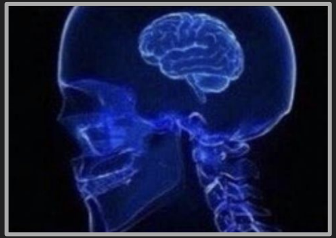
Scaling paradigms for large language models

Jason Wei
Research Scientist
OpenAI

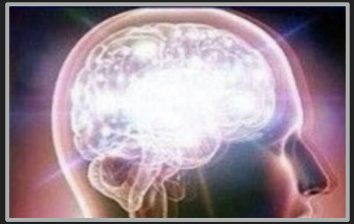
(Opinions are my own and do not reflect my employer.)

2019



- Can barely write a coherent paragraph
- Can't do any reasoning

2024



- Can write an essay about almost anything
- Competition-level programmer and mathematician

Scaling has been the engine of progress in AI and will continue to dictate how the field advances.

Outline

What is scaling and why do it?

Paradigm 1: Scaling next-word prediction

The challenge with next-word prediction

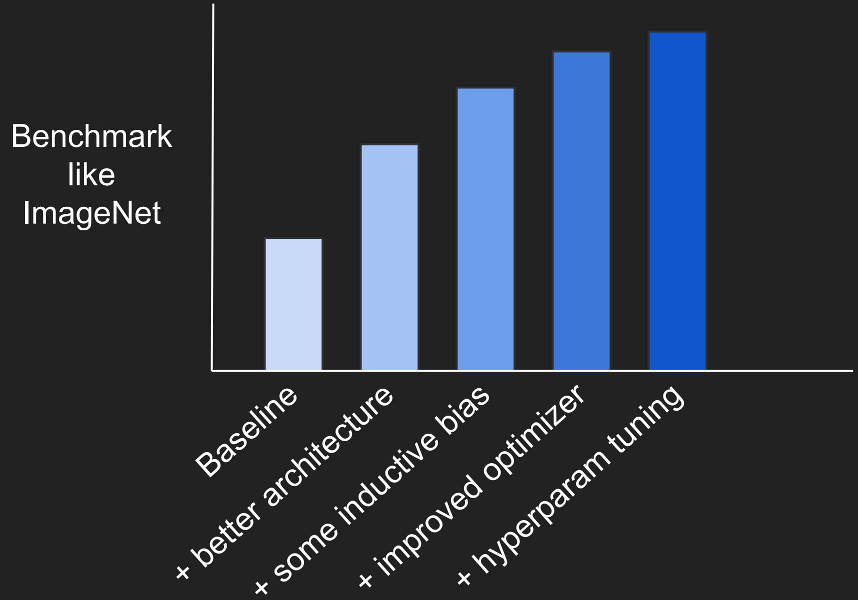
Paradigm 2: Scaling RL on chain-of-thought

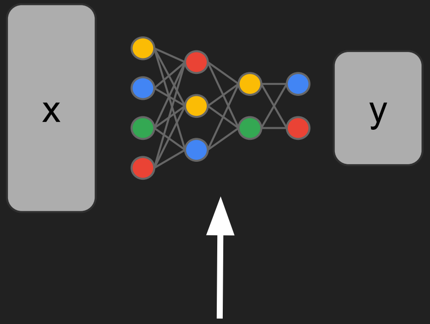
How scaling changed AI culture & what's next?

“Studying the past tells you what’s special about the current moment.”

How we made progress, early 2010s to 2017

(pre-transformer deep learning)



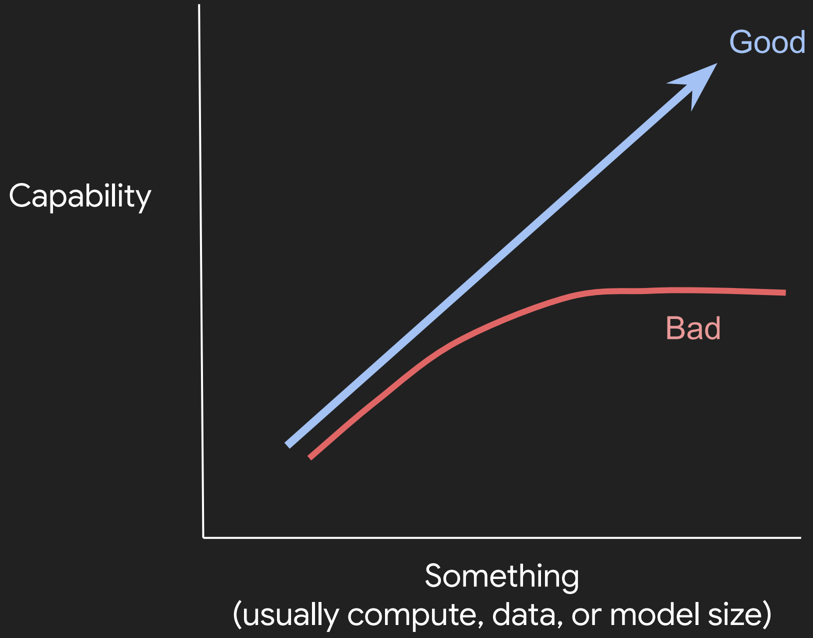


Make this as good as possible.

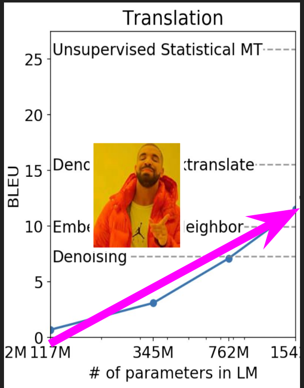
Success looks like “On the Im
dataset, our m
the baseline
the compute.

What is scaling?

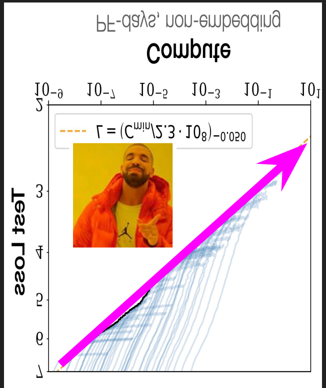
Scaling is when you put yourself in a situation where you move along a continuous axis and expect sustained improvement.



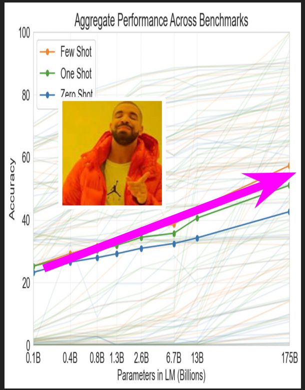
Scaling is everywhere



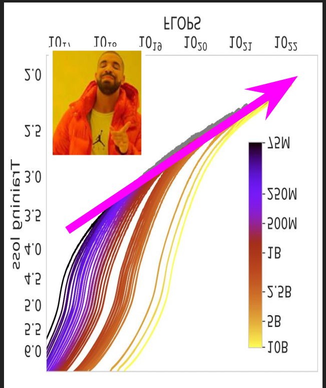
GPT-2 (2019)



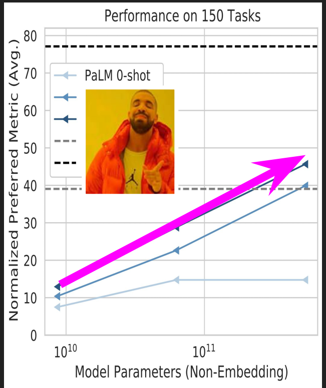
Scaling laws (2020)



GPT-3 (2021)



Chinchilla (2022)



PaLM (2022)

Scaling is hard and was not obvious at the time

Technical & operational challenges

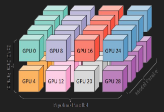
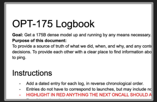


Image source: HF

(1) Distributed training requires a lot of expertise


OPT-175 Logbook

Goal: Get a 175B dense model up and running by any means necessary.

