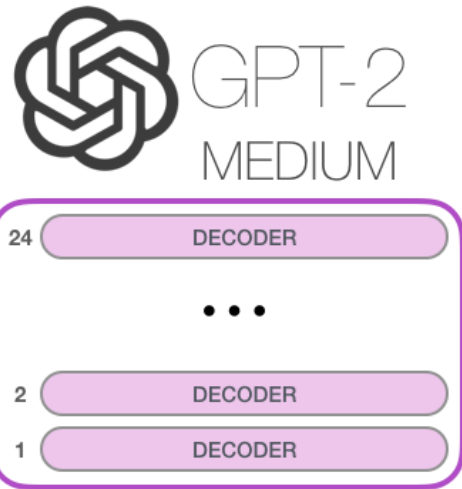
# GPT2: Model Sizes

Play with it here: <https://huggingface.co/gpt2>



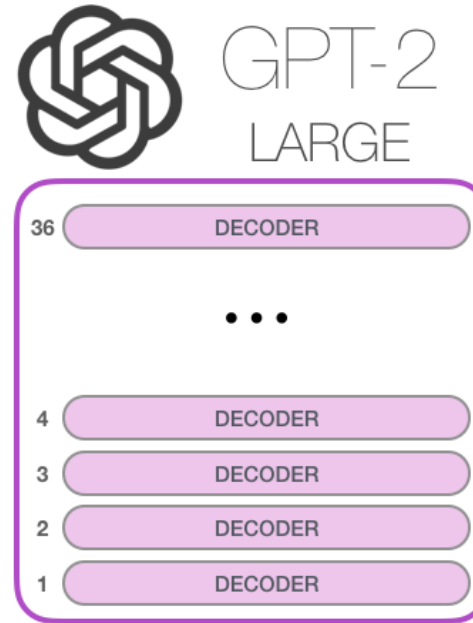
Model Dimensionality: 768

117M parameters



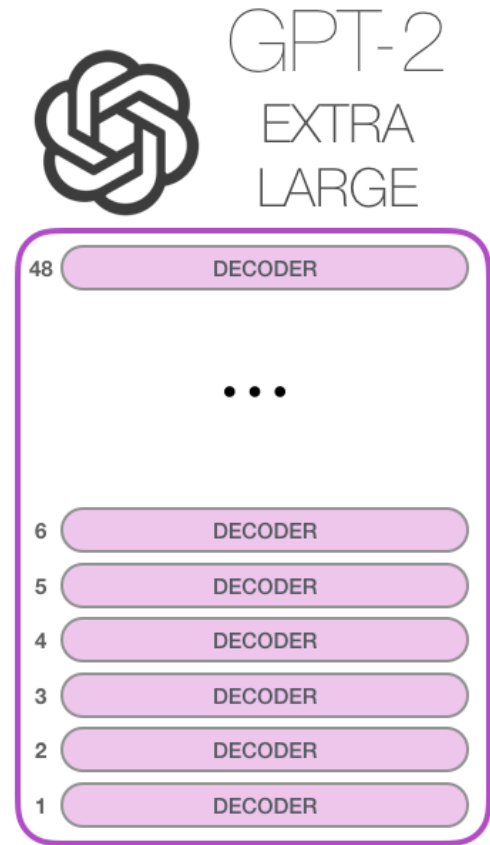
Model Dimensionality: 1024

345M



Model Dimensionality: 1280

762M



Model Dimensionality: 1600

1542M

# GPT2: Some Results

**Language Models are Unsupervised Multitask Learners**

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	56.25	85.7	82.3	39.14	46.54	0.99	1.08	18.3	<b>21.8</b>
117M	<b>35.13</b>	45.99	<b>87.65</b>	<b>83.4</b>	<b>29.41</b>	65.85	1.16	1.17	37.50	75.20
345M	<b>15.60</b>	55.48	<b>92.35</b>	<b>87.1</b>	<b>22.76</b>	47.33	1.01	<b>1.06</b>	26.37	55.72
762M	<b>10.87</b>	<b>60.12</b>	<b>93.45</b>	<b>88.0</b>	<b>19.93</b>	<b>40.31</b>	<b>0.97</b>	<b>1.02</b>	22.05	44.575
1542M	<b>8.63</b>	<b>63.24</b>	<b>93.30</b>	<b>89.05</b>	<b>18.34</b>	<b>35.76</b>	<b>0.93</b>	<b>0.98</b>	<b>17.48</b>	42.16

Table 3. Zero-shot results on many datasets. No training or fine-tuning was performed for any of these results. PTB and WikiText-2 results are from (Gong et al., 2018). Other language model results are from (Dai et al., 2019).

**Article:** Prehistoric man sketched an incredible array of prehistoric beasts on the rough limestone walls of a cave in modern day France 36,000 years ago.

Now, with the help of cutting-edge technology, those works of art in the Chauvet-Pont-d'Arc Cave have been reproduced to create the biggest replica cave in the world.

The manmade cavern named the Caverne du Pont-d'Arc has been built a few miles from the original site in Vallon-Pont-D'arc in Southern France and contains 1,000 painstakingly-reproduced drawings as well as around 450 bones and other features...

Cavemen and women sketched an incredible array of prehistoric beasts on the rough limestone walls of a cave 36,000 years ago and now a replica has been created (pictured)

...

**GPT-2:** The original site in Vallon-Pont-D'arc in Southern France is a Unesco World Heritage site and is the oldest known and the best preserved cave decorated by man. The replica cave was built a few miles from the original site in Vallon-Pont-D'Arc in Southern France. The cave contains images of 14 different species of animals including woolly rhinoceros, mammoths, and big cats.

**Reference:** Cave mimics famous Caverne du Pont-d'Arc in France, the oldest cave decorated by man and the best preserved. The replica contains all 1,000 paintings which include 425 such as a woolly rhinoceros and mammoths. Minute details were copied using 3D modelling and anamorphic techniques, often used to shoot widescreen images. The modern cave also includes replica paw prints of bears, bones and details preserved in the original cave.

# GPT-3: A Very Large Language Model (2020)

- More layers & parameters
- Bigger dataset
- Longer training
- Larger embedding/hidden dimension
- Larger context window



Title: United Methodists Agree to Historic Split  
Subtitle: Those who oppose gay marriage will form their own denomination  
Article: After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.

