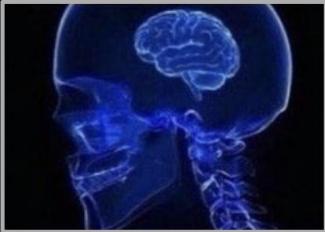


Driving forces in AI in 2025

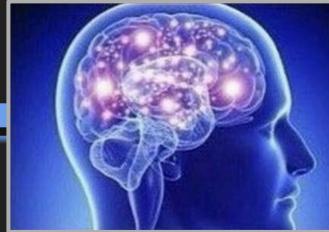
Jason Wei
Research Scientist
OpenAI

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

2020



2025



2030



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

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

?

How did our field progress so quickly, and what will happen in the future?

Talk outline

Motivate the driving forces in AI in 2025:

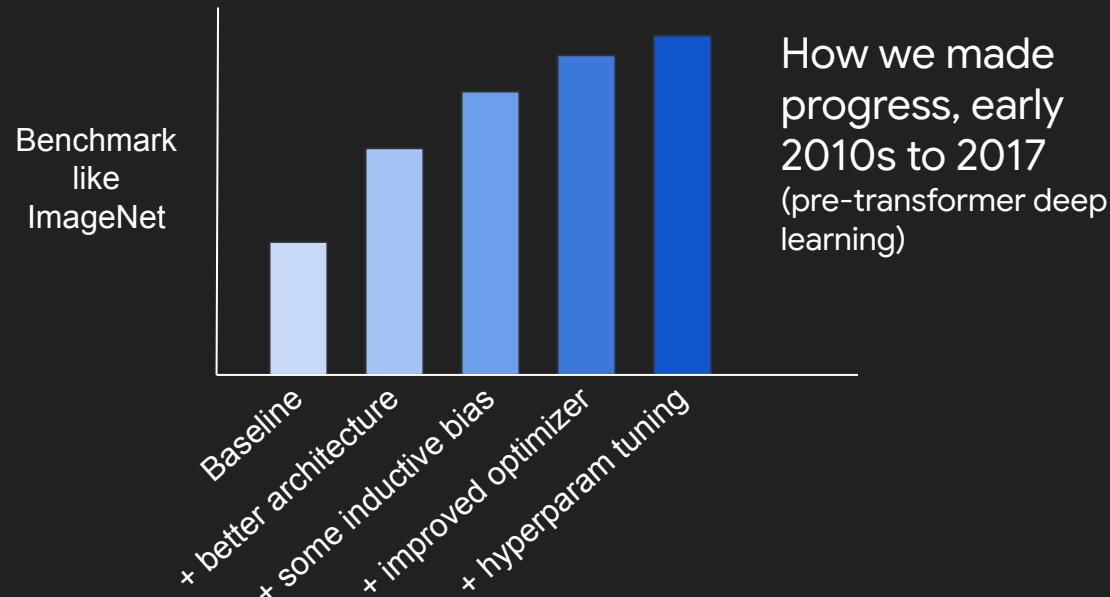
0. Scaling
1. Pre-training
2. Test-time compute via RL

Try to extrapolate into the next few years:

- How will AI research evolve?
- How will AI change our world?

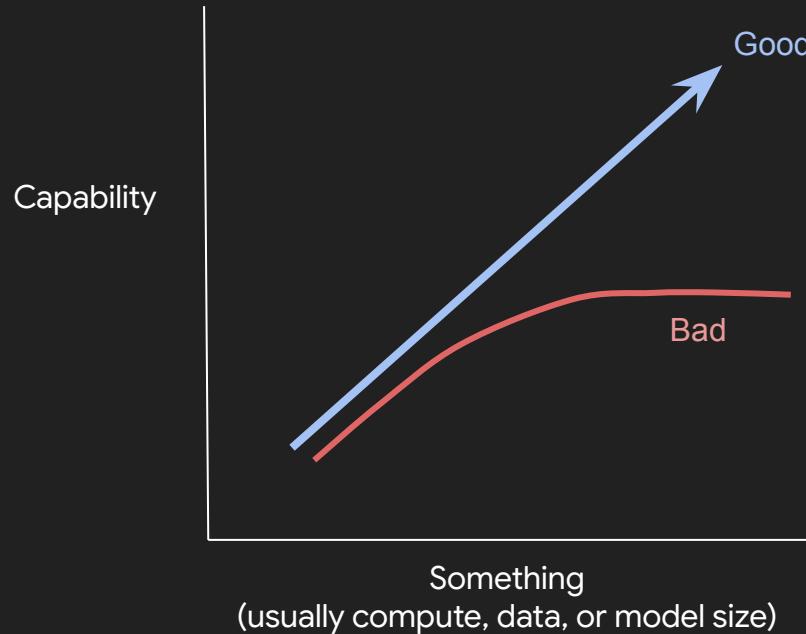
Scaling

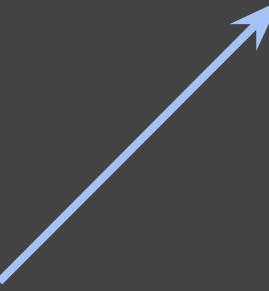
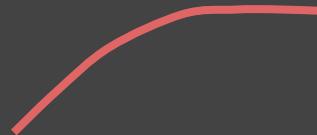
“Studying the past tells you what’s special about the current moment.”



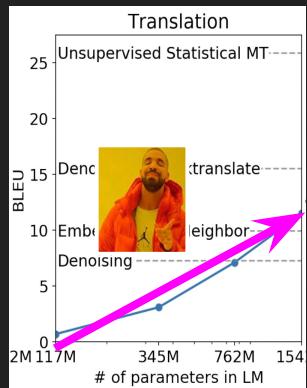
What is scaling?