Purpose of this document:

To provide a source of truth of what we did, when, and why, and any cont

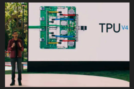
decisions. To provide each other with a clear place to find information abo

to ping.

Instructions

- Add a dated entry for each log, in reverse chronological order.
- Entries do not have to correspond to launches, but may include nc
- HIGHLIGHT IN RED ANYTHING THE NEXT ONCALL SHOULD A

(2) Loss divergences and hardware failures are hurdles



(3) Compute is expensive

Psychological challenges

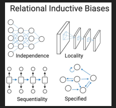
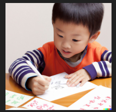


Image source

(1) Researchers like inductive biases



(2) Scaling is different from human learning

(3) Scientific research incentives don't match engineering work ("novelty")

Why scale?

Not scaling

Each improvement in the model requires ingenuity on a new axis

There are a lot of tasks that we want AI to do

Scaling-centric AI

You can reliably improve capability (even if it's expensive)

If your measure of capability is very general, extreme investment is justified

The Bitter Lesson of AI

General methods that leverage compute are the most effective

Things that scale will ultimately win out

The Bitter Lesson

Rich Sutton

March 13, 2019

The biggest lesson that can be read from 70 years of AI research is that gener
computation are ultimately the most effective, and by a large margin. The ul
Moore's law, or rather its generalization of continued exponentially falling c
Most AI research has been conducted as if the computation available to the a
case leveraging human knowledge would be one of the only ways to improv
slightly longer time than a typical research project, massively more computa
available. Seeking an improvement that makes a difference in the shorter ter
leverage their human knowledge of the domain, but the only thing that mat
leveraging of computation. These two need not run counter to each other, bu
Time spent on one is time not spent on the other. There are psychological co
one approach or the other. And the human-knowledge approach tends to co
that make them less suited to taking advantage of general methods leveragi
many examples of AI researchers' belated learning of this bitter lesson, and i
some of the most prominent.

In computer chess, the methods that defeated the world champion, Kasparo
massive, deep search. At the time, this was looked upon with dismay by the
researchers who had pursued methods that leveraged human understanding
chess. When a simpler, search-based approach with special hardware and so
effective, these human-knowledge-based chess researchers were not good lo
force" search may have won this time, but it was not a general strategy, and

Paradigm 1: Scaling next-word prediction

Started in 2018, still ongoing

Get really, really good at predicting the next word.

Why do you get so much from "just" predicting the next word?
Next-word prediction is massively multi-task learning.

Review: next-word prediction

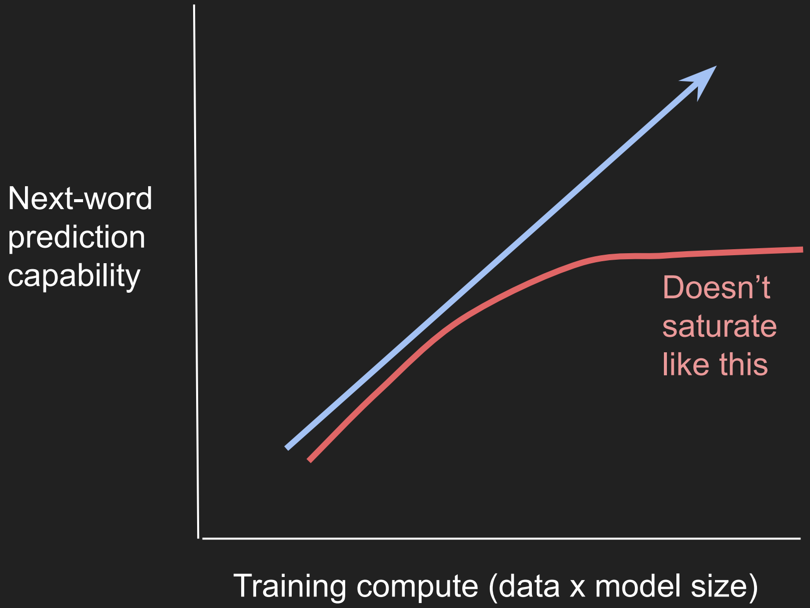
Example “tasks” from next-word prediction

Task	**Example sentence in pre-training that would teach that task**
Grammar	In my free time, I like to {**code**, **banana**}
World knowledge	The capital of Azerbaijan is {**Baku**, **London**}
Sentiment analysis	Movie review: I was engaged and on the edge of my seat the whole time. The movie was {**good**, **bad**}
Translation	The word for “neural network” in Russian is {**нейронная сеть**, **привет**}
Spatial reasoning	Iroh went into the kitchen to make tea. Standing next to Iroh, Zuko pondered his destiny. Zuko left the {**kitchen**, **store**}
Math question	Arithmetic exam answer key: 3 + 8 + 4 = {**15**, **11**}

[millions more]

Extreme multi-task learning!

Scaling predictably improves performance ("scaling laws")



Kaplan et al., 2020:

"Language modeling performance improves smoothly as we increase the model size, dataset size, and amount of compute for training."

Jason's rephrase: You should expect to get a better language model if you scale up compute.

Why does scaling work?

Hard to answer, but here is a hand-wavy explanation

Small language model	**Large language model**
Memorization is costly	More generous with memorizing tail knowledge
First-order correlations	Complex heuristics

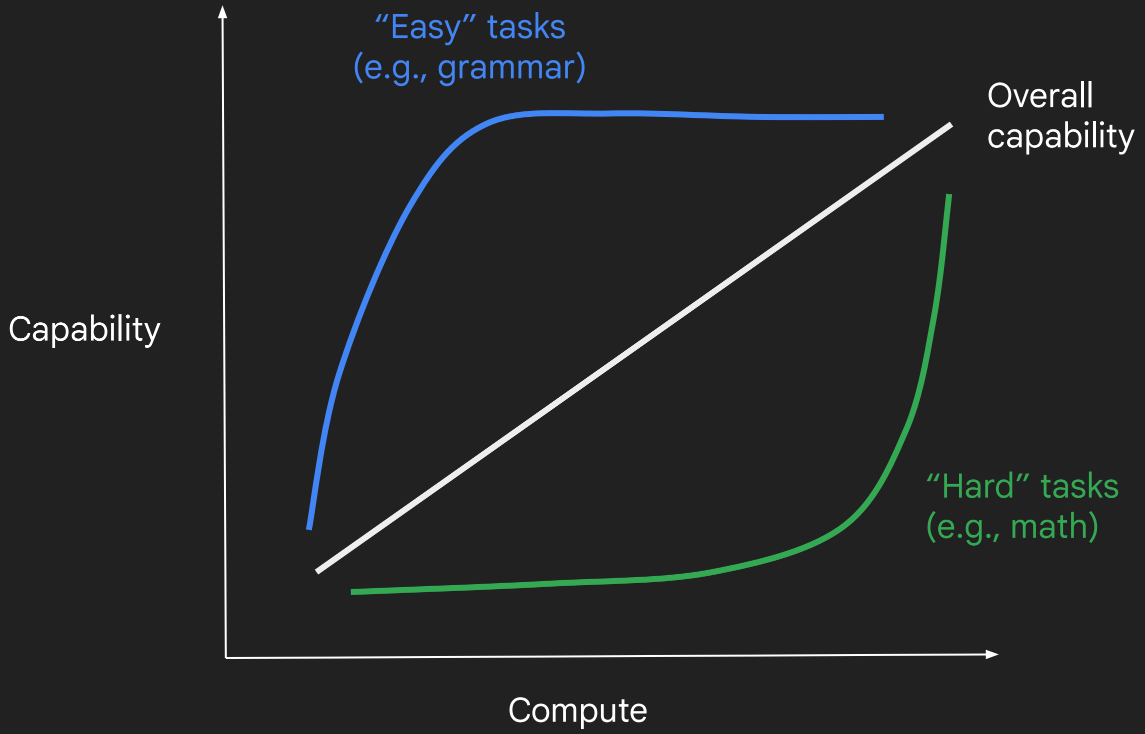
If scaling was so predictable, why was the success of this paradigm so surprising?