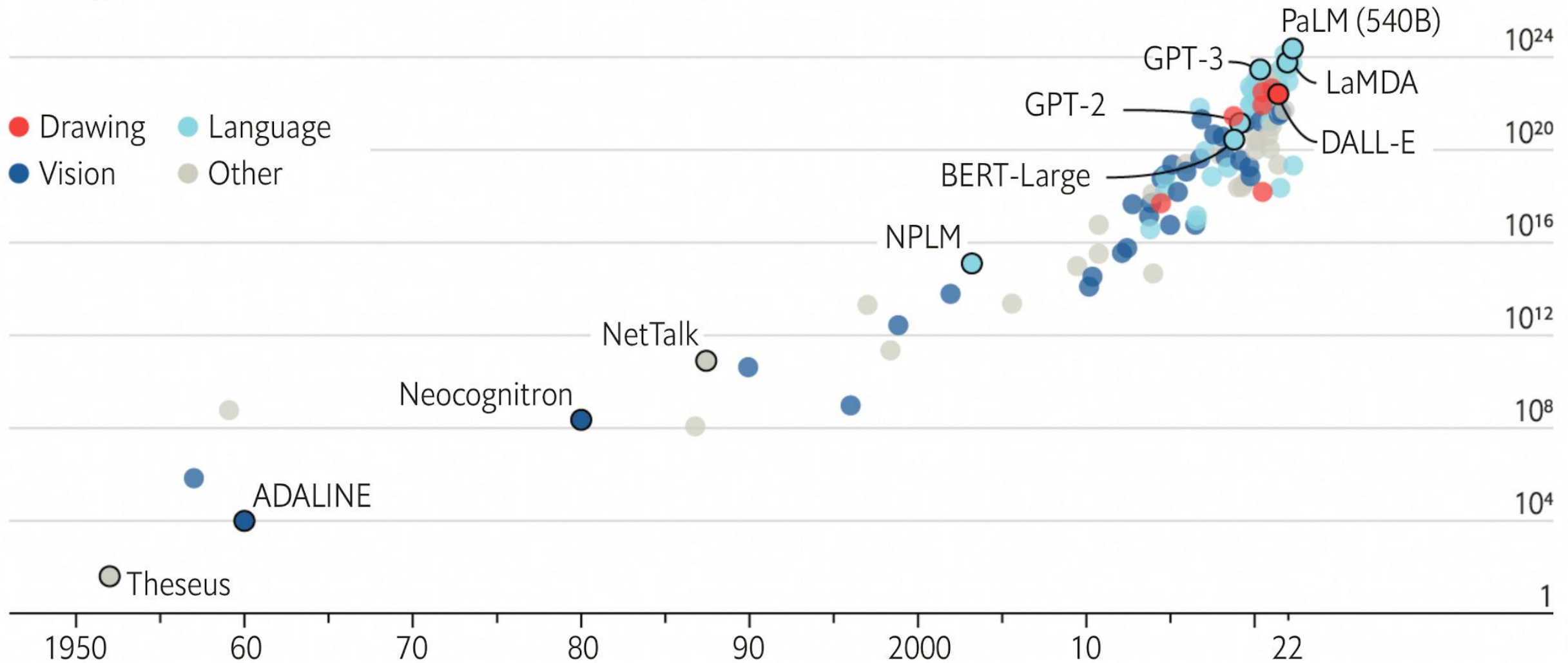
The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church's history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.

**Figure 3.14:** The GPT-3 generated news article that humans had the greatest difficulty distinguishing from a human written article (accuracy: 12%).