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

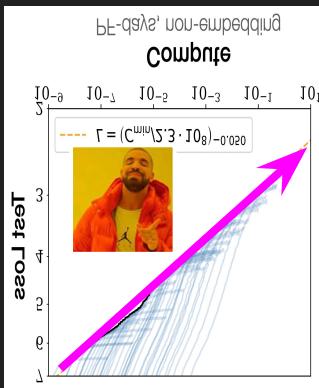




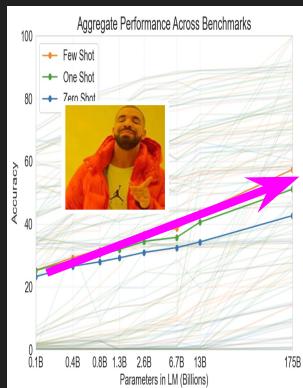
Scaling is everywhere



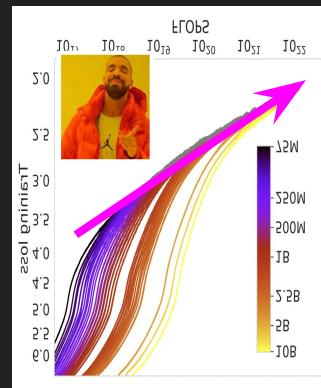
GPT-2 (2019)



Scaling laws (2020)



GPT-3 (2021)



Chinchilla (2022)



PaLM (2022)

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 general computation are ultimately the most effective, and by a large margin. The ultimate generalization of Moore's law, or rather its generalization of continued exponentially falling costs, is that computation is the most effective way to solve problems. Most AI research has been conducted as if the computation available to the algorithm were the limiting factor. In many cases, leveraging human knowledge would be one of the only ways to improve performance significantly in a slightly longer time than a typical research project, massively more computation available. Seeking an improvement that makes a difference in the shorter term requires leveraging their human knowledge of the domain, but the only thing that matters is the cost of computation. These two need not run counter to each other, but they do. Time spent on one is time not spent on the other. There are psychological costs to one approach or the other. And the human-knowledge approach tends to come at a cost that make them less suited to taking advantage of general methods leveraging computation. Many examples of AI researchers' belated learning of this bitter lesson, and it is one of the most prominent.

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

Pre-training

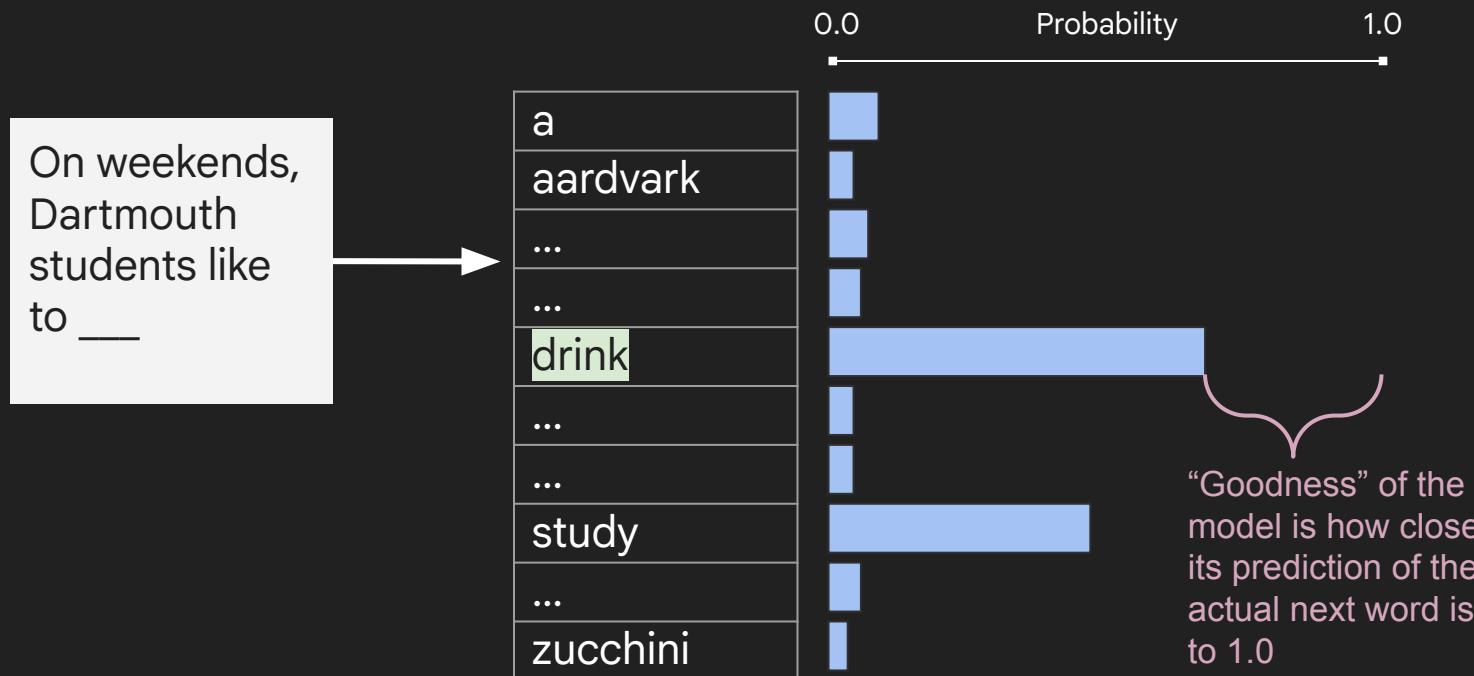
What training task do we scale?

Scaling is great when there is a task that:

1. Has a lot of data
2. Does not get saturated, and there is a gradient of difficulty

Golden task: *predict the next word* in text from the internet

Next-word prediction



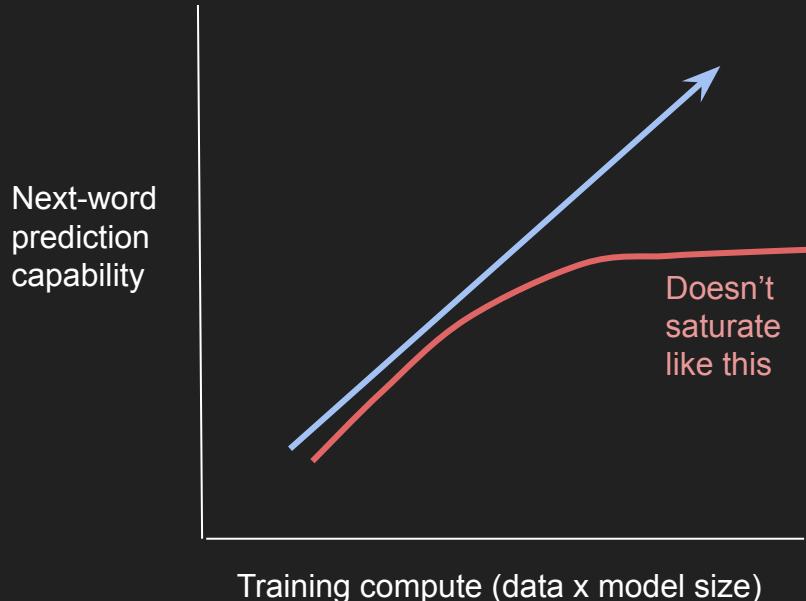
You learn a lot about the world by predicting the next word

<u>Task</u>	<u>Example sentence in pre-training that would teach that task</u>
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 “pretty” in Spanish is { bonita , hola }
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.”