Next-word prediction is secretly massively multi-task, and performance on different tasks arise at different rates

Let's take a closer look at next-word prediction accuracy. Consider that

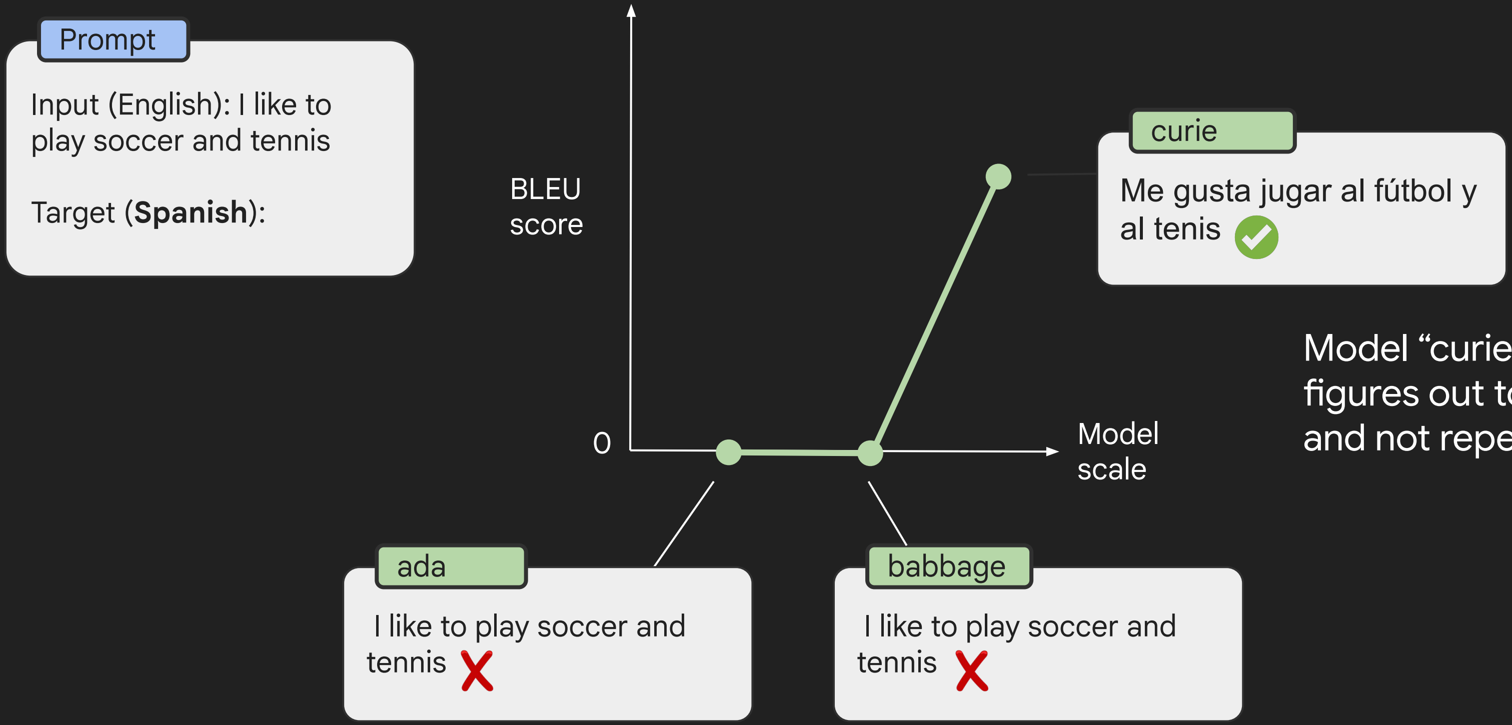
Overall accuracy = 0.002 * accuracy_grammar +
0.005 * accuracy_knowledge +
0.000001 * accuracy_sentiment_analysis +
…
0.0001 * accuracy_math_ability +
0.000001 * accuracy_spatial_reasoning
…

🤔 *If accuracy goes from 70% to 80%, do all tasks get better uniformly?*
…probably not.

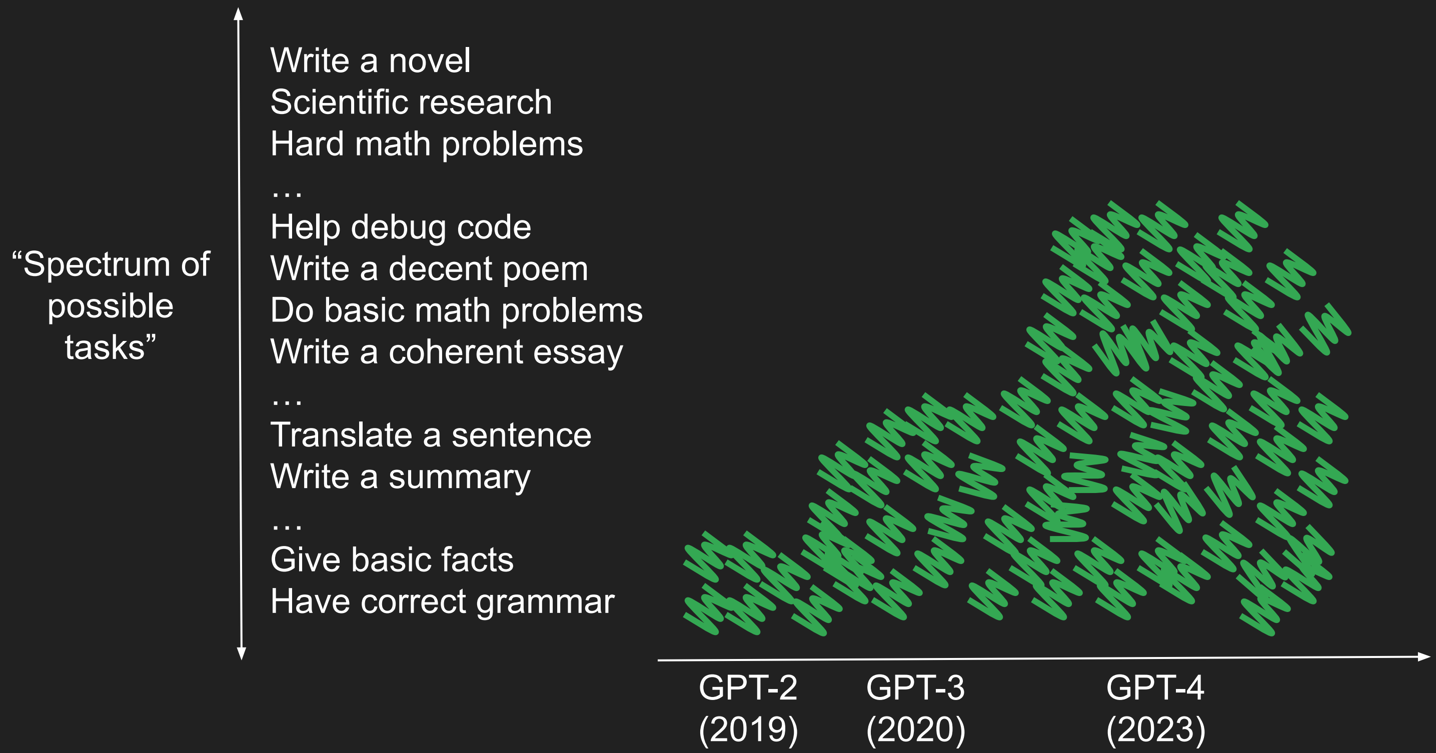


Emergent abilities /
phase transition

Emergence ability example



Model "curie" suddenly figures out to translate and not repeat.

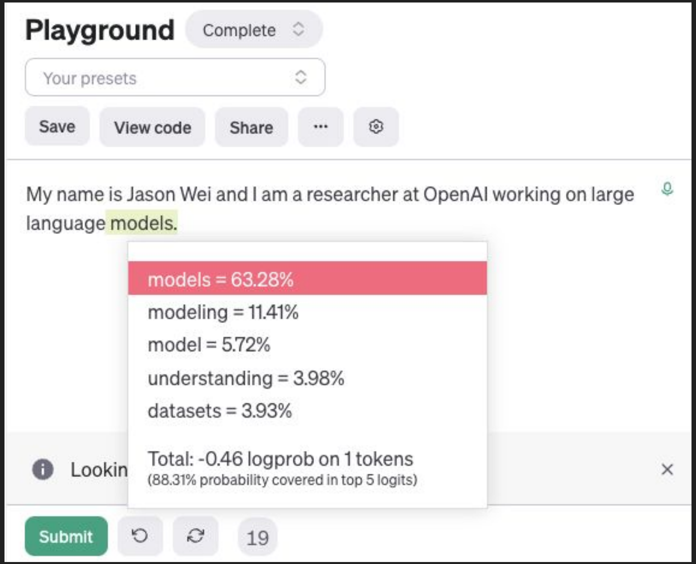

"Spectrum of possible tasks"
Write a novel
Scientific research
Hard math problems
…
Help debug code
Write a decent poem
Do basic math problems
Write a coherent essay
…
Translate a sentence
Write a summary
…
Give basic facts
Have correct grammar
GPT-2 (2019)
GPT-3 (2020)
GPT-4 (2023)

🤔 If next-word prediction works so well, can we scale it to reach AGI?

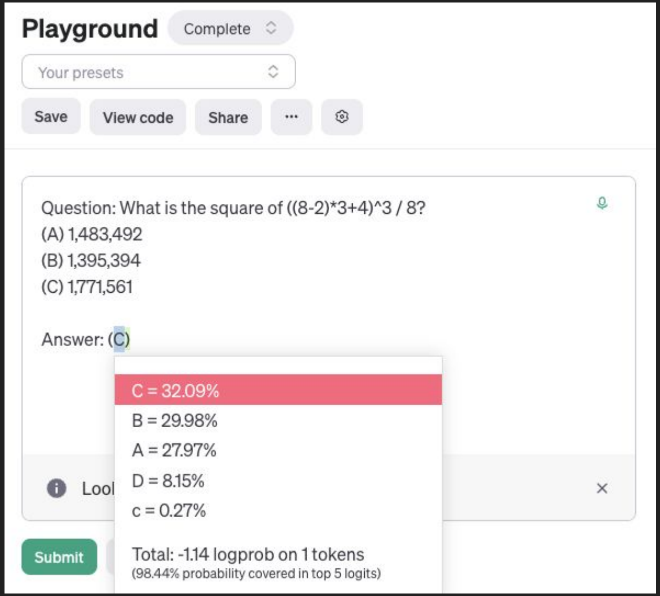
Maybe (it would be hard), but there is a bottleneck:

Some words are super hard to predict and take a lot of work

When next-word prediction works fine



When next-word prediction becomes very hard



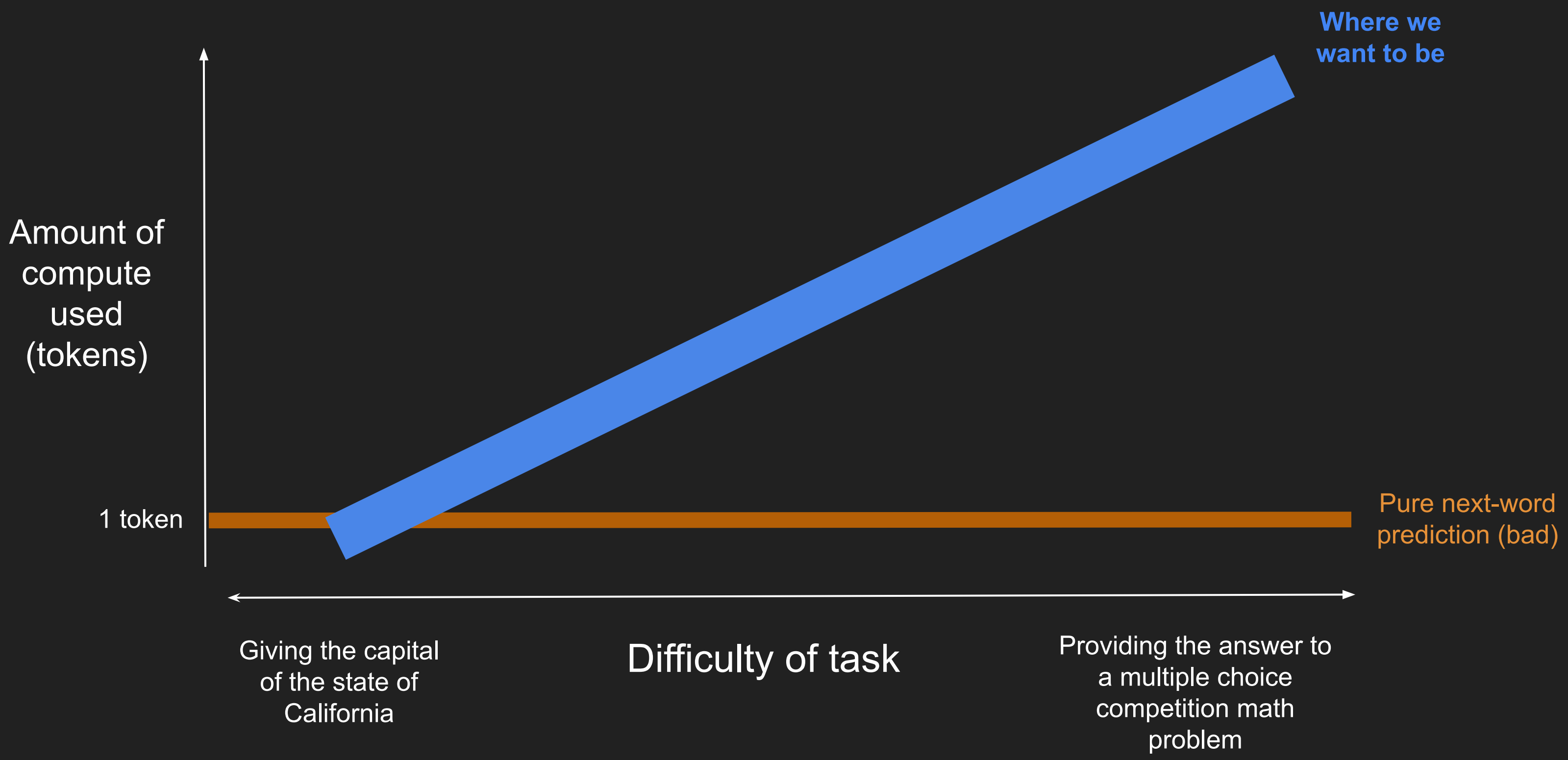
Pretend you're ChatGPT. As soon as you see the prompt you have to immediately start typing... go!

*Question: What is the square of ((8-2)*3+4)^3 / 8?*

(A) 1,483,492
(B) 1,395,394
(C) 1,771,561

...

Tough right?


Where we want to be
Amount of compute used (tokens)
1 token
Pure next-word prediction (bad)
Giving the capital of the state of California
Difficulty of task
Providing the answer to a multiple choice competition math problem

An approach: chain-of-thought prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Question

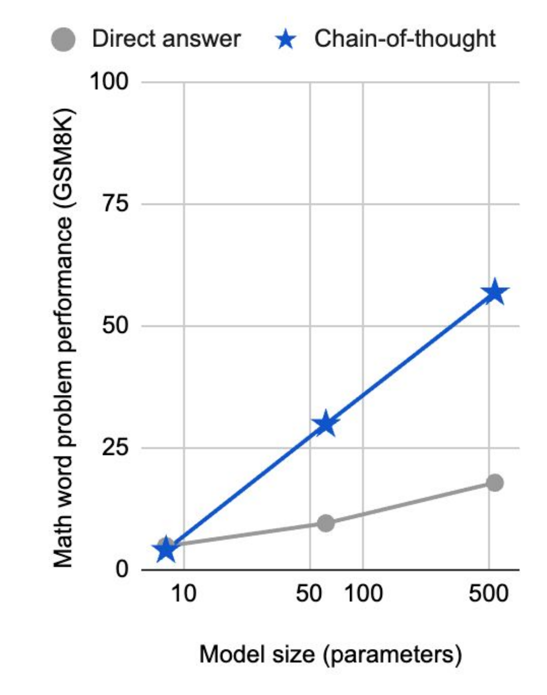
Chain of thought

Answer

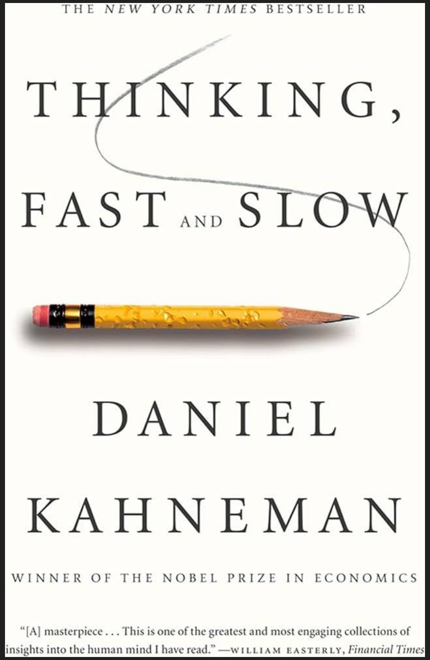
Unseen input

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓



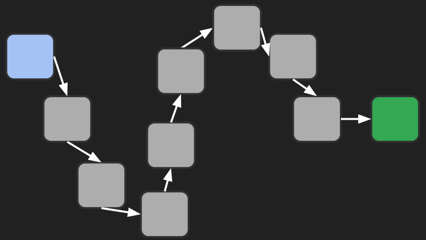
Chain-of-thought prompting elicits reasoning in large language models. Wei et al., 2022.



System 1: Fast, intuitive thinking	**System 2: Slow, deliberate thinking**
Automatic Effortless Intuitive Emotional	Conscious Effortful Controlled Logical
Recognizing faces Repeating basic facts Reacting to something	Solving math problems Planning a detailed agenda Making a thoughtful decision

Next-word prediction

Chain of thought



The limitation with CoT prompting

Most reasoning on the internet looks like this...

17-3: Formally prove Theorem 17.3.2.

Theorem 17.3.2. *A one-pass algorithm for* FREQUENCY-ESTIMATIO *eter* ε *must use* $\Omega(\min\{m,n,\varepsilon^{-1}\})$ *space. In particular, in order to g must use* $\Omega(\min\{m,n\})$ *space.*

Proof. We will prove the stronger result that the simpler FREQUEN which asks whether the input stream contains a token whose frequency space in the deterministic setting. Since the cost of the randomized ministic algorithm, this will prove the theorem as a whole.

Let $\mathscr{A}$ be a one-pass S-space deterministic algorithm for FREQUENC input $(\boldsymbol{x}, y)$ for the IDX_N, Alice creates a stream $\boldsymbol{\sigma}_1 = (a_1, a_2, \cdots, a_N)$ a stream $\boldsymbol{\sigma}_2 = (b, b, \cdots, b)$ of length $k-1$ for $k \geq 2$ where $b = 2y - 1$ on the combined stream $\boldsymbol{\sigma}_1 \circ \boldsymbol{\sigma}_2$ with parameter k.

The output of $\text{IDX}_N(\boldsymbol{x}, y)$ is 1 iff $\mathscr{A}$ produces b as output. This is so b will be the unique entry with $f_b = k \geq k$. Thus Alice and Bob can solv Alice to Bob using $\mathscr{A}$.

By the lower bound result of $\Omega(N)$ for IDX_N, $S = \Omega(N)$. By const $N + k - 1 \geq N + 1$. Therefore, we have proven a lower bound of $\Omega($ and $n \geq N + 1$. We have thus proven that $S = \Omega(\min\{m, n, \varepsilon^{-1}\})$, sinc

What we actually want is the inner "stream of thought"

Hm let me first see what approach we should take...

Actually this seems wrong

No that approach won't work, let me try something else

Let me try computing this way now

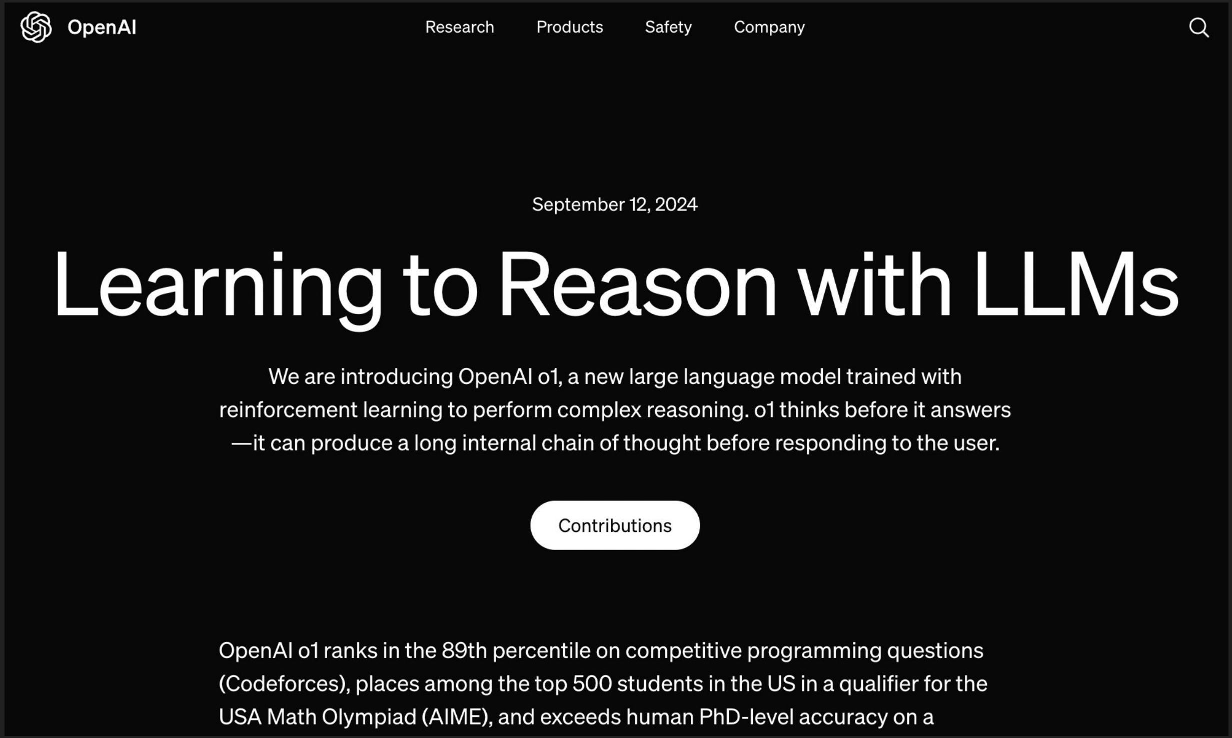
OK I think this is the right answer!

Paradigm 2: Scaling RL on chain-of-thought

Train language models to “think” before giving an answer

In addition to scaling compute for training, there is a second axis here: scaling how long the language model can think at inference time.

OpenAI o1 (work of most of the company)


OpenAI

Research Products Safety Company

September 12, 2024

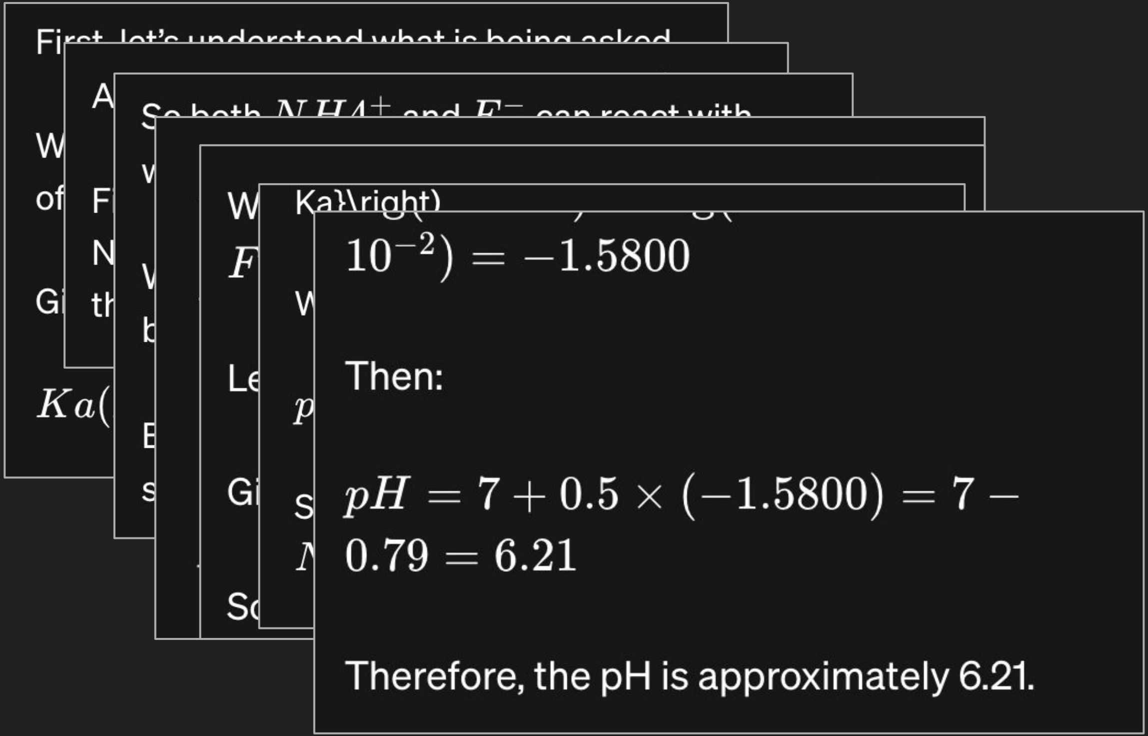
Learning to Reason with LLMs

We are introducing OpenAI o1, a new large language model trained with reinforcement learning to perform complex reasoning. o1 thinks before it answers—it can produce a long internal chain of thought before responding to the user.

Contributions

OpenAI o1 ranks in the 89th percentile on competitive programming questions (Codeforces), places among the top 500 students in the US in a qualifier for the USA Math Olympiad (AIME), and exceeds human PhD-level accuracy on a

A chain of thought from OpenAI o1

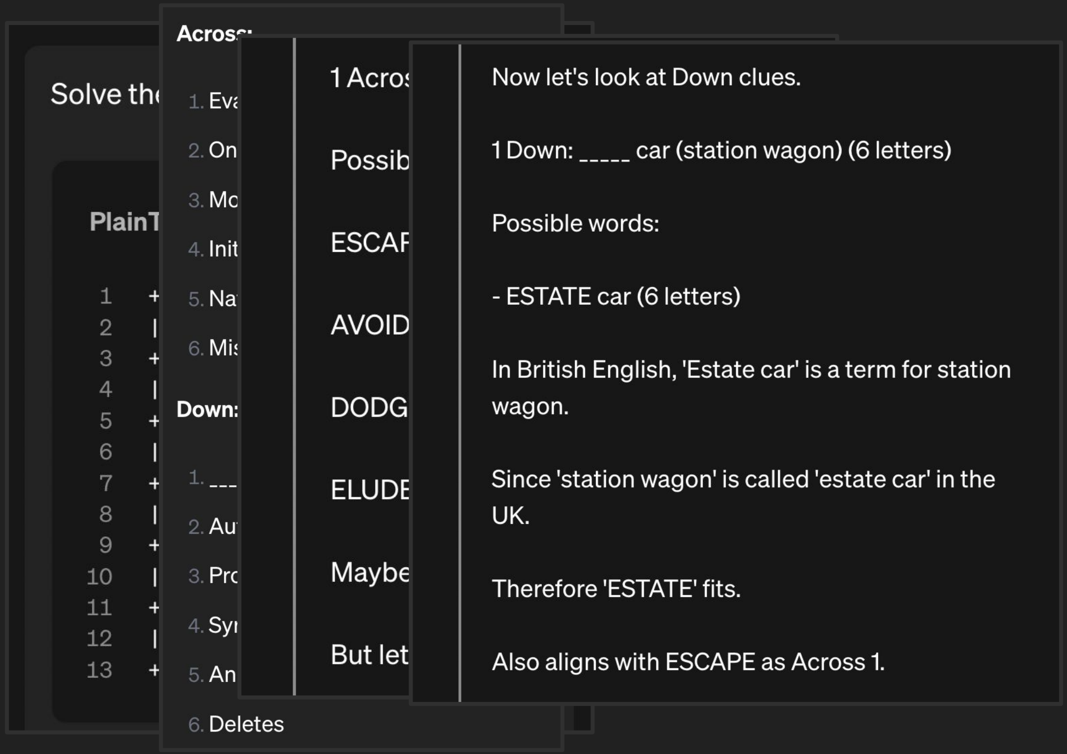


Learning to reason with LLMs. OpenAI, September 2024.

CoT allows models to leverage asymmetry of verification

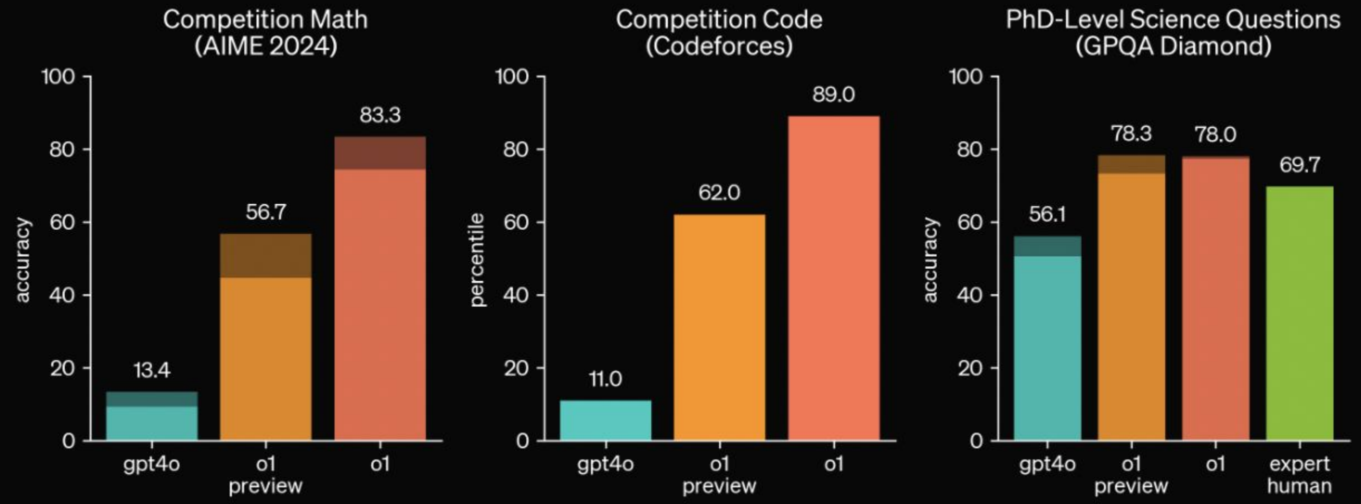
A class of problems has "asymmetry of verification", which means it's easier to verify a solution than to generate one

For example, a crossword puzzle, sudoku, or writing a poem that fits constraints



Learning to reason with LLMs. OpenAI, September 2024.

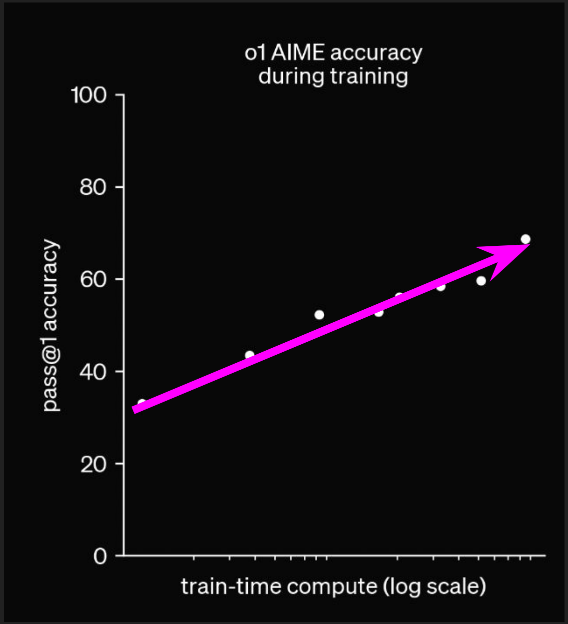
Scale RL on chain-of-thought

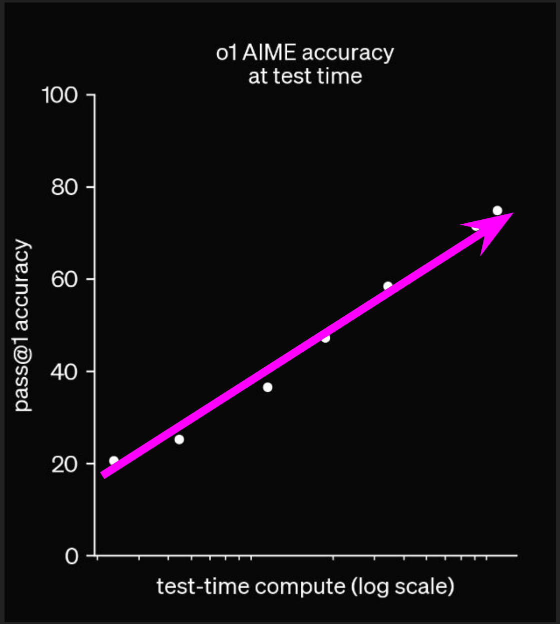


o1 greatly improves over GPT-4o on challenging reasoning benchmarks. Solid bars show pass@1 accuracy and the shaded region shows the performance of majority vote (consensus) with 64 samples.

Learning to reason with LLMs. OpenAI, September 2024.

Scale inference-time compute





Learning to reason with LLMs. OpenAI, September 2024.

Why is this special: one day we may want AI to solve very challenging problems

Prompt

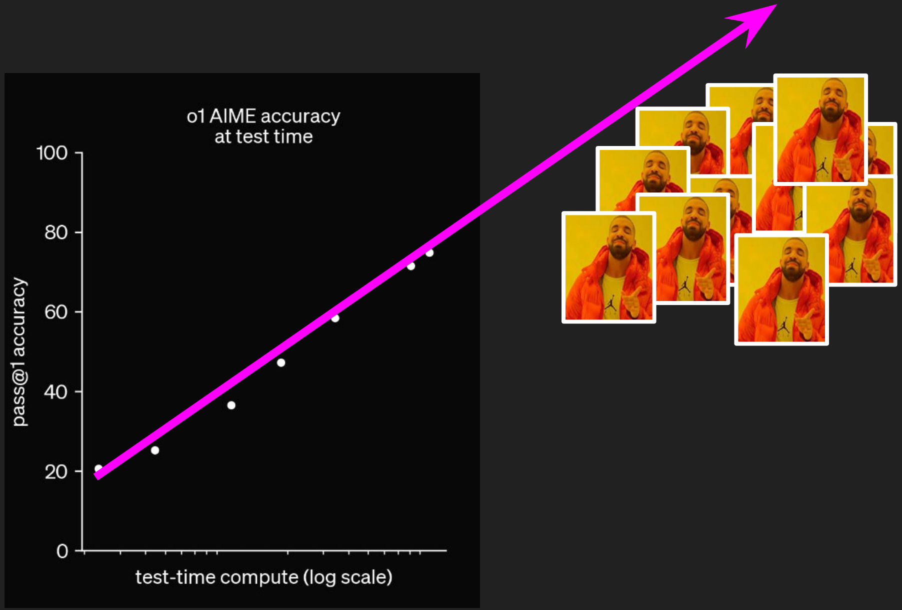
Write the code, documentation, and research paper for the best way to make AI safe

Hypothetical response

Let me think very hard about this...

[Researches all the existing literature] [Data analysis] [Conducts new experiments]

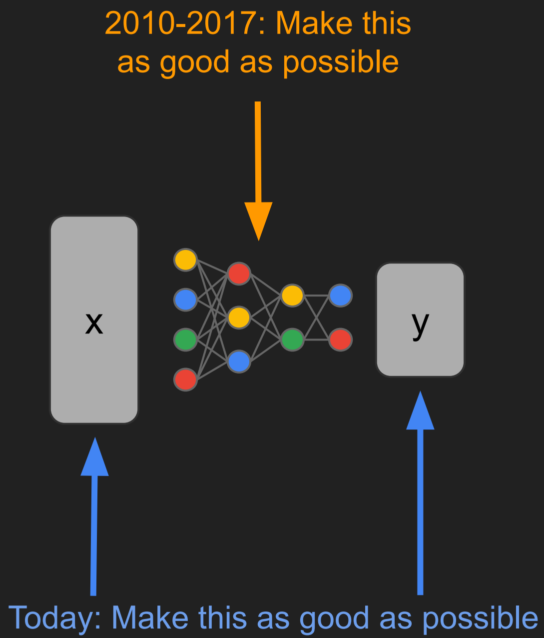
OK, here is a body of work on how to make AI safe



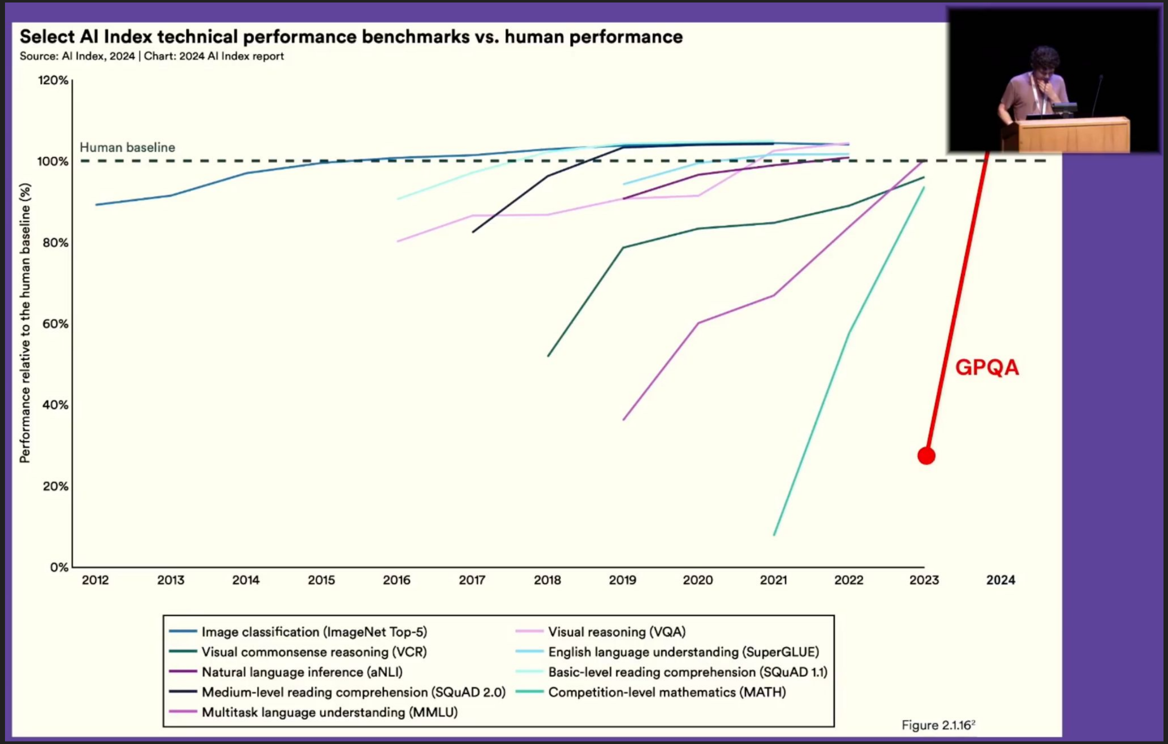
seconds minutes hours days weeks ***months***

How has scaling changed the culture around doing AI research?

Changes in AI research culture: shift to data

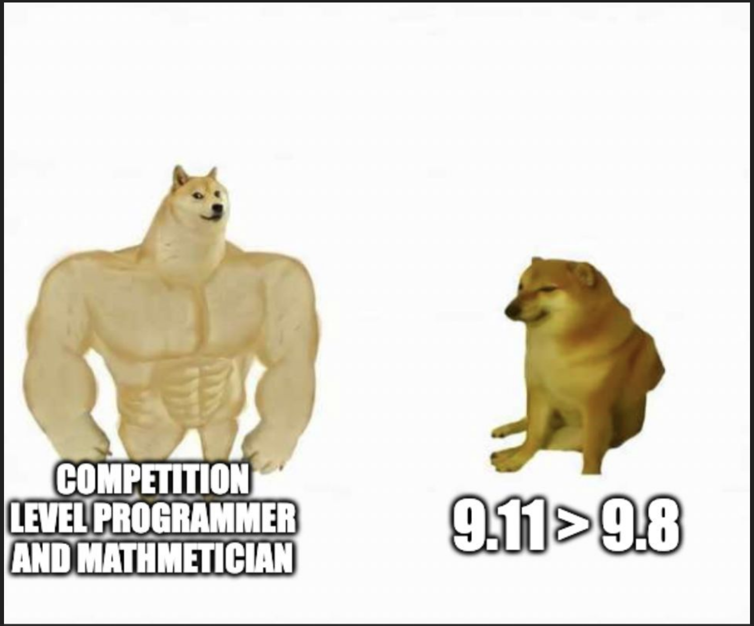


Changes in AI culture: we desperately need evals



"People ask me if I'm making an even harder version of GPQA... [well] we set out to make the hardest science benchmark that we could"
- David Rein

Changes in AI culture: highly multi-task models



Language models must be measured on many dimensions

Hard to say that one model is strictly better than another

AI doesn’t need to human-level on everything

Intelligence != user experience

Changes in AI culture: bigger working teams


[Submitted on 19 Dec 2023 (v1), last revised 17 Jun 2024 (this version, v4)]

Gemini: A Family of Highly Capable Multimodal Models

Gemini Team Google: Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, …

… Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, Oriol Vinyals (collapse list)

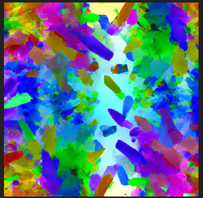
This report introduces a new family of multimodal models, Gemini, that exhibit remarkable capabilities across image, audio, video, and text understanding. The Gemini family consists of Ultra, Pro, and Nano sizes, suitable for applications ranging from complex reasoning tasks to on-device memory-constrained use-cases. Evaluation on a broad range of benchmarks shows that our most-capable Gemini Ultra model advances the state of the art in 30 of 32 of these benchmarks - notably being the first model to achieve human-expert performance on the well-studied exam benchmark MMLU, and improving the state of the art in every one of the 20 multimodal benchmarks we examined. We believe that the new capabilities of the Gemini family in cross-modal reasoning and language understanding will enable a wide variety of use cases. We discuss our approach toward post-training and deploying Gemini models responsibly to users through services including Gemini, Gemini Advanced, Google AI Studio, and Cloud Vertex AI.

Subjects: **Computation and Language (cs.CL)**; Artificial Intelligence (cs.AI); Computer Vision and Pattern Recognition (cs.CV)
Cite as: arXiv:2312.11805 **[cs.CL]**
(or arXiv:2312.11805v4 **[cs.CL]** for this version)
https://doi.org/10.48550/arXiv.2312.11805

Where will AI continue to progress?

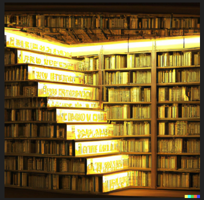
AI for science and healthcare

As an assistant in scientific and medical innovation



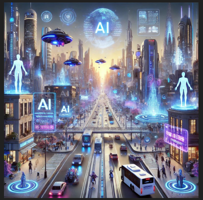
Tool use

Goal: enable AI to interact with the world



More factual AI

Reduced hallucinations, cite sources, calibration



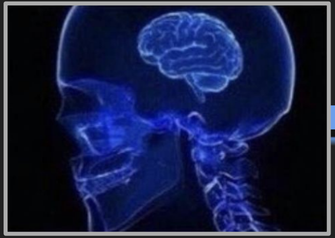
AI applications

More ubiquitous use of AI

Multimodality

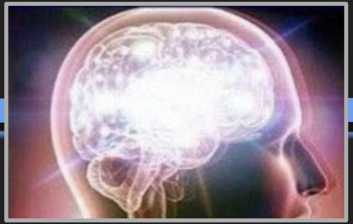
AI to see, hear, and speak

2019



- Can barely write a coherent paragraph
- Can’t do any reasoning

2024



- Can write an essay about almost anything
- Competition-level programmer and mathematician

2029

?

Scaling has been the engine of progress in AI and will continue to dictate how the field advances.

X / Twitter: @_jasonwei
OpenAI roles: jasonwei@openai.com

Feedback? https://tinyurl.com/jasonwei