# GPT<sub>4</sub>: Try it yourself!

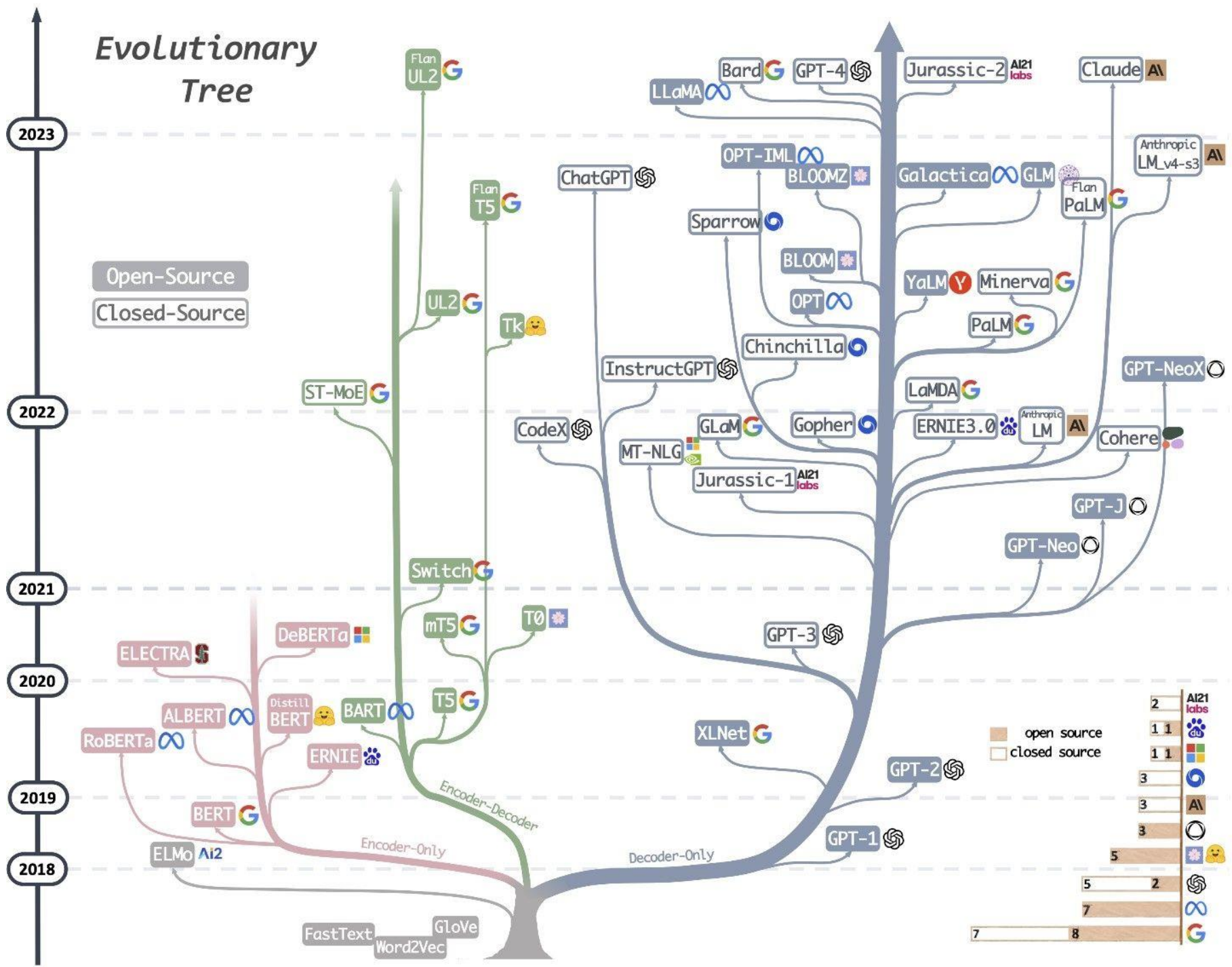
<https://beta.openai.com/playground>

Floating-point operations, selected systems, by type, log scale



Sources: "Compute trends across three eras of machine learning", by J. Sevilla et al., arXiv, 2022; Our World in Data







# Scaling Self-Supervised Models

- Scaling models generally improves their performance!

# Scaling Self-Supervised Models

- Scaling models generally improves their performance!
- Larger pre-training datasets