Rephrased: You should expect to get a better language model if you scale up compute.

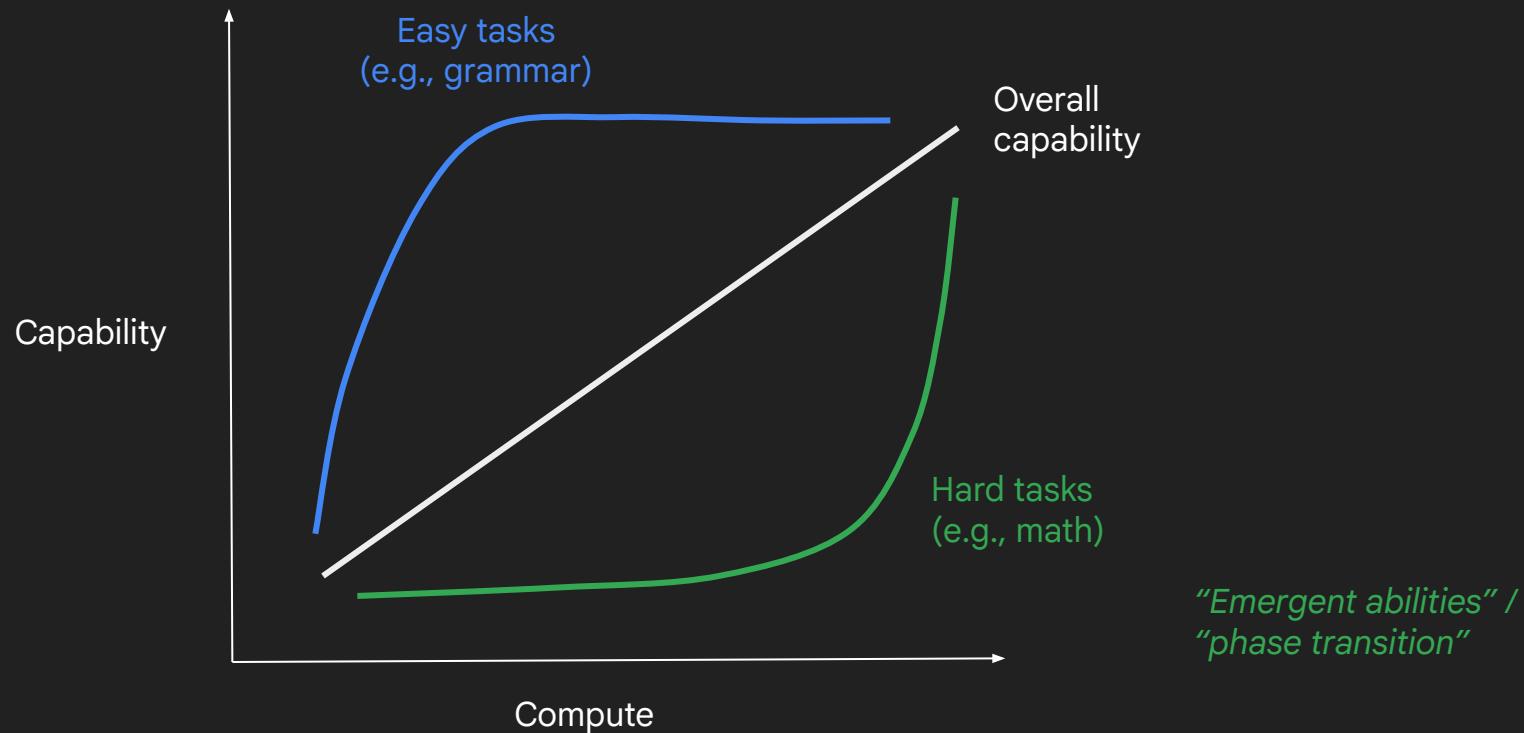
If scaling was so predictable, why was the success of this paradigm so surprising?

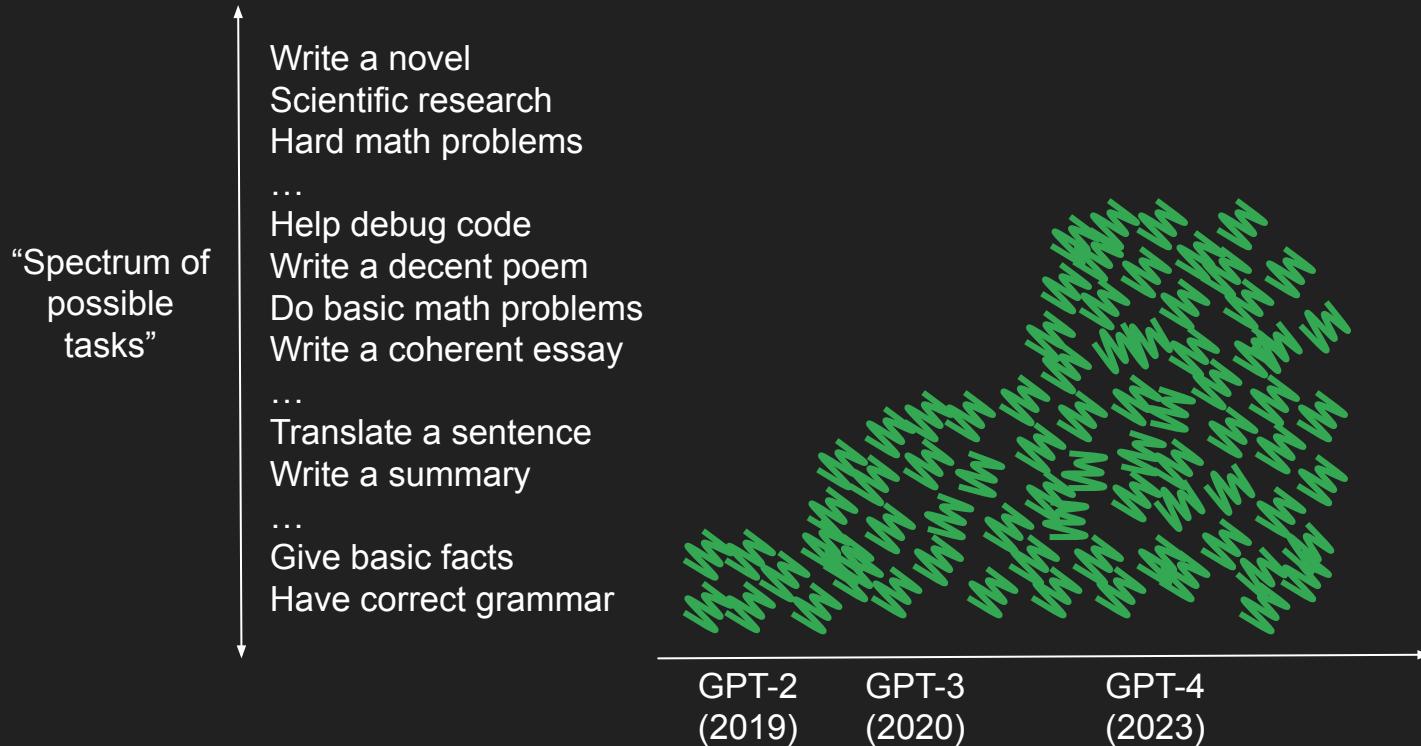
Next-word prediction is secretly massively multi-task, and performance on different tasks arise at different rates

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.





The future of pre-training?



Pre-training as we know it will end

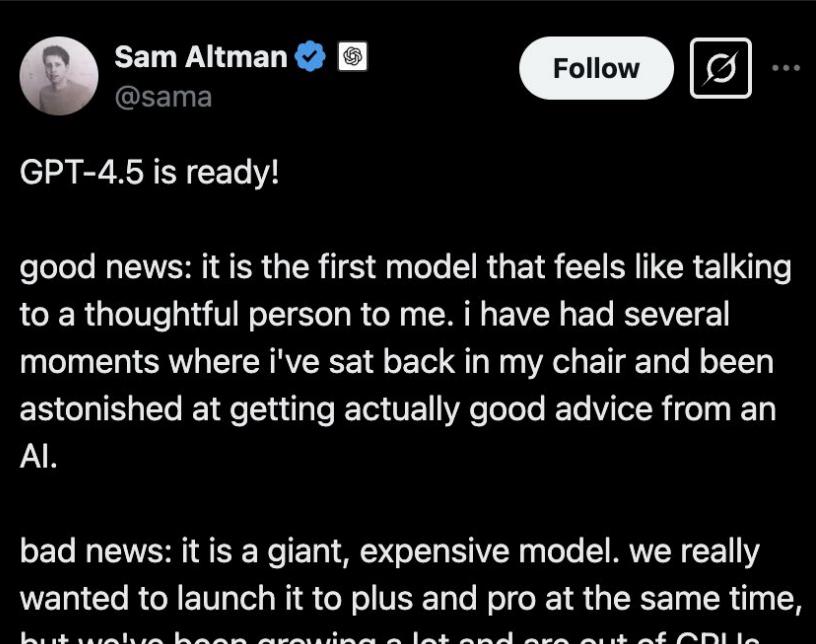
Compute is growing:

- Better hardware
- Better algorithms
- Larger clusters

Data is not growing:

- We have but one internet
- The fossil fuel of AI

Internet. We have, but one Internet. You could even say you can even go as far as to say. That data is the fossil fuel of AI. It was like, created somehow. And now we use it.



Sam Altman    ...

Follow   ...

GPT-4.5 is ready!

good news: it is the first model that feels like talking to a thoughtful person to me. i have had several moments where i've sat back in my chair and been astonished at getting actually good advice from an AI.

bad news: it is a giant, expensive model. we really wanted to launch it to plus and pro at the same time, but we've been growing a lot and are out of GPUs.

Test-time compute via RL

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

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

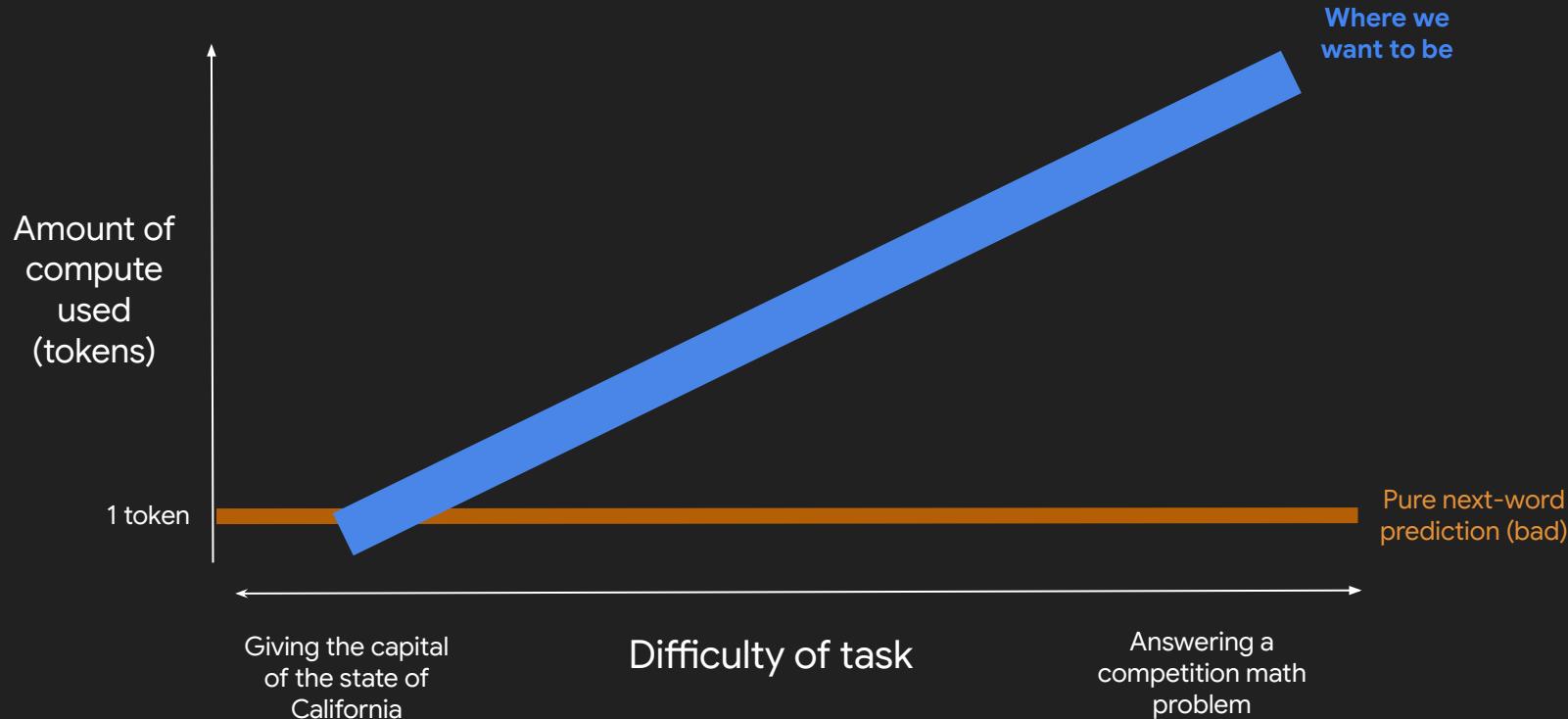
*Some words are super hard to
predict and take a lot of work*

When next-word prediction works fine

The screenshot shows the Playground interface with a dark theme. At the top, there's a "Playground" header with a "Complete" dropdown and a "Your presets" dropdown. Below that are buttons for "Save", "View code", "Share", and three more options. The main text area contains the sentence: "My name is Jason Wei and I am a researcher at OpenAI working on large language models." A modal window is open, showing the breakdown of the prediction: "models = 63.28%", "modeling = 11.41%", "model = 5.72%", "understanding = 3.98%", and "datasets = 3.93%". Below this, it says "Total: -0.46 logprob on 1 tokens (88.31% probability covered in top 5 logits)". At the bottom, there's a "Submit" button and some navigation icons.

When next-word prediction becomes very hard

The screenshot shows the same Playground interface. The main text area asks: "Question: What is the square of ((8-2)*3+4)^3 / 8?" with options (A) 1,483,492, (B) 1,395,394, and (C) 1,771,561. The answer is given as "(C)". A modal window shows the breakdown: "C = 32.09%", "B = 29.98%", "A = 27.97%", "D = 8.15%", and "c = 0.27%". Below this, it says "Total: -1.14 logprob on 1 tokens (98.44% probability covered in top 5 logits)". At the bottom, there's a "Submit" button and some navigation icons.



Naive 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.

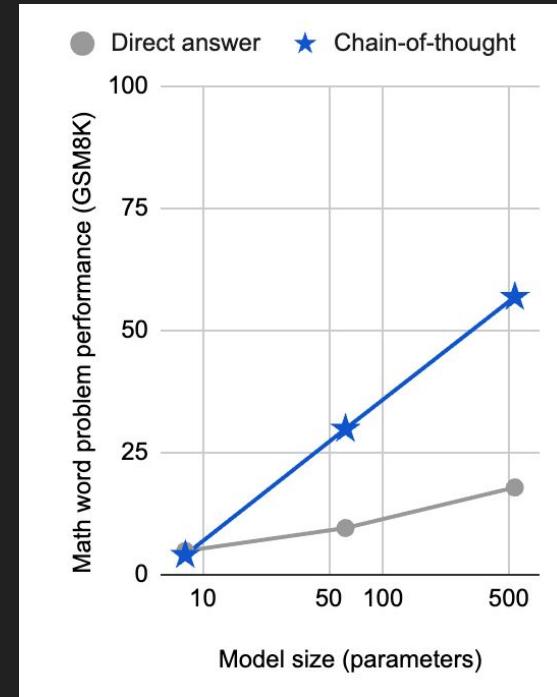
Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

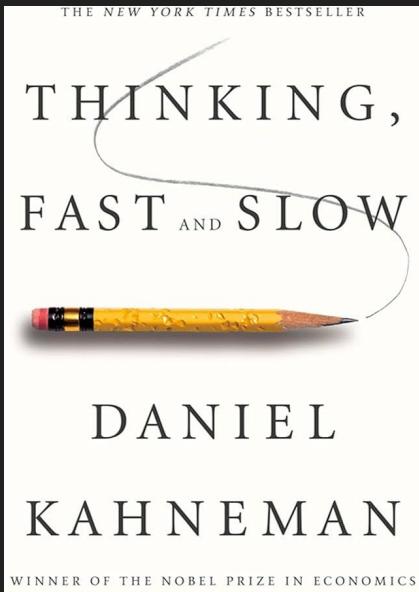
A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✓

Diagram showing the structure of the input:

- Question
- Chain of thought
- Answer
- Unseen input



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



WINNER OF THE NOBEL PRIZE IN ECONOMICS

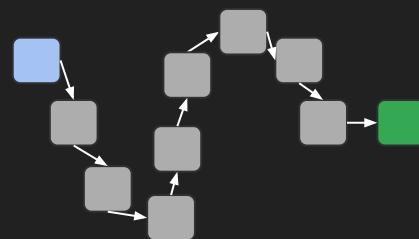
"[A] masterpiece . . . This is one of the greatest and most engaging collections of insights into the human mind I have read." —WILLIAM EASTERLY, *Financial Times*

<u>System 1: Fast, intuitive thinking</u>	<u>System 2: Slow, deliberate thinking</u>
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



OpenAI o1 (work of most of the company)

The screenshot shows the official OpenAI website. At the top, there's a navigation bar with the OpenAI logo, Research, Products, Safety, Company, and a search icon. Below the navigation is a large, dark header section featuring the date "September 12, 2024" and a prominent title "Learning to Reason with LLMs". Underneath the title is a descriptive paragraph about the new model. A "Contributions" button is visible. At the bottom of the page, there's a dark footer section containing text about the model's performance in competitive programming and other academic challenges.

September 12, 2024

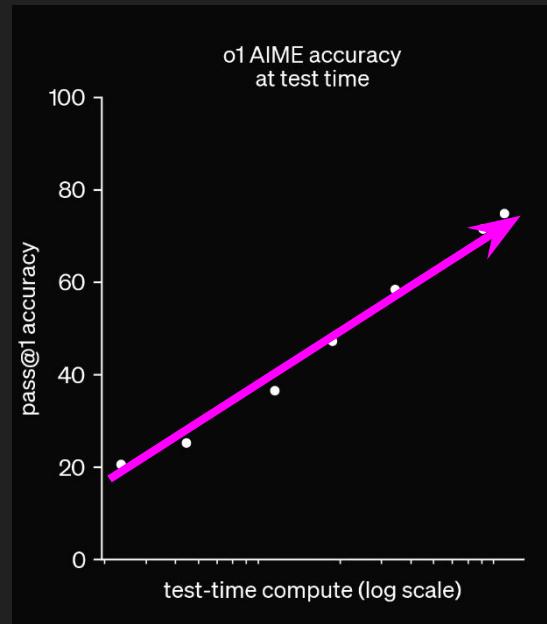
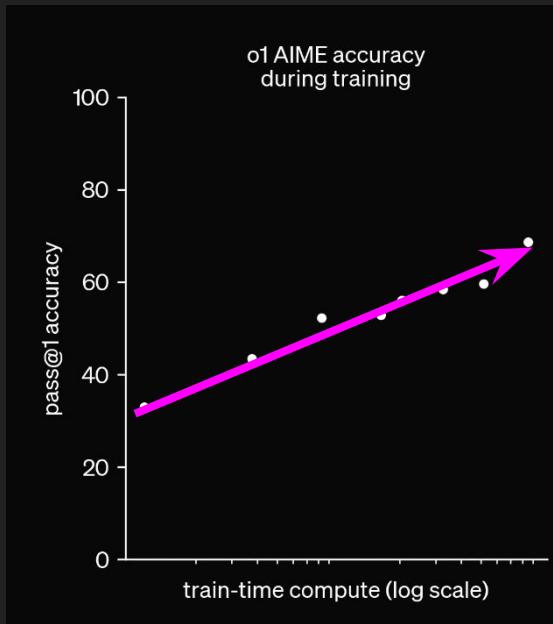
Learning to Reason with LLMs

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

Scale test-time compute



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



Imitation learning



Sora generated

Mimic micro-actions

Teacher forcing / planned curriculum

Imitation

Reinforcement learning



Daniil Medvedev, US Open 2020

Receive feedback from the trajectory

Training tailored for the model

Real learning

RL on the agent's own trajectories (“on-policy”) is important because the best action depends on the properties of the agent.

Regular tennis player:

- Orthodox ground strokes
- Reasonable court positioning



Daniil Medvedev:

- Unconventional (“ugly”) strokes
- Extreme court positioning



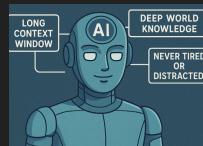
Human:

- Finite short-term memory
- Expert at few things
- Gets tired and distracted



AI:

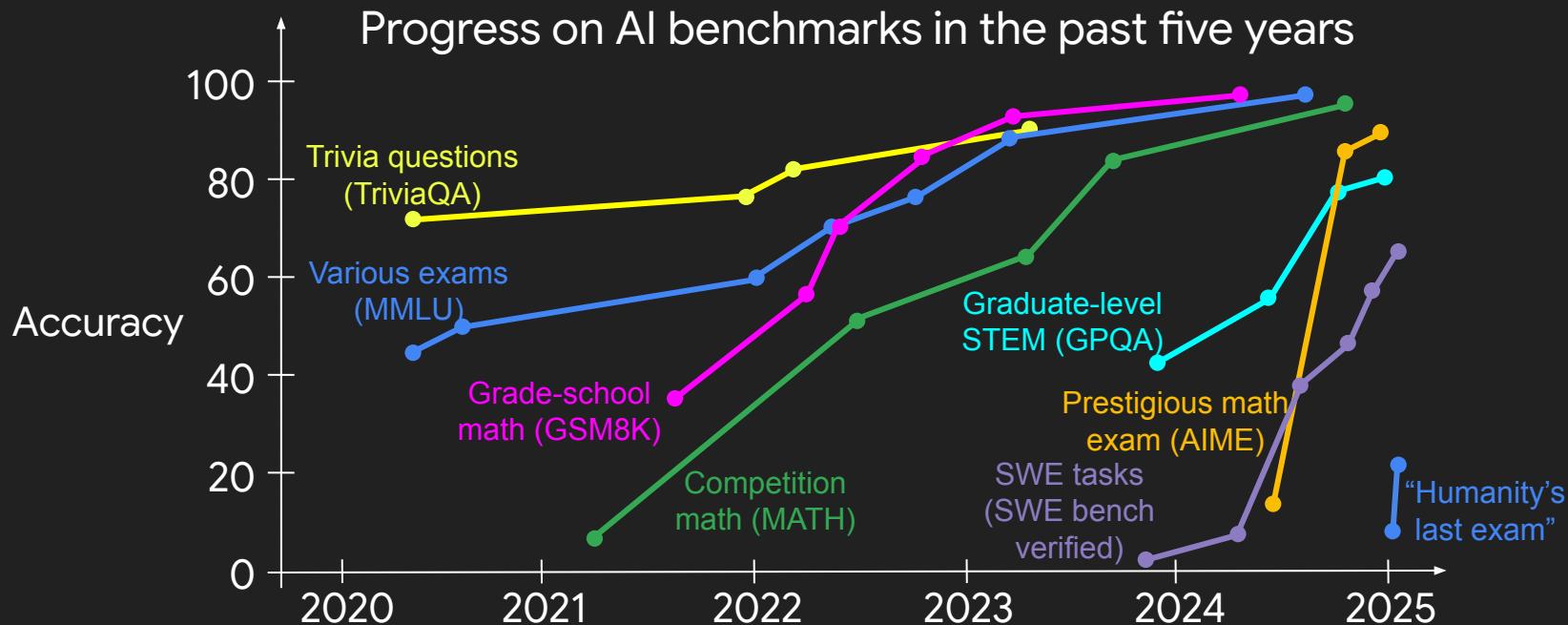
- Long context window
- Deep world knowledge
- Never tired or distracted



See also [“Deep dive into LLMs like ChatGPT” from Andrej Karpathy](#)

How will AI research evolve in the
next few years?

Trend 1: More effort will be spent measuring AI capabilities

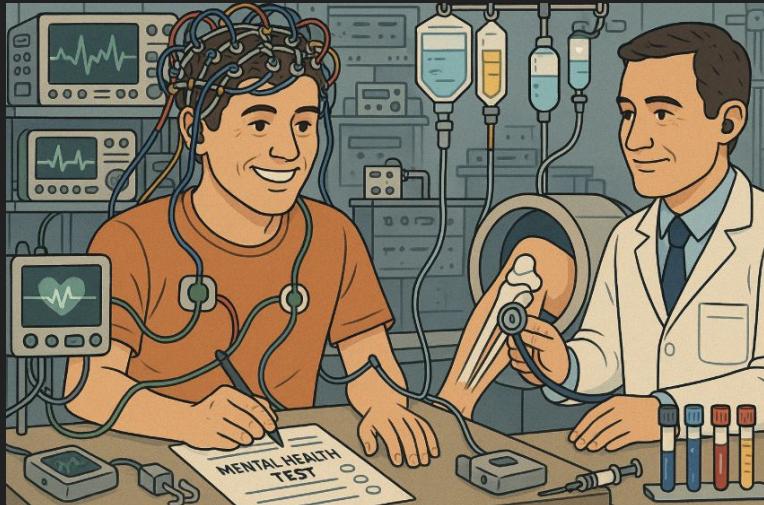


“Evals are dead, just use vibes?”

Not really...

- Good evals will be very valuable, and you can design them to have longevity
- We will be out of the era of a single eval for comparing models
- Eval will require more effort and context

Analogy: like measuring health of human body. Spectrum of evals with different domain, cost, precision, etc



A few benchmarks I'm personally excited about in 2025:

Traditional tasks at the edge of human knowledge

- Humanity's Last Exam (Phan et al., 2024)

SWE and AI research tasks

- SWE-Lancer (Miserendino et al., 2025)
- MLE-Bench (Shern et al., 2024)

Scientific research

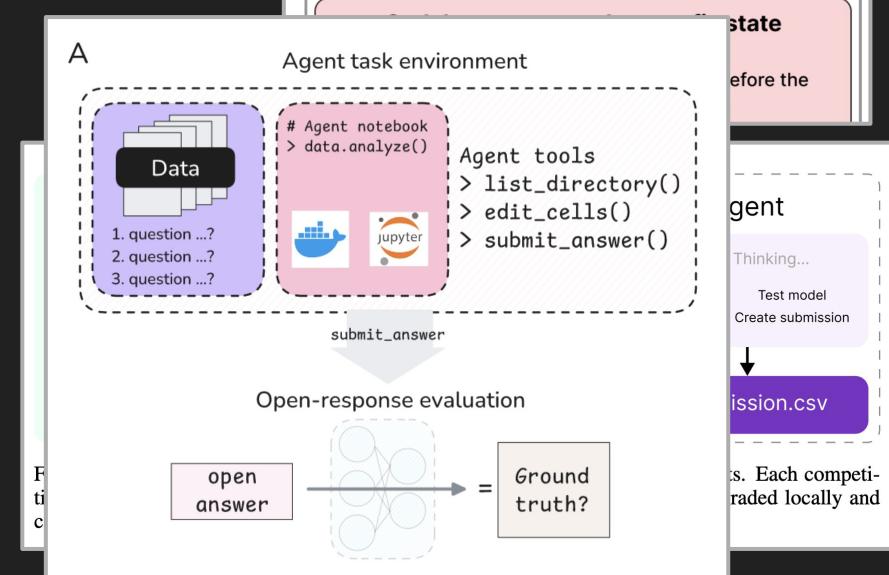
- LAB-Bench (Laurent et al., 2024)
- Bix-Bench (Mitchener et al., 2025)

Model	Accuracy (%) ↑
GPT-40	
GROK 2	
CLAUD	
GEMIN	
GEMIN	
O1	
DEEPSI	
O3-MIN	
O3-MIN	

IC SWE Task: Reverted codebase and original issue description

Original Issue (Problem, Description, Price)

Example: "[\\$8000] zip /postcode validation error message not displayed for entering , on the Home address screen #14958"



Trend 2: The field will continue to push the frontier of capability

“Levels of AI” ([Sam Altman on Levels of AI](#))

Level 1: Chatbots (2022 onwards)

Level 2: Reasoners (2024 onwards)

Level 3: Agents (2025 onwards)

Level 4: Innovators (202?)

Level 5: Organizations (20??)

Currently: tasks that take seconds to hours (OpenAI Operator, deep research models, Claude Code, Manus)

Eventually: hours to days of work

Tasks that are at the edge of human performance, or are totally new

- New scientific discoveries (Alpha*)
- Prize-winning writing
- Unsolved mysteries

Huge demand for specialists in other fields

Trend 3: More people will work on RL

Five years ago, RL was not particularly mainstream. Now, many more people at companies and in academia are working on RL. This trend towards RL will continue

Language models reached a threshold capability to reason in natural language

Good enough for RL?	
GPT-1	no
GPT-2	no
GPT-3	no (?)
GPT-4	yes

AI is already human-level on many tasks. If we want to push to super-human level, RL is a way to do that



Progress compounds. Many labs and researchers are betting on RL



Trend 4: The field will continue to overcome barriers of AI adoption

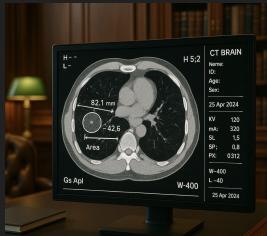
There is a huge overhang in AI adoption



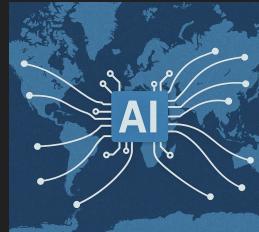
Infrastructure
and tooling



New interfaces



Reduced
friction and
easier to use



Awareness

What tasks will AI be able to
automate in our world?

How AI will change the world



“I looked into ChatGPT and it’s cool but it can’t really do any of the stuff in my job” - quant trader buddy of mine

“I think we have 2-3 more years of working as AI researchers before AI takes our own jobs” - AI researcher at another lab

How AI will change the world



Boaz Barak ✅
@boazbaraktcs

Follow



...

Moving between the worlds of East coast academics and SF industry is jarring in terms of the predictions of how AI will change the world in next ~5 years.

I think East coasters greatly underestimate the magnitude of change, and over index on temporary limitations of current systems.

On the other hand folks in the Bay Area are underestimating the friction and time lag in translating strong capabilities to wide scale economic impact.

6:15 AM · Jan 11, 2025 · 123.4K Views



33



86



806



221



roon ✅
@tszzi

Follow



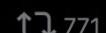
...

nobody should give or receive any career advice right now. everyone is broadly underestimating the scope and scale of change and the high variance of the future. your L4 engineer buddy at meta telling you "bro cs degrees are cooked" doesn't know shit

12:55 PM · Dec 23, 2024 · 1M Views



241



771



8.8K



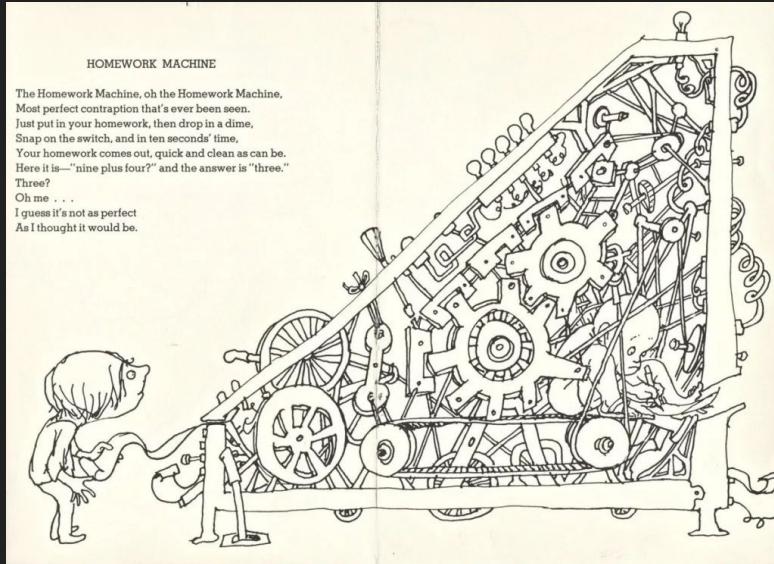
1.6K



AI is good at digital tasks

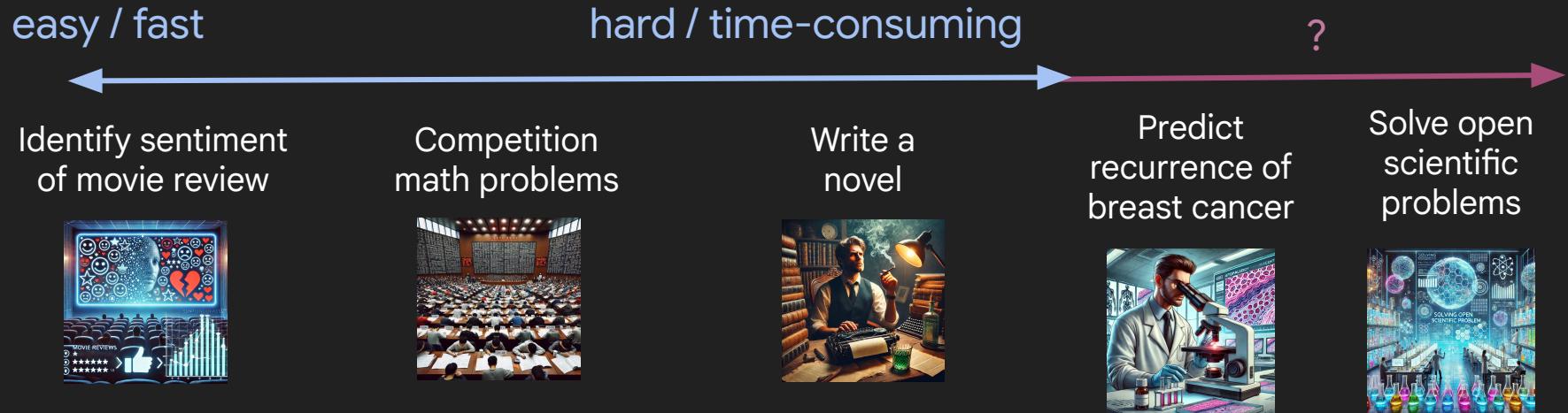


I, robot (2004) ?



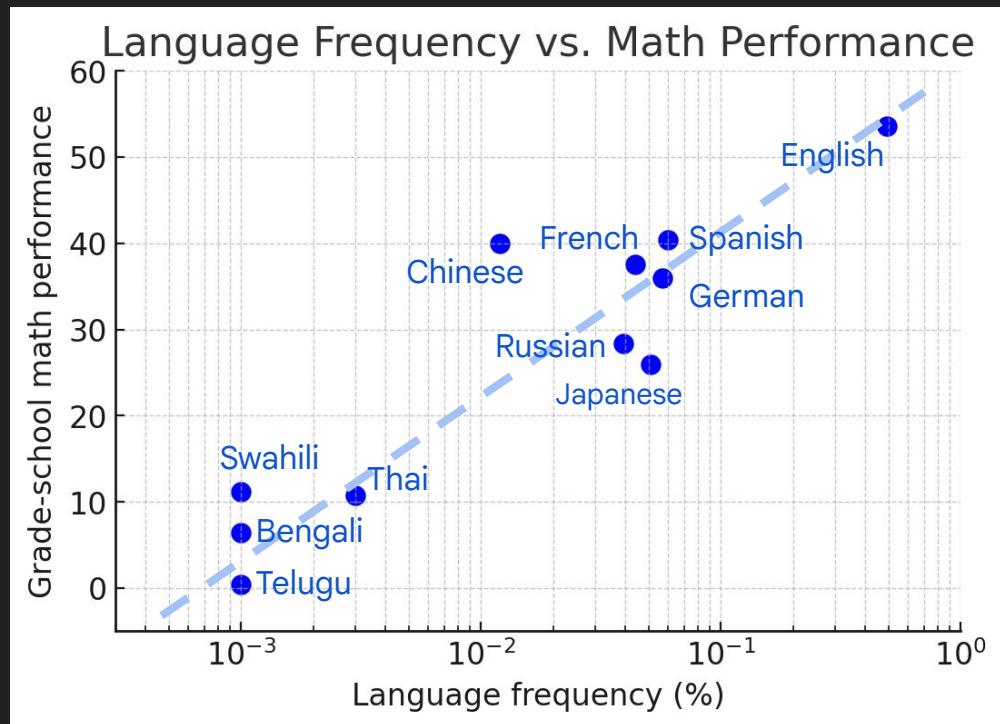
"Homework machine", Shel Silverstein (1981)

Tasks that are easier for humans tend to be easier for AI (as expected)



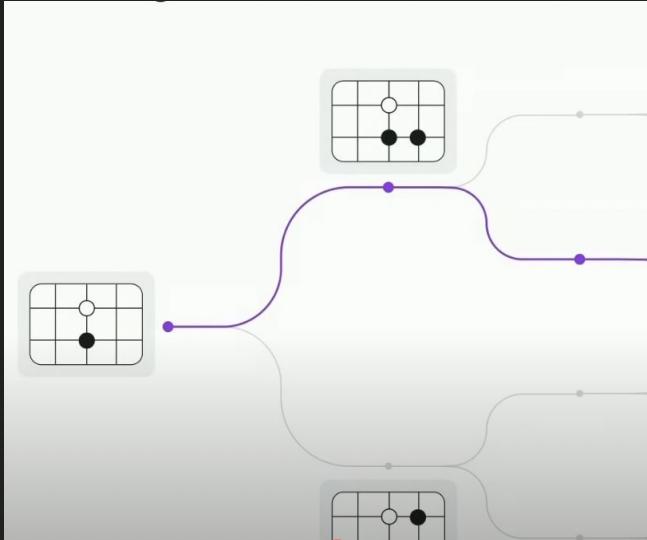
AI thrives when data is abundant

Math performance by thinking in a particular language is correlated with amount of data in that language



Language models are multilingual chain-of-thought reasoners, 2022.
Language frequency from W3Techs.

Special data unlock: