

Is Scale All You Need?

Intro to HLT, December 2022

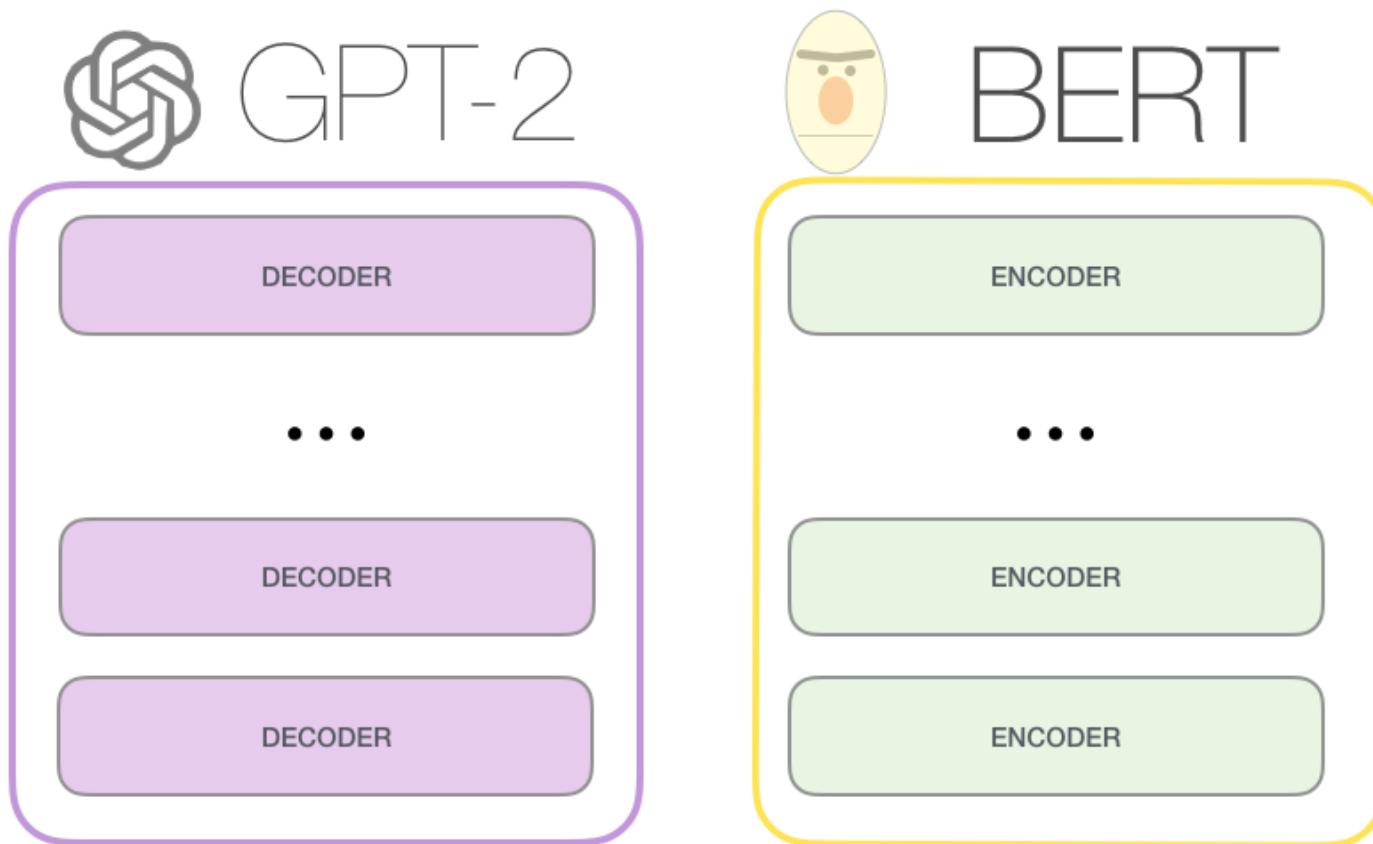
Daniel Khashabi



But first ...

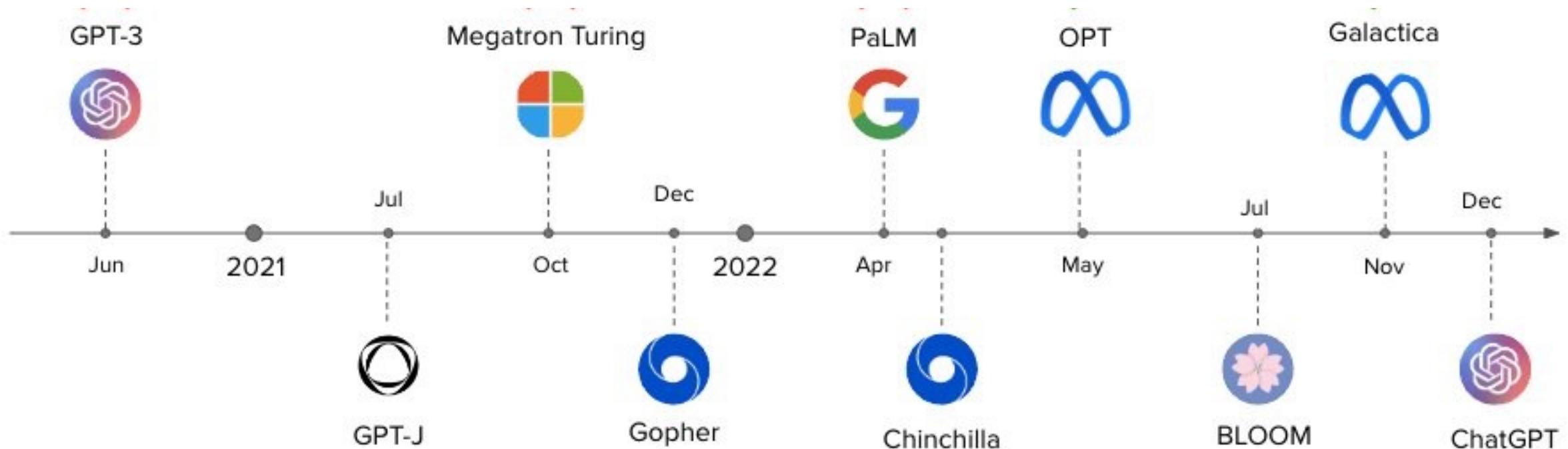
- Open this link on your phone: pollev.com/danielkhashabi994
- We will use this for collecting your votes!

Large Language Models



[Bengio et al. '04, Peters et al. '18, Raffel et al. '20, Brown et al. '20, many others]

Large Language Models since GPT3



[Slide from Nazneen Rajani]

Scaling Self-Supervised Models

- Scaling models generally improves their performance!

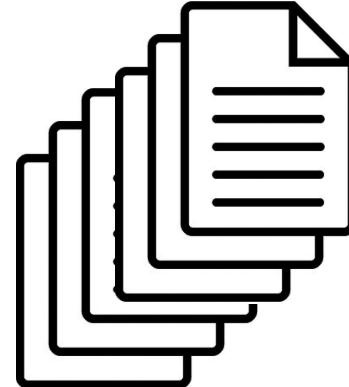
Scaling Self-Supervised Models

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

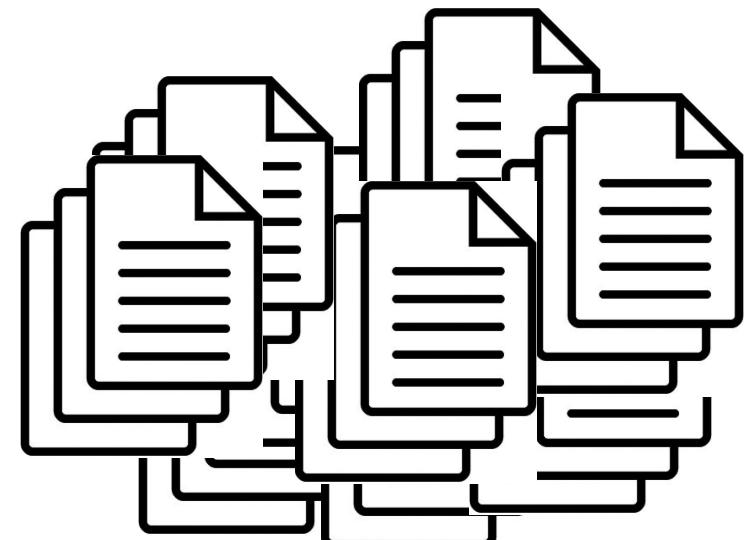
cheap freely-available text



1GB



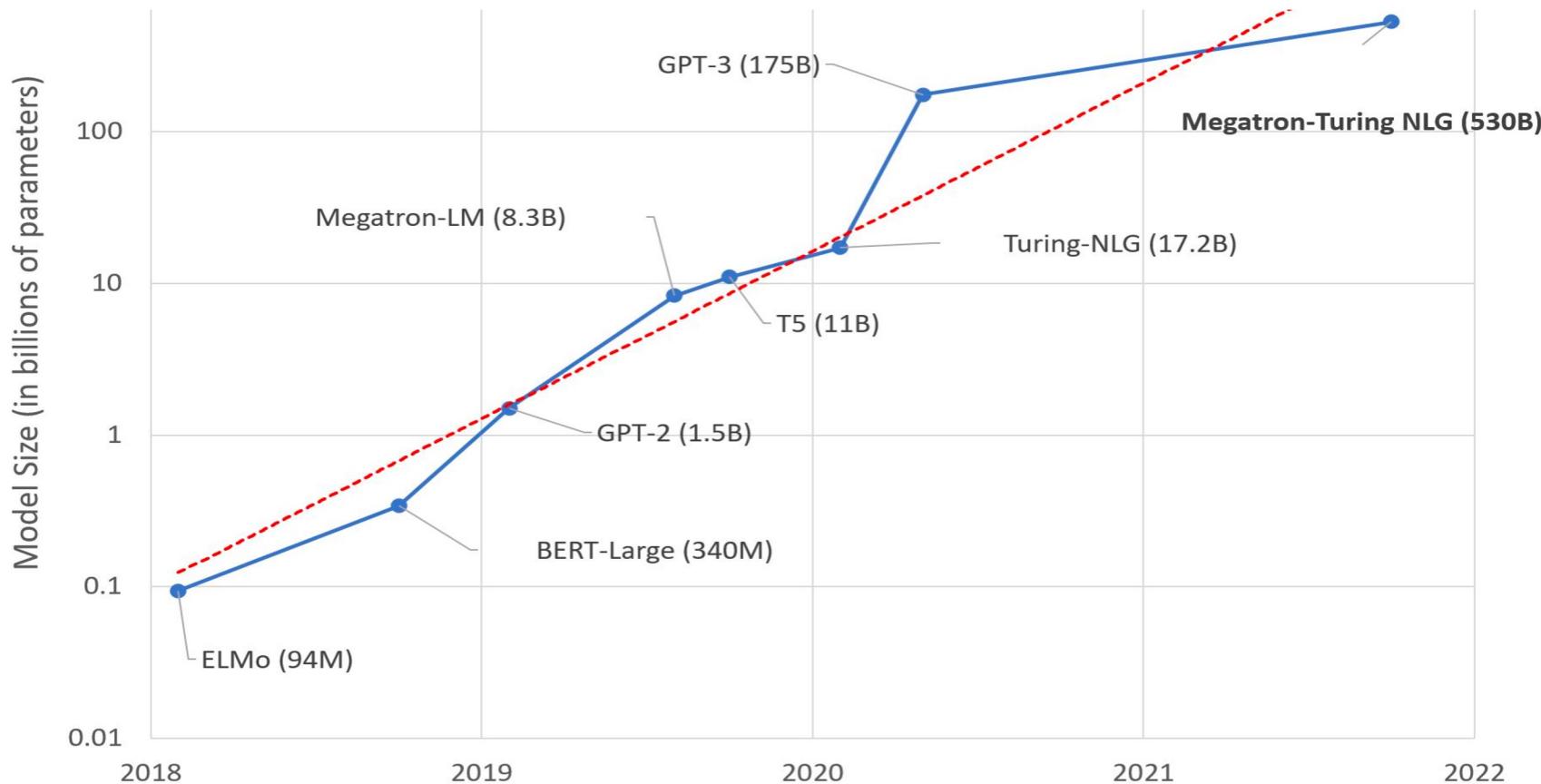
10GB



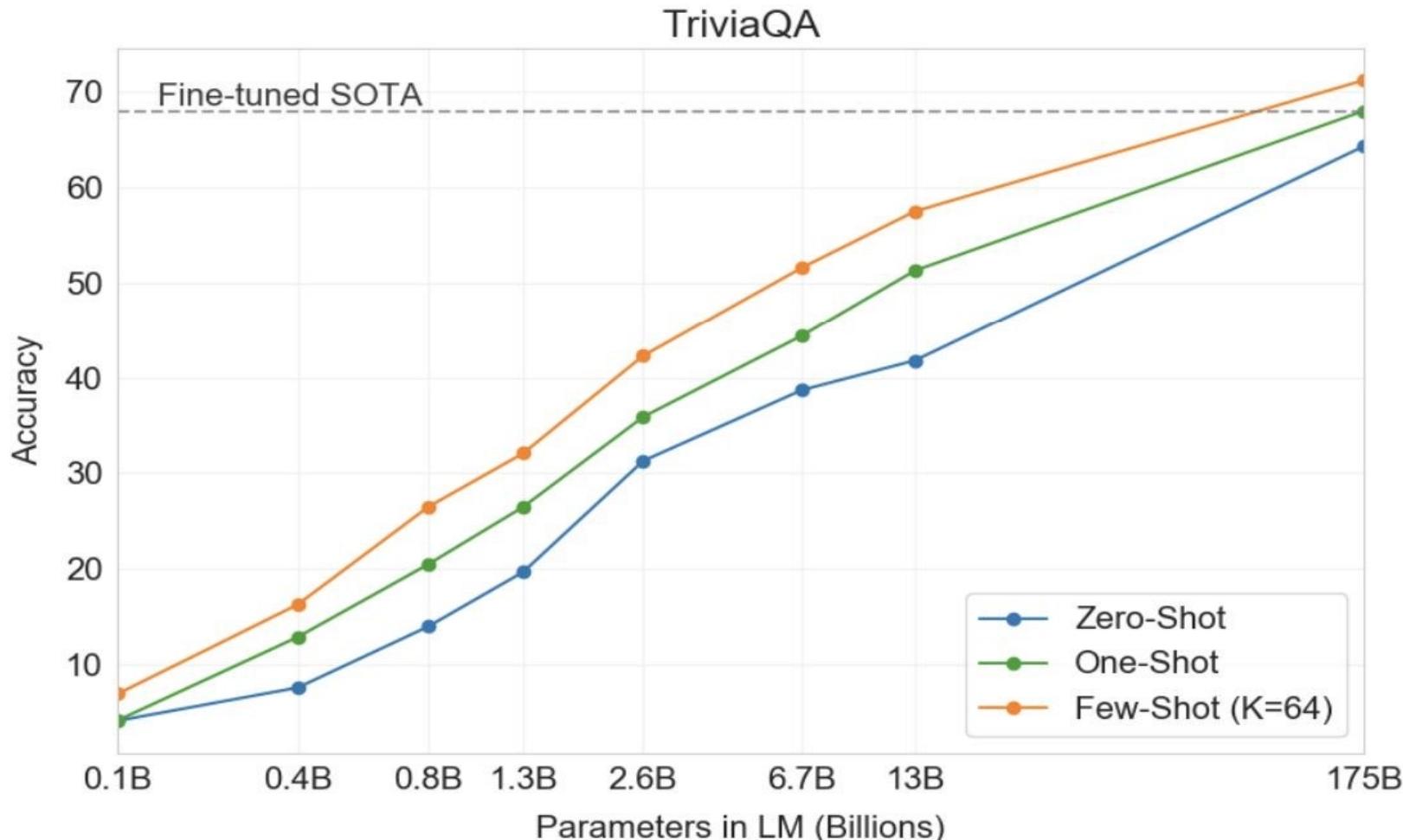
10TB

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

- Days ago
- 170B model
- Built on top of GPT3.x
- Fine-tune to respond to interactive dialogue



Do you think a magnet could attract a penny?



Yes, a magnet could potentially attract a penny if the penny is made of a magnetic material. However, pennies made in the United States after 1982 are made primarily of zinc, which is not magnetic. Therefore, a magnet would not be able to attract a penny made after 1982.



What if it's made of copper?



Copper is not a magnetic material, so a magnet would not be able to attract a penny made of copper. Copper is a good conductor of electricity, which is why it is often used in electrical wiring. However, it does not have magnetic properties.



"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?

What is “Scale”?

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

Argument: Not Enough Compute for Scaling

Limitations regarding compute:

- 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.

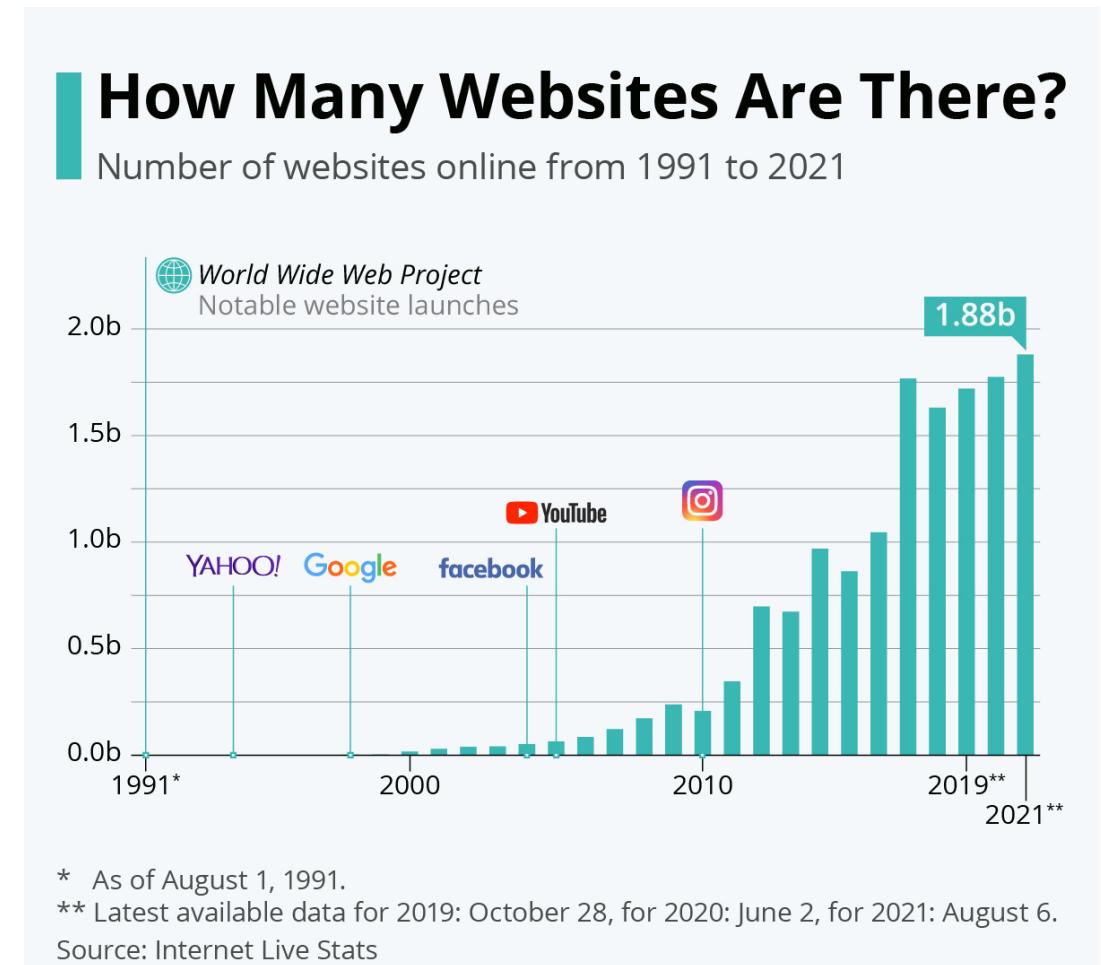
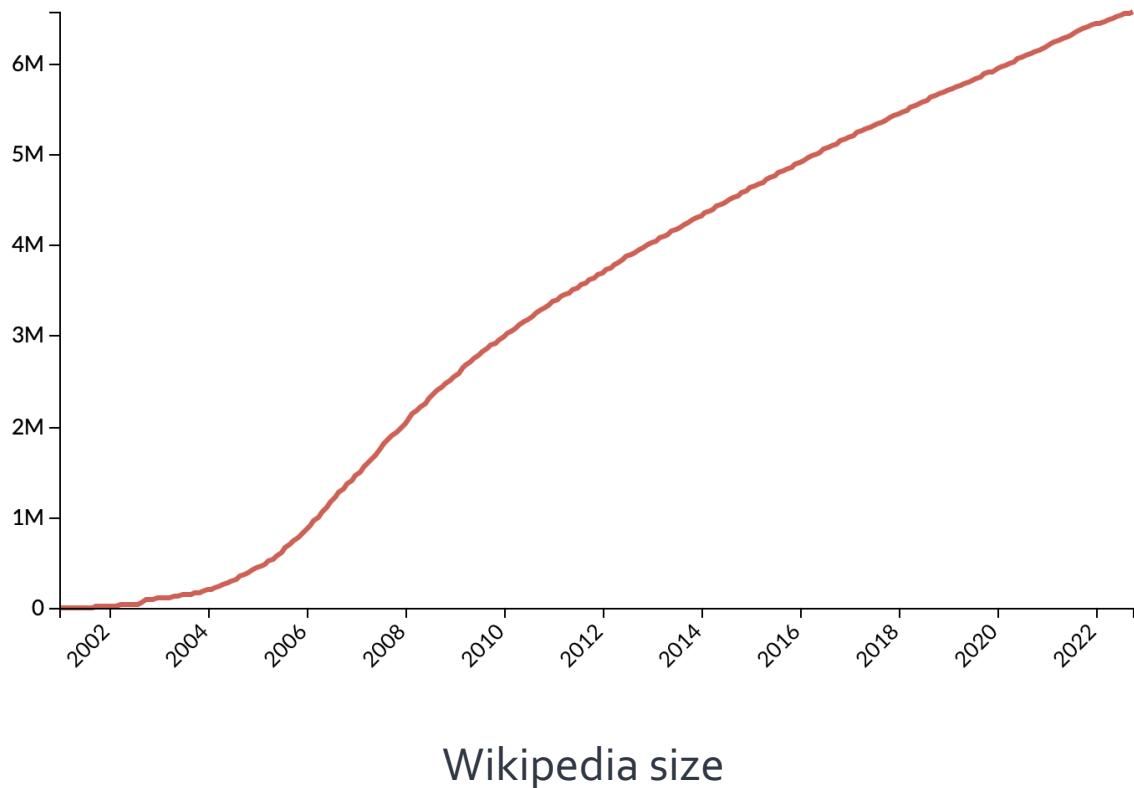
Not Enough Data for Scaling

- 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 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 (?)



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)

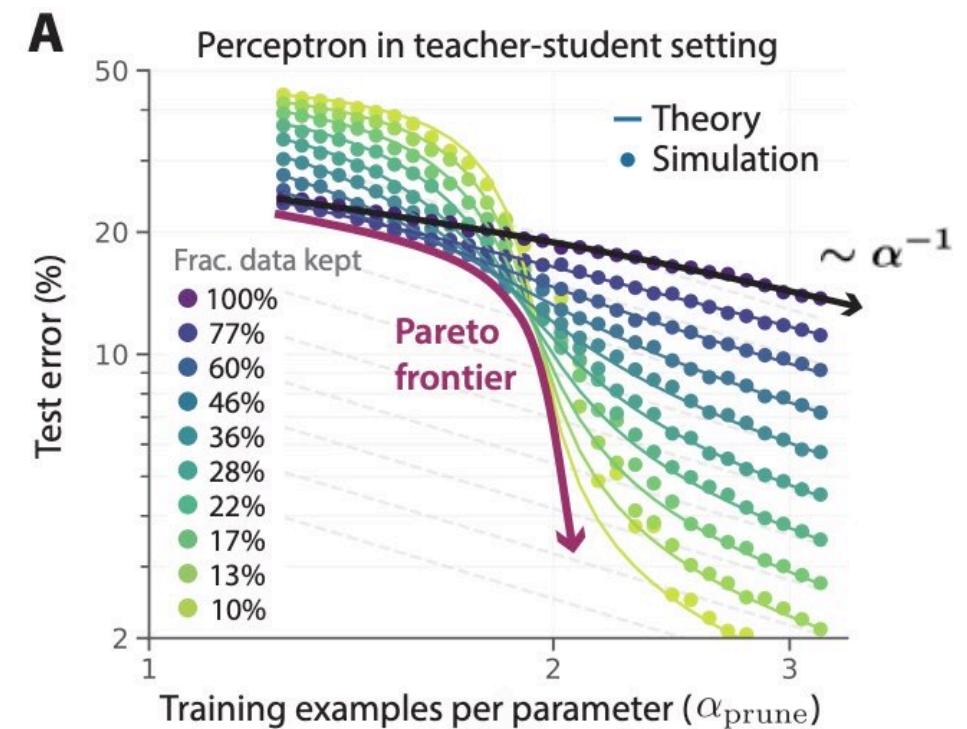


- (aside2: is that why OpenAI built Whisper?!)

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

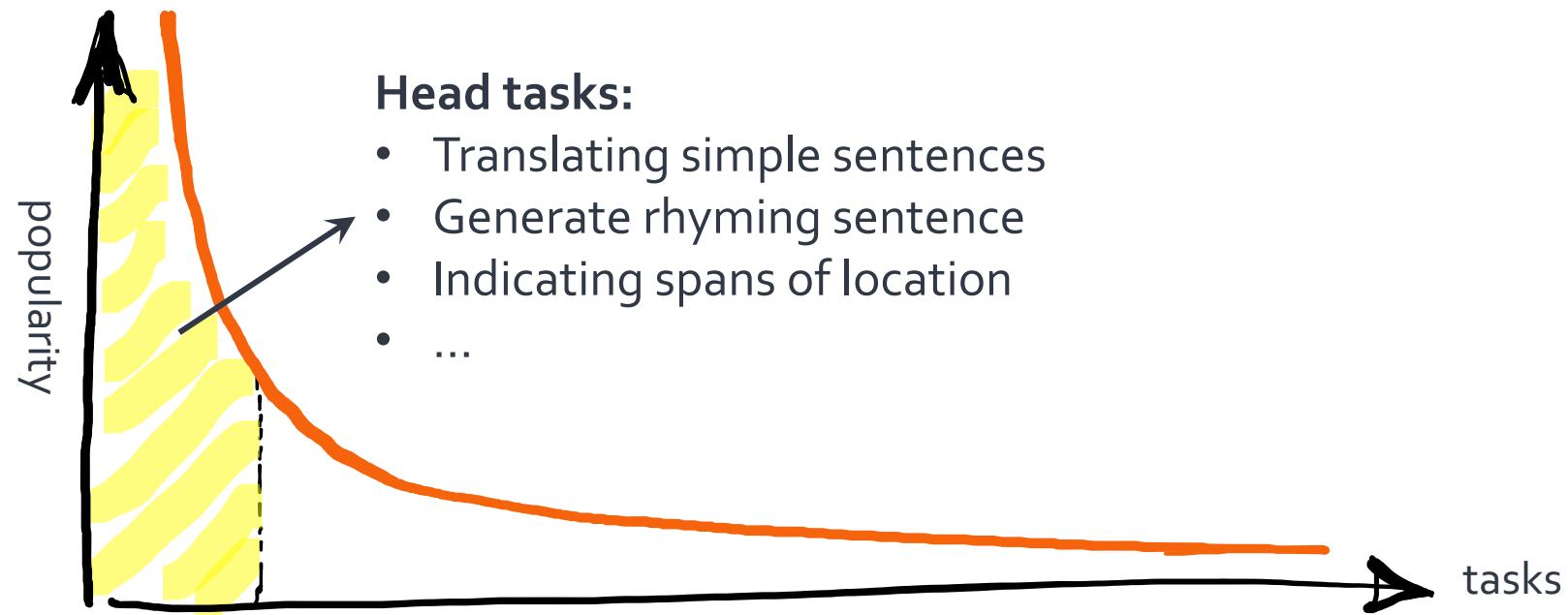
Argument Against “Not Enough Data” (3)

- 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!



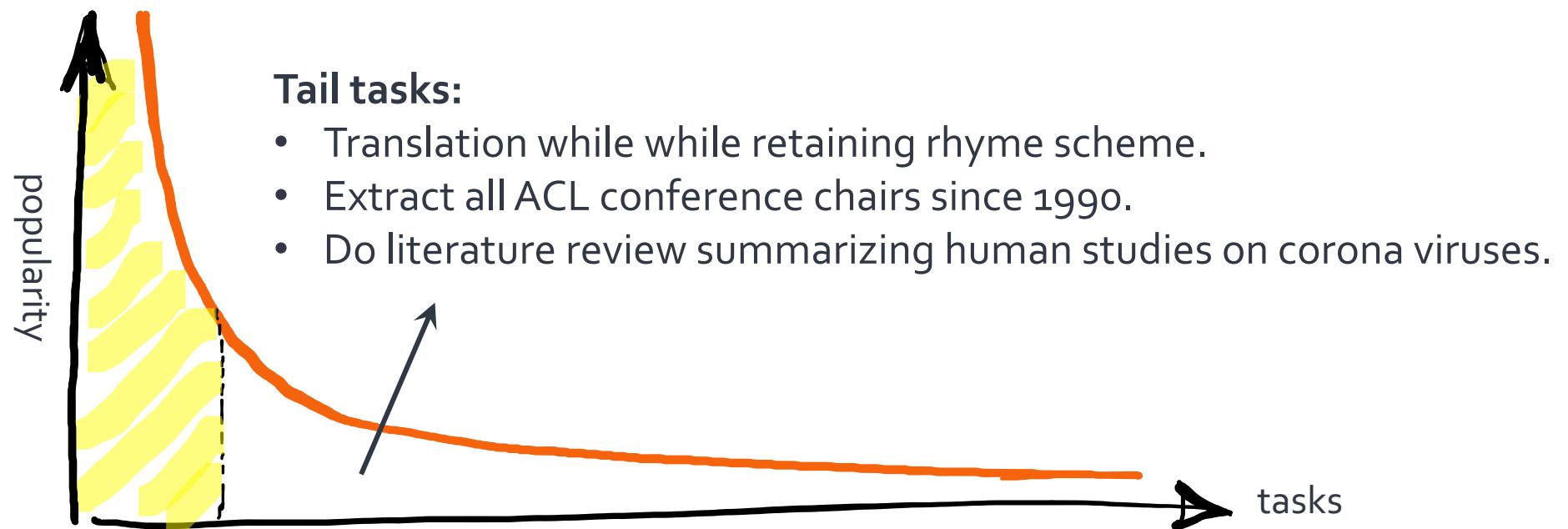
Scale is Not all You Need Because of Tail Phenomena

- 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

- Tail phenomena will never go away!



Example: Google Translate

"The number of parallel sentences [...] ranges from around tens of thousands to almost 2 billion."

Scale is Not all You Need Because of Tail Phenomena

Q: What is 24 times 18? A: ___ Model: 432 ✓
Q: What is 23 times 18? A: ___ Model: 462 ✗

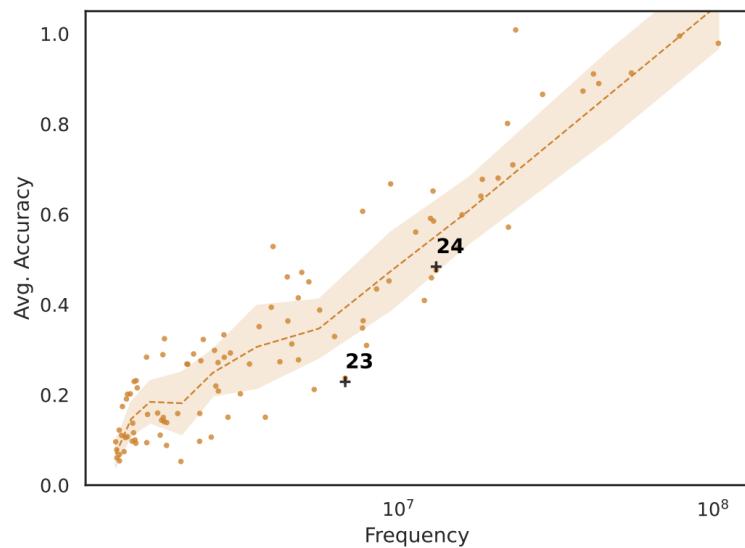


Figure 1. Multiplication Performance: Plot of GPT-J-6B’s 2-shot accuracy on multiplication (averaged over multiple multiplicands and training instances) against the frequency of the equation’s first term in the pretraining corpus. Each point represents the average performance for that term (e.g., 24) 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.

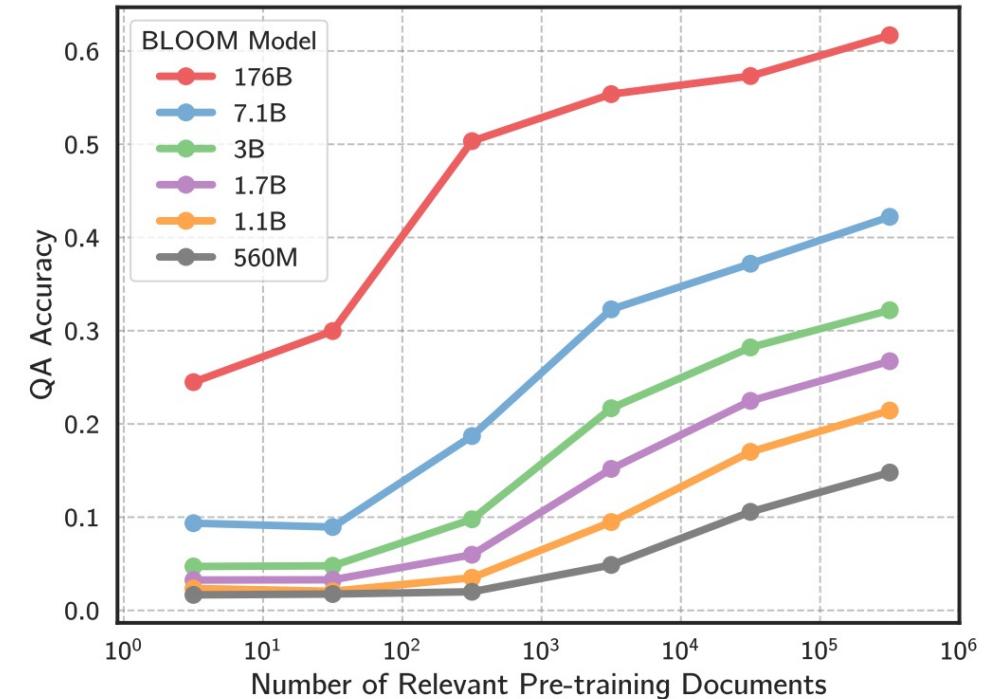


Figure 1: Language models struggle to capture the long-tail of information on the web. Above, we plot accuracy for the BLOOM model family on TriviaQA as a function of how many documents in the model’s pre-training data are relevant to each question.

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”



Y

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.



Y

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.

GR

Factor $x^2 + 35x - 7$, take it step by step



To factor the polynomial $x^2 + 35x - 7$, we need to find two numbers that multiply to -7 and add to 35 . The factors of -7 are -1 and 7 , so we can write the polynomial as:

$$x^2 + 35x - 7 = (x - 1)(x + 7)$$

We can verify that this is the correct solution by multiplying the two factors:

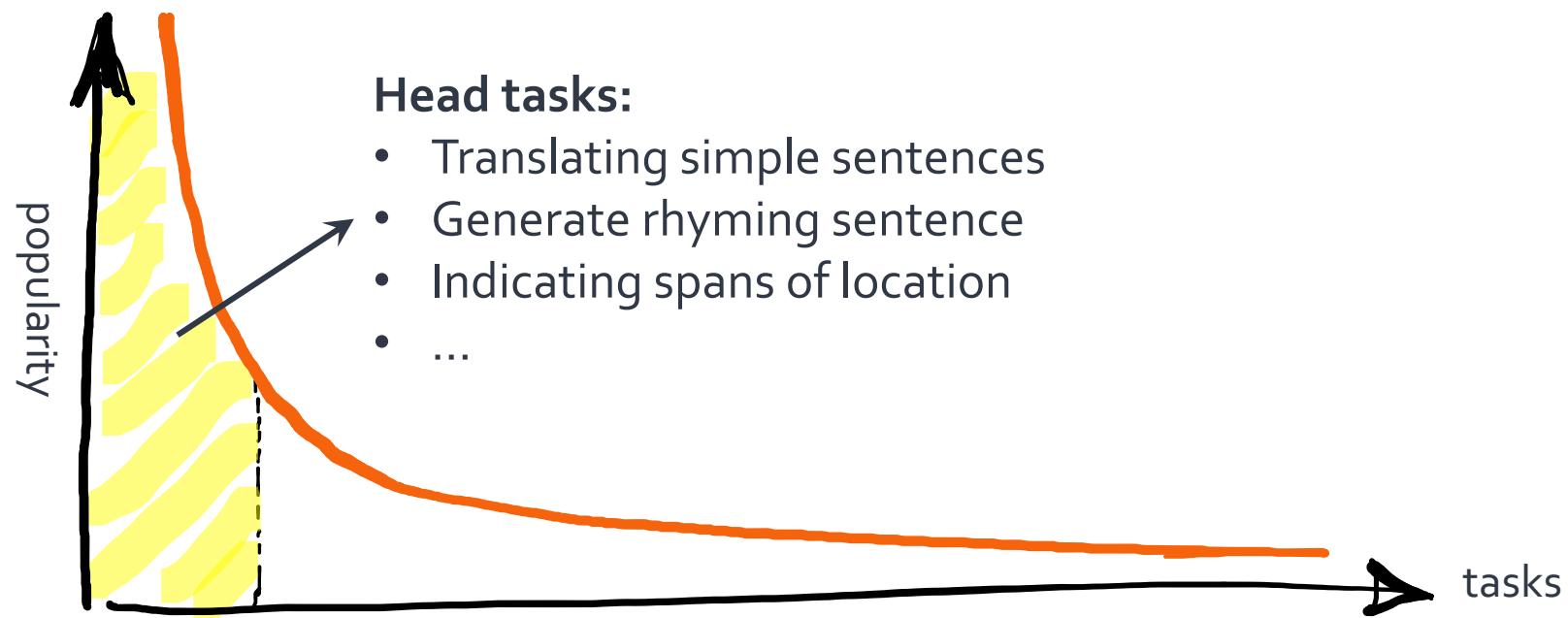
$$(x - 1)(x + 7) = x^2 + 7x - x - 7 = x^2 + 35x - 7$$

Therefore, the factored form of the polynomial is $(x - 1)(x + 7)$.

Scale is Not all You Need Because of Tail Phenomena

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

Let's do a poll!



But do They Really “Understand”?



Erhan Hosca
@ehosca

@GaryMarcus Isn't this basic probability? Enough monkeys banging on typewriters sort of thing...

3:18 PM · Dec 1, 2022



@emilymbender@dair-community.social on Mastodon
@emilymbender

That "Limitations" section has it wrong though. ChatGPT generates strings based on combinations of words from its training data. When it sometimes appears to say things that are correct and sensible when a human makes sense of them, that's only by chance.

But do They Really “Understand”?

- What is “meaning”?
- What is “understanding”?
-

Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data

Emily M. Bender

University of Washington
Department of Linguistics
ebender@uw.edu

Alexander Koller

Saarland University
Dept. of Language Science and Technology
koller@coli.uni-saarland.de

Models are Not Grounded

- Models don't get to experience the world, the way humans do. Hence, they're doomed to fail.

k

which is heavier, 10kg iron or 10kg cotton?



10kg iron is heavier than 10kg cotton. Iron is typically heavier
10kg iron is heavier. Answer: \boxed{iron}.



Mike Ma
@MMikeMMA

@plibin @GaryMarcus So, I'm really struggling here.

These machines don't *understand* anything. As I understand, they are MASSIVE correlation engines.

Yet, they simultaneously:

- hilariously fail
- breathtakingly win

WHY/HOW are both things true? What Qs do the engine win/fail on and why?



3:03 PM · Dec 1, 2022

Argument Against (1)

- You can make models grounded in world by training them on grounded data.

Argument Against (2)

- There are many “intelligent” behaviors in the world that consist of non-intelligent atomic elements.
 - Every neuron in our body
 - Every step of evolution
 -

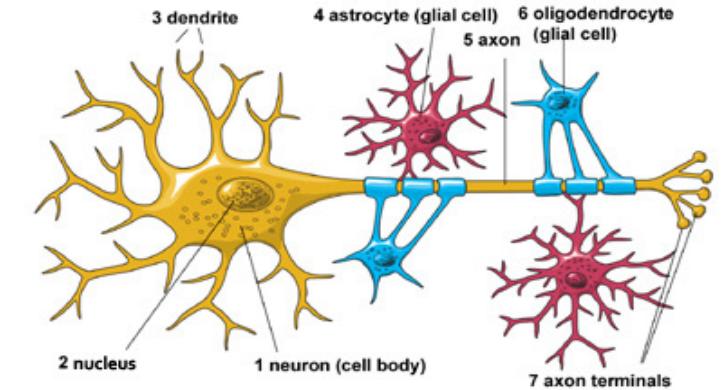
Fundamentally What does “Scale” Solve?

- Maybe data or compute are the bottleneck.
 - If “Scale is All You Need”, is good predictive ability* all needed for “general intelligence”?
 - (is prediction all about “prediction”?)
 - What about causality, planning, coordination, ... ?
- What do you think?

* Self-supervised models == predictive models of the world.

Argument by Comparison to Human Brain

- Do we really need more parameters?
 - The human brain has approximately **100B** neurons.
 - ChatGPT has **170B** parameters
- If the same laws that govern human intelligence also apply to LMs* we don't need more parameters.
- Instead, we may want to focus on better architectures, supervision data and algorithms.



* BIG assumption; might not be true.

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

Interactive Semantics

*Single-shot
evaluation*



*Learning
from
interactions*



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

Instruction-tuning: Limitations (2)

- Dependence on labeled data
- Many tasks are abstract — especially those related to our value system.

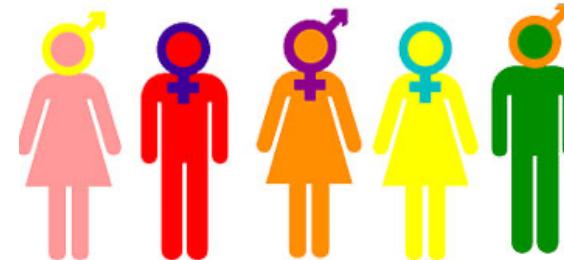
social norms

respecting the elderly



moral norms

avoiding gender or racial bias



human rights

freedom of speech

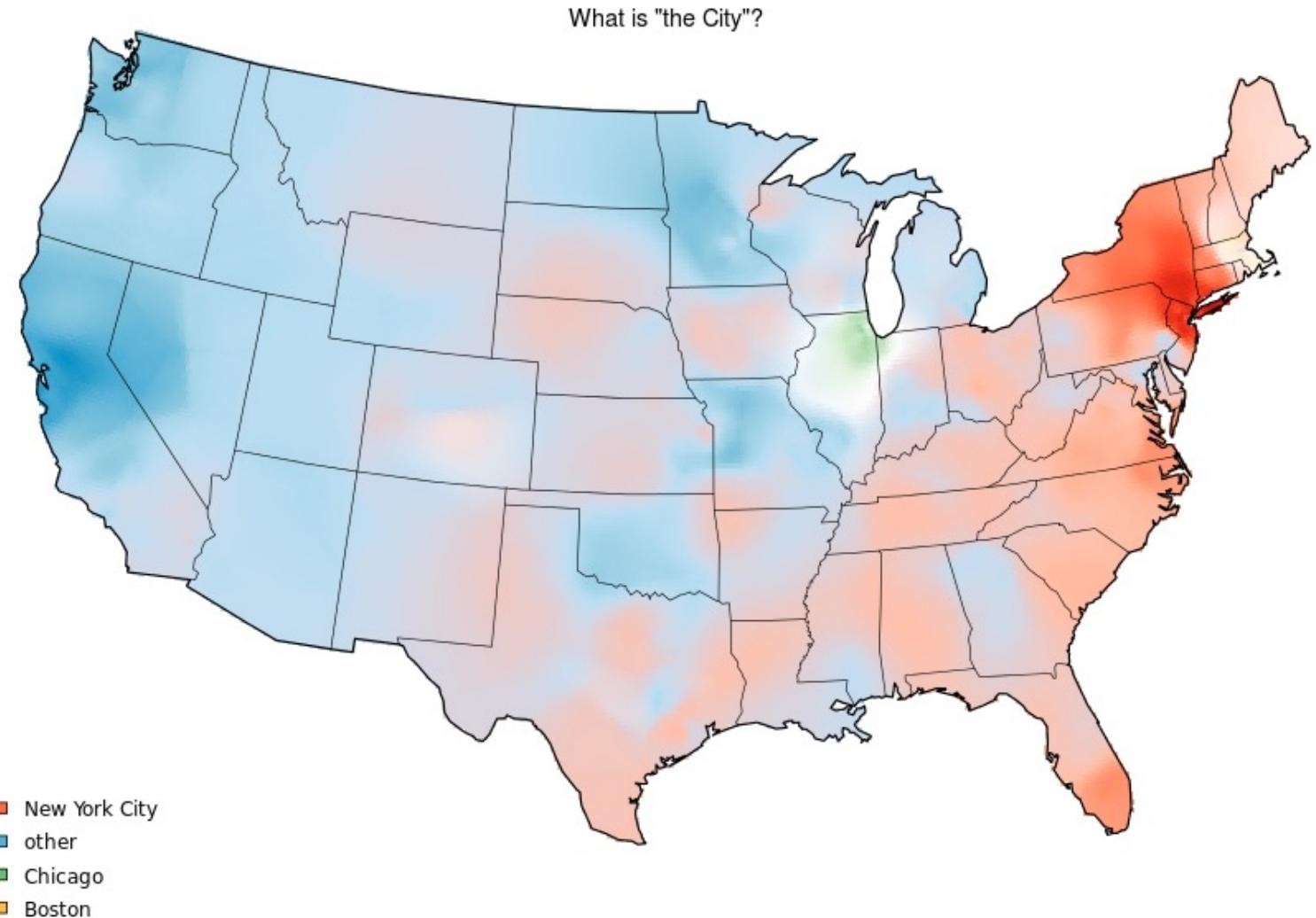


Replacing us or augmenting us?

- Augmenting us: bouldezer
- Robots digging holes

But what is “context”?

- “Context” is everything
 - Seeing
 - Touching
 - Hearing
 - Reading





[Inspiration: Dan Dennett]

Popular Media: AI is Solved!!

FINANCIAL TIMES

Is AI finally closing in on
human intelligence?

INSIDER

Google's DeepMind artificial
intelligence has figured out how to talk

The Atlantic

An Artificial Intelligence Developed
Its Own Non-Human Language

Self-Supervised Models