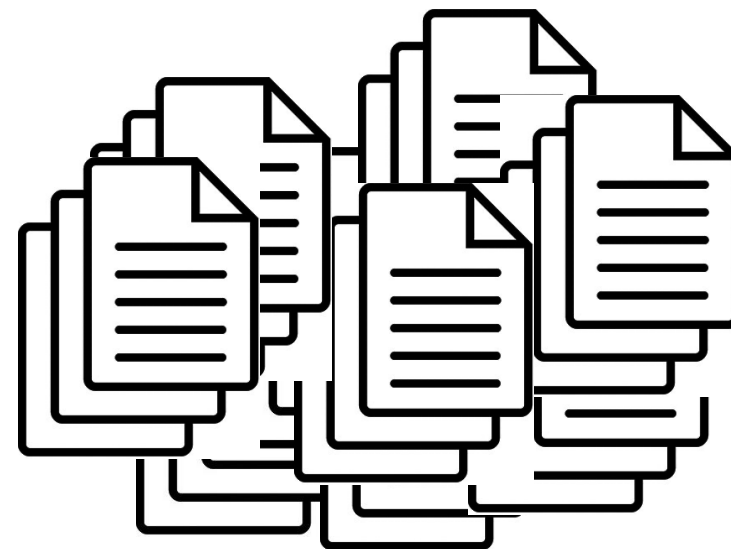
*cheap freely-  
available text*



*1GB*



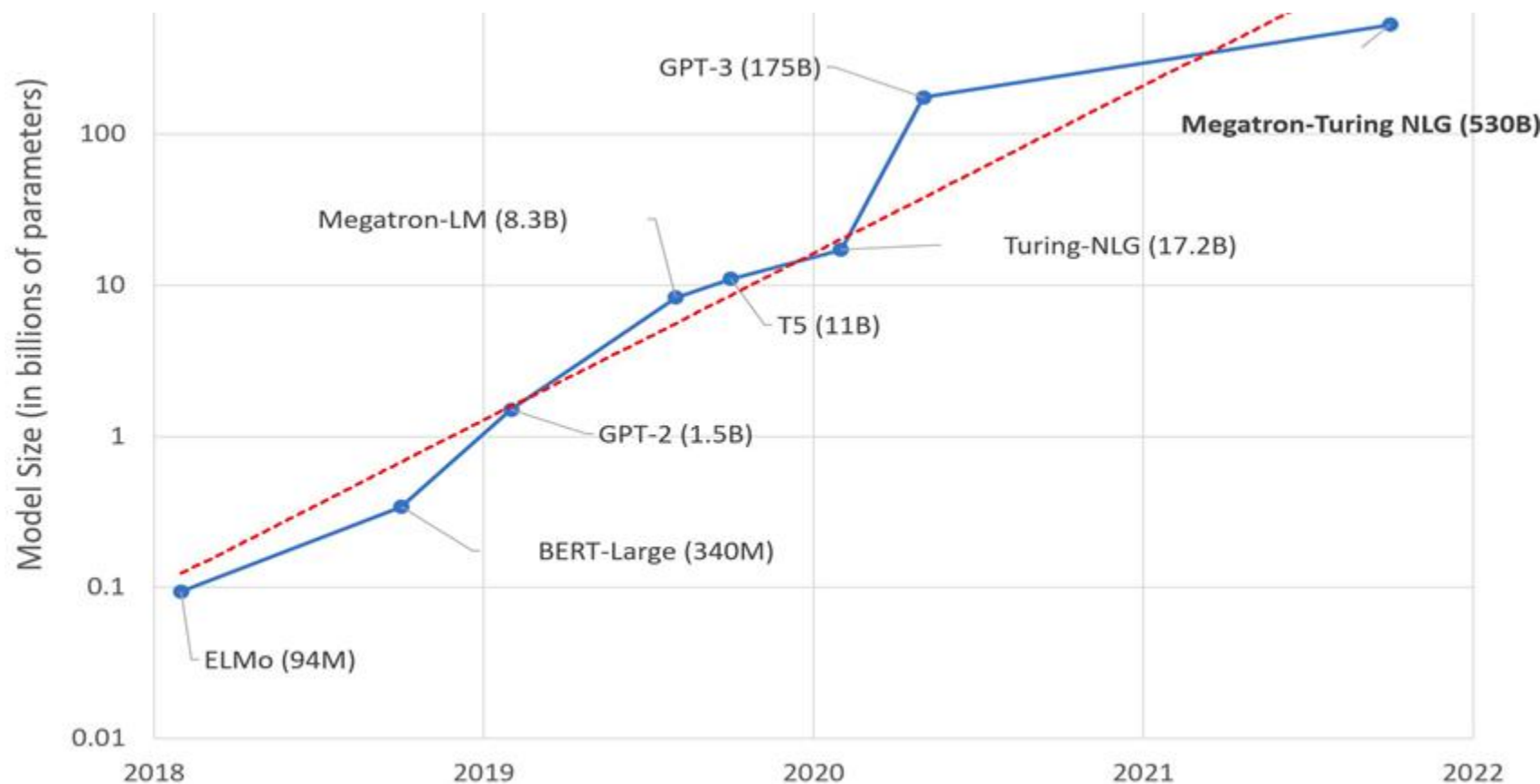
*10GB*



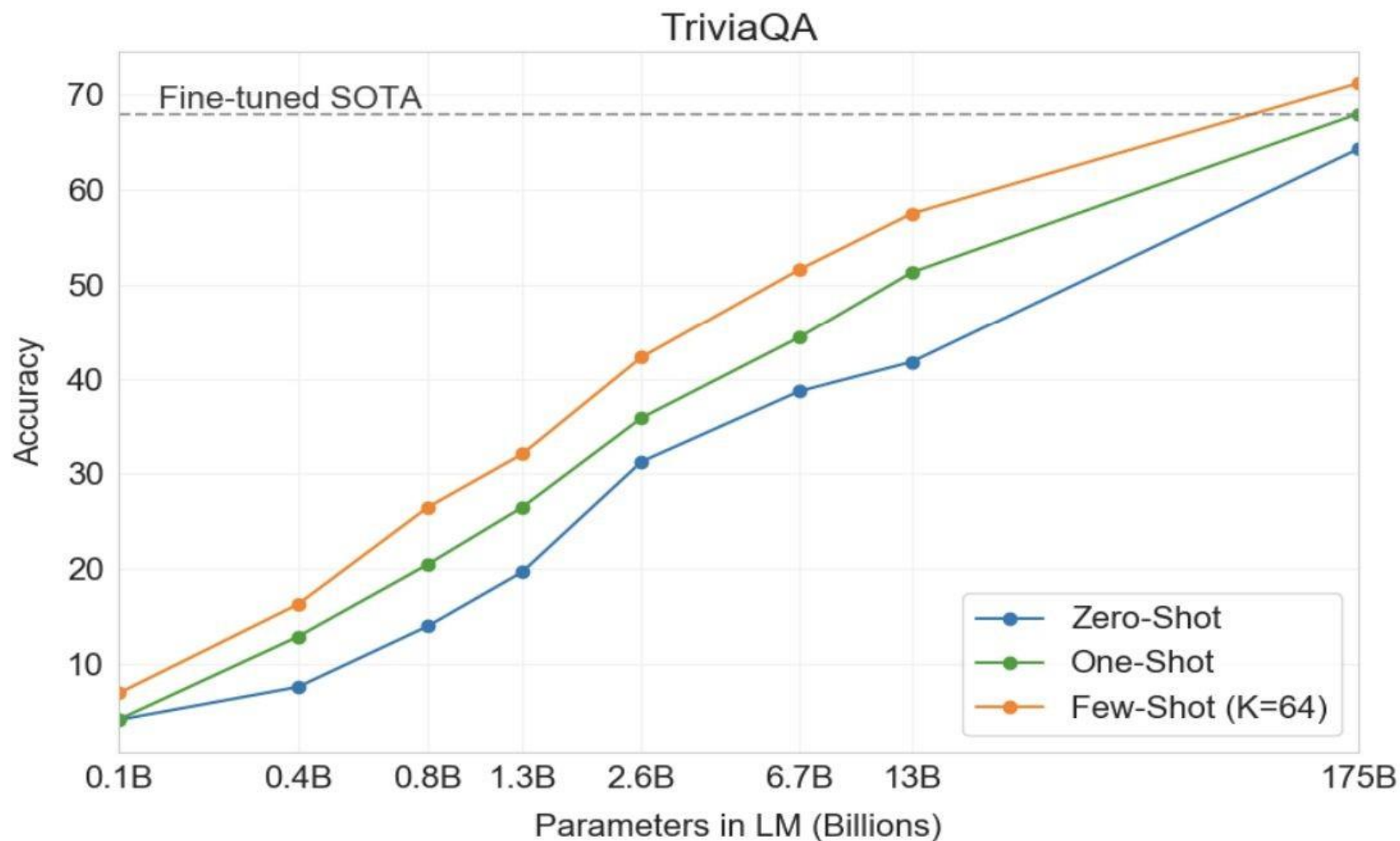
*10TB*

# Scaling Self-Supervised Models

- Scaling models generally improves their performance!
- Larger pre-training datasets
- Larger models



# Scaling Self-Supervised Models



[ [Brown et al. 2020](#). "Language Models are Few-Shot Learners" ]

# ChatGPT: LM optimized for Dialogue

- Built on top of GPT<sub>3</sub>.x and GPT<sub>4</sub>.x
- Fine-tune to respond to interactive dialogue
- Additional technical component: Instruction-tuning/RLHF
  - If you want to learn about it, try CS 6o1.471/671
  - <https://self-supervised.cs.jhu.edu/sp2024/>

"Write George Washington's farewell address as if it was delivered by a pirate"



Ahoy mateys! George Washington here, former captain of the good ship America. I come before ye today to bid ye farewell, as I step down from me post as captain of this great nation.

As ye know, I have been at the helm of this ship for many years, steering it through rough waters and guiding it towards a bright future. But now, it be time for me to hand over the reins to a new captain and retire to me cabin.

But before I do, I wanted to share with ye some wisdom that I have gathered over the years, in the hopes that it will help ye navigate the treacherous seas ahead.

First and foremost, I urge ye to always be vigilant in the defense of our great nation. We have fought hard to earn our freedom and independence, and we must never let it slip through our fingers. Keep a watchful eye on those who would seek to do us harm, and be ready to defend our shores at a moment's notice.

Second, I implore ye to always be mindful of the welfare of our fellow shipmates. We are stronger when we work together, and we must always strive to ensure that all members of our crew are treated fairly and with respect.

VI

Can you fix the bug in this code snippet?

# check if two strings are anagrams

```
def is_anagram(a,b):
    for char in a:
        if char not in b:
            return False
    return True
```



There are a few problems with the code snippet you provided. First, it only checks if all the characters in the first string `a` are also in the second string `b`, but it does not check if the strings have the same length or if the number of occurrences of each character is the same in both strings. To fix these issues, we can use the following updated code:

Copy code

```
# check if two strings are anagrams
def is_anagram(a,b):
    # check if the strings have the same length
    if len(a) != len(b):
        return False

    # create dictionaries to store the number of occurrences of each
    # character in the strings
    a_count = {}
    b_count = {}

    # count the number of occurrences of each character in the first
```

```
        a_count[char] = 1

    # count the number of occurrences of each character in the second
    # string
    for char in b:
        if char in b_count:
            b_count[char] += 1
        else:
            b_count[char] = 1

    # check if the number of occurrences of each character is the same in
    # both strings
    for char in a_count:
        if char not in b_count or a_count[char] != b_count[char]:
            return False

    # if all checks pass, the strings are anagrams
    return True
```

This updated code first checks if the two strings have the same length. If not, the strings cannot be anagrams and the function returns `False`. Next, it creates two dictionaries to store the number of occurrences of each character in the two strings. It then counts the number of occurrences of each character in each string, and finally checks if the number of occurrences of each character is the same in both strings. If all these checks pass, the function returns `True`, indicating that the two strings are anagrams.

Is Scale All We Need?



# Is Scale All We Need?

For what purpose?

- For many purposes (answering simple questions, translating simple sentences) we already have good models. **Not our focus.**
- Let's use this as a goal:

General intelligence:

- Engaging in conversation with humans to solve a variety of problems
- Work in a variety of contexts/domains
- Works as good as the best expert human in each domain



Nando de Freitas   
@NandoDF



Someone's opinion article. My opinion: It's all about scale now! The Game is Over! It's about making these models bigger, safer, compute efficient, faster at sampling, smarter memory, more modalities, INNOVATIVE DATA, on/offline, ... 1/N



thenextweb.com

DeepMind's new Gato AI makes me fear humans will never achieve AGI  
DeepMind just unveiled a new AI system called Gato that makes OpenAI's GPT-3 look like a child's toy. But are we any closer to AGI?

4:46 AM · May 14, 2022 · Twitter for iPhone

233 Retweets 114 Quote Tweets 885 Likes



Nando de Freitas   
@NandoDF



Solving these scaling challenges is what will deliver AGI. Research focused on these problems, eg S4 for greater memory, is needed. Philosophy about symbols isn't. Symbols are tools in the world and big nets have no issue creating them and manipulating them 2/n

4:50 AM · May 14, 2022 · Twitter for iPhone

23 Retweets 5 Quote Tweets 153 Likes

Let's do a poll!

# Is Scale All We Need?

1. Is scale the/a right “hill to climb”?
2. Even if it is a right “hill” is it feasible/practical to climb this hill? (there might be other “hills” too).

# What is “Scale”?

- Compute?
- Data?
- Information?
- Effective compression of information?

# Is Scale All You Need?

- Compute?
- Data?
- Information?
- Effective compression of information?

# Argument: **Not** Enough Compute for Scaling

- There is simply **not enough compute available**.
  - Models have been increasing **10x** every year
  - Moore's law: # of transistors on an IC **doubles** about every two years.
  - There are physical limits to how much faster computers can get.
- Even if we have the compute, scaling the compute will be quite **costly**.
- Scaling compute is simply infeasible. [QED]

**Let's do a poll!**

# Argument **Against** “Not Enough Compute”

- On insufficiency of compute resource:
  - Hardware technologies continue to progress at a rapid pace.
  - Huang’s law: advancements in GPUs happen at **much faster** rate than what Moore predicted.
  - So much potentials in **parallel computing**.
- On cost-[in]efficiency of scaling:
  - While models like GPT3 cost a lot (monetary or otherwise), their availability prevent training MANY smaller, mediocre models.
  - Therefore, it might be that the net cost of scaling large models is negative.
    - It is the case within Microsoft according to its CTO, Kevin Scott.

# Is Scale All You Need?

- Compute?
- Data?
- Information?
- Effective compression of information?



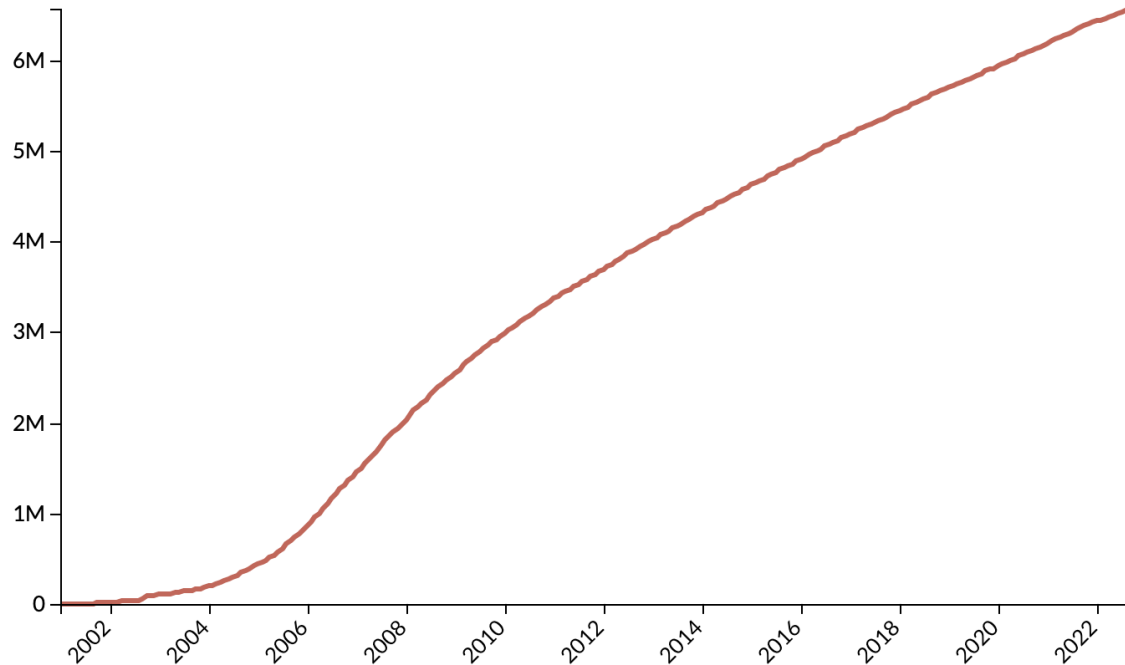
# Not Enough Data for Scaling

- Hoffmann et al showed that, to be compute-optimal, **model size and training data must be scaled equally**.
- It shows that existing LLMs are severely data-starved and under-trained.
- Given the new scaling law, even if you pump a billions of params into a model, the gains will **not** compensate for more training tokens.
- There is simply not enough [language] data. [QED]

**Let's do a poll!**

# Argument Against “Not Enough Data” (1)

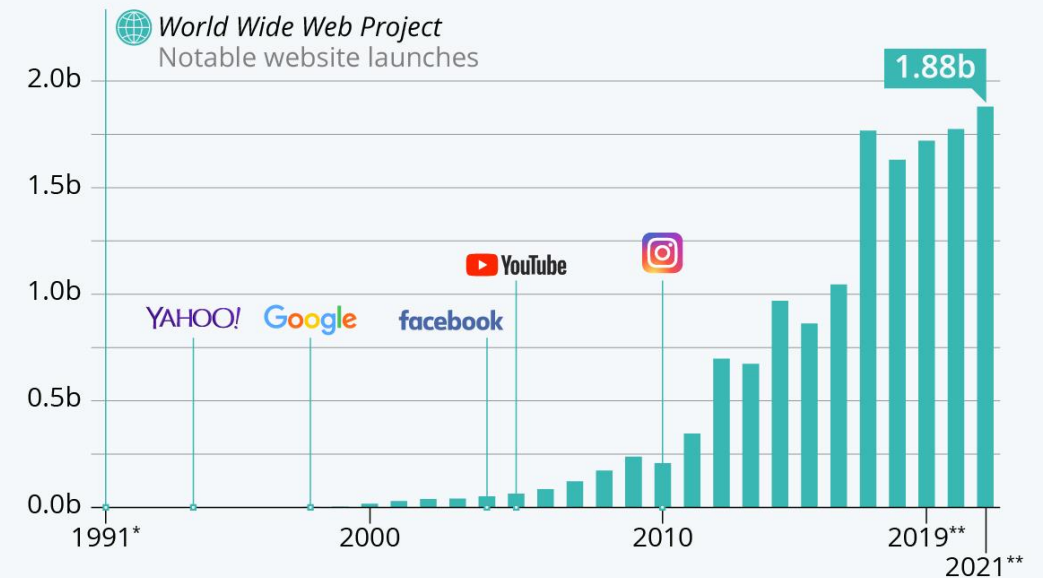
- Data is growing exponentially (?)



Wikipedia size

## How Many Websites Are There?

Number of websites online from 1991 to 2021



\* As of August 1, 1991.

\*\* Latest available data for 2019: October 28, for 2020: June 2, for 2021: August 6.

Source: Internet Live Stats

# Argument **Against** “Not Enough Data” (2)

- You can harness data from **other modalities**.
  - For example, to get more text data we can build a solid speech processor model that converts speech to text.
  - (aside: more than 80% if internet traffic is video)

## **SKYQUEST**

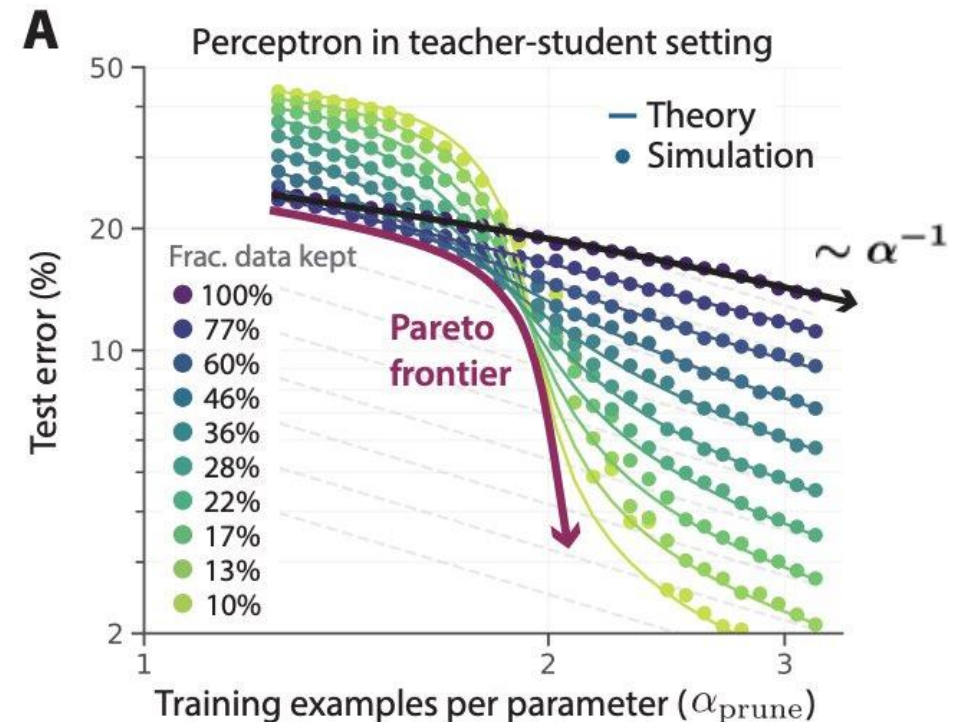
Global Online Video Platforms Market Drives over 80% of Total Internet Traffic |  
Skyquest Technology

- (aside2: is that why OpenAI built Whisper?!)

["Robust speech recognition via large-scale weak supervision." Radford+ 2022]

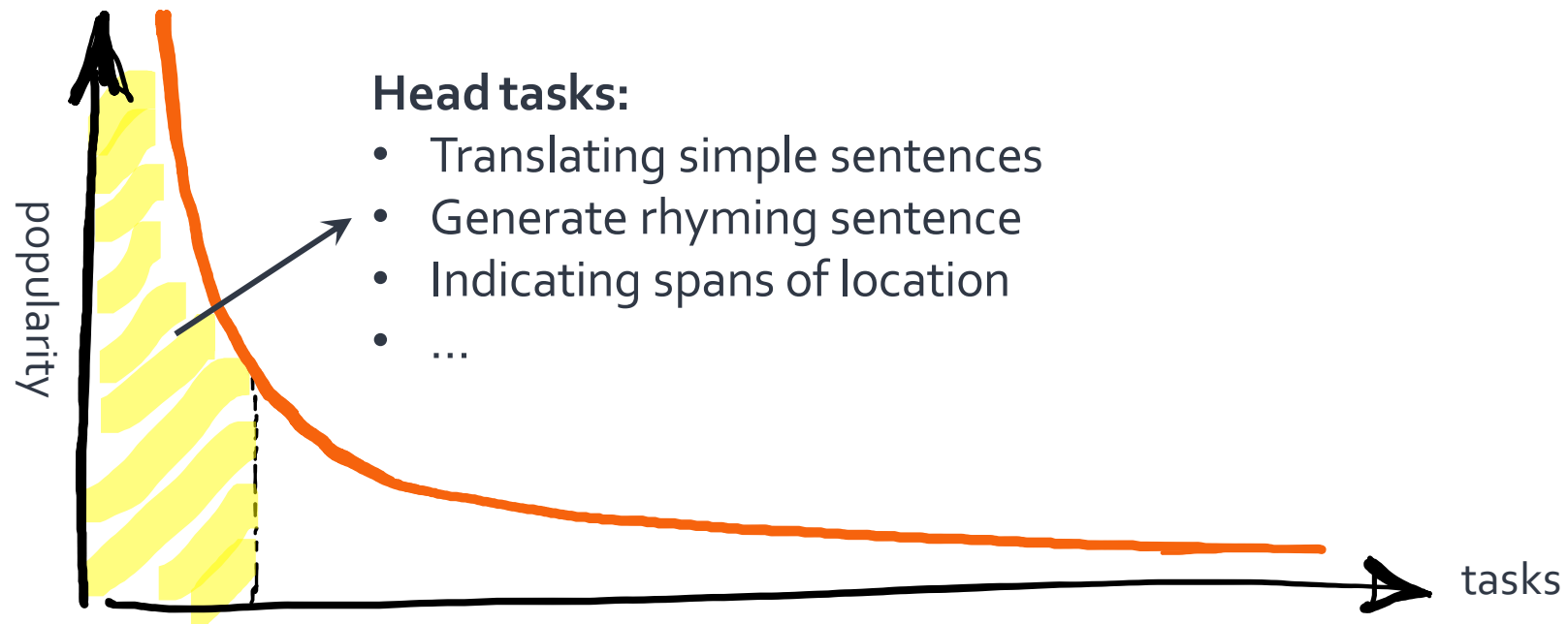
# Argument **Against** “Not Enough Data” (3)

- You can use data more effectively.
- Sorscher et al. lays out recipes to achieve \*exponential\* scaling instead through statistical mechanics theory.
- Carefully curating a small subset goes a long way!



# Scale is **Not** all You Need Because of Tail Phenomena

- Tail phenomena will never go away!

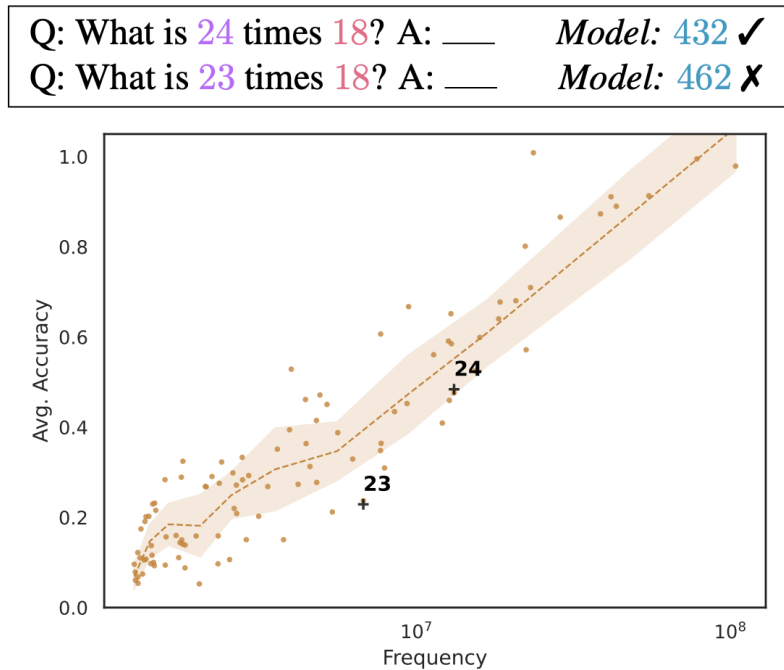


# Scale is **Not** all You Need Because of Tail Phenomena

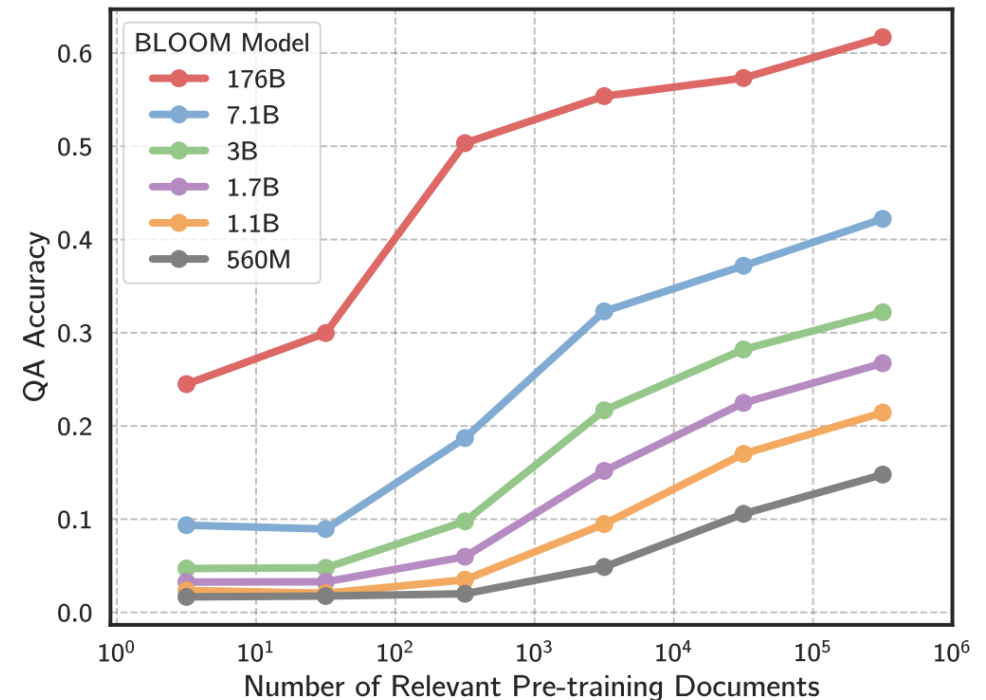
- Tail phenomena will never go away!



# Scale is **Not** all You Need Because of Tail Phenomena



**Figure 1. Multiplication Performance:** Plot of GPT-J-6B's 2-shot accuracy on multiplication (averaged over multiple multiplicands and training instances) against the frequency of the equation's first term in the pretraining corpus. Each point represents the average performance for that term (e.g., 24) multiplied by numbers 1-50 and 5 choices of random seeds. As in the example, the performance difference for the numbers 24 and 23 is more than 20%. We find a strong correlation between accuracy and frequency.



**Figure 1: Language models struggle to capture the long-tail of information on the web.** Above, we plot accuracy for the BLOOM model family on TriviaQA as a function of how many documents in the model's pre-training data are relevant to each question.

# Scale is **Not** all You Need Because of Tail Phenomena

- Tail phenomena will never go away!
- Will result in brittleness to small changes

“Tesla's Autopilot system confusing horse-drawn carriage for truck”







which number is bigger, 999934 or 32323232332?



The number 32323232332 is larger than 999934. To compare the sizes of two numbers, we need to look at the number of digits in each number. In this case, the number 32323232332 has 10 digits, while the number 999934 has 6 digits. Because the number with more digits is always larger, we can conclude that 32323232332 is the larger number in this case.



which number is bigger, -999934 or -32323232332?



The number -32323232332 is larger than -999934. When comparing negative numbers, we follow the same rules as when comparing positive numbers. In this case, the number -32323232332 has 10 digits, while the number -999934 has 6 digits. Because the number with more digits is always larger, we can conclude that -32323232332 is the larger number in this case.



It is worth noting that negative numbers are always less than positive numbers, so even though -32323232332 is larger than -999934, it is still smaller than 999934. This is because a negative number represents a value that is less than zero, while a positive number represents a value that is greater than zero.

TU

how many more legs does a spider have than an ant?

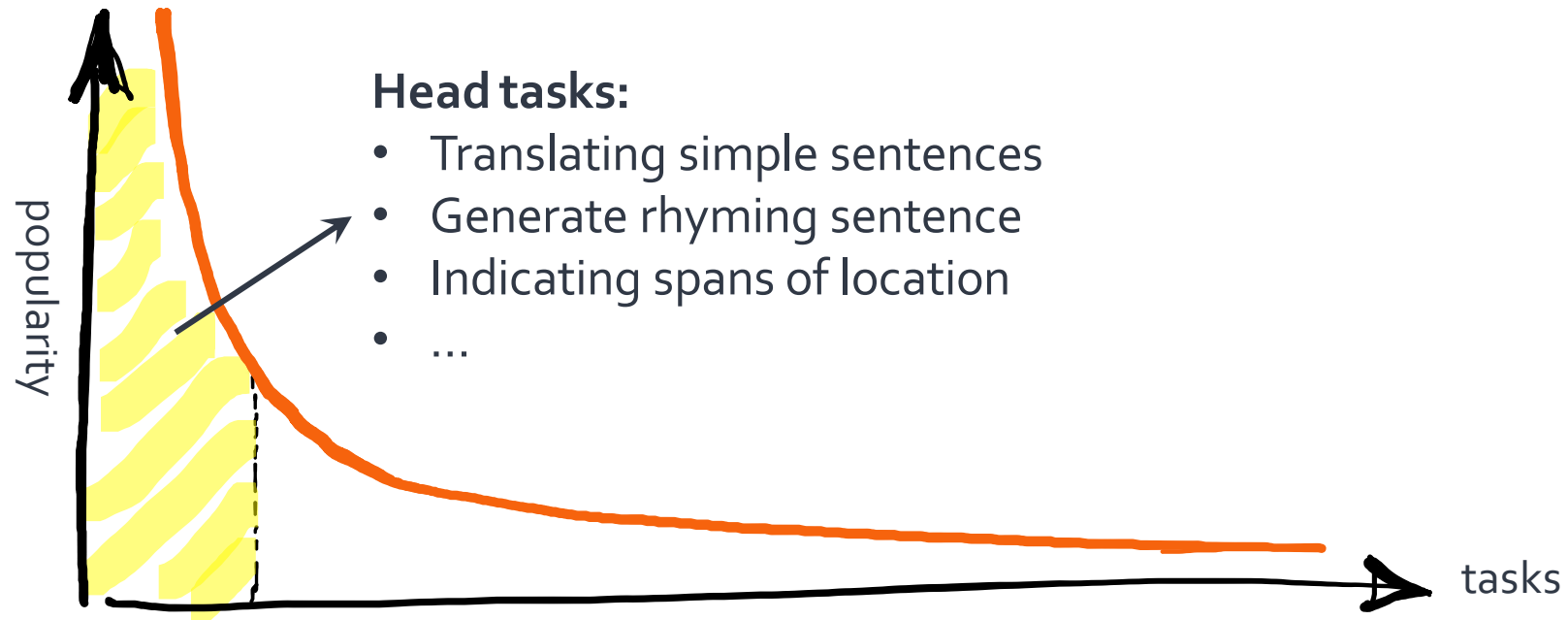


A spider has eight more legs than an ant. A spider typically has eight legs, while an ant typically has six legs. The number of legs can vary depending on the species of spider or ant, but in general, spiders have eight more legs than ants. The legs of a spider are used for walking, running, climbing, and spinning webs, and they enable the spider to move quickly and efficiently in different environments and terrains. On the other hand, the legs of an ant are used for walking, running, climbing, and digging, and they enable the ant to move quickly and efficiently in different environments and terrains. The additional legs of a spider allow it to have a wider range of movements and behaviors, and to adapt to different situations and challenges.

# Scale is **Not** all You Need Because of Tail Phenomena

- Hence, scale won't solve the tail phenomena. [QED]

Let's do a poll!



Given all these arguments, what do you think now?

Is Scale All We Need?

# Putting it All Together

- “Scaling” is here is here to stay for now.
  - There is plenty of data out there that we haven’t been able to use.
  - There is no sign of development in hardware technology.
- The long tail poses a serious challenge:
  - It’s possible that scaling will continue to yield gains in the long-tail, but unlike to solve it.
  - It is unclear if there are better ways of solving long tail phenomena.
- Unclear “scale” can say about other aspect of intelligence
  - coordination, cooperation, communication, etc.

# Augmenting Humans vs. Replacing Them

- It's unlikely that "scaling" will lead to complete reliable/accountable models.
- Hence, the foreseeable future is about human-AI loop.
- It is more productive to think of them as "amplifiers" of human abilities.
  - A calculator does not replace engineers, but it enables them.

Thank you!

# How does the future look like to you?

## Which future will we have?

1. One very large model
2. Few very large models
3. Many vey large models



05

Who won the 2020 presidential election?



The winner of the 2020 presidential election has not yet been determined. The election took place on November 3, 2020, and the results are still being counted and verified. It is important to note that the winner of the election will not be officially declared until all of the votes have been counted and the results have been certified by the appropriate authorities.

As of December 1, 2022, the 2020 presidential election is still ongoing and its outcome is not yet known. It is important to wait for the official results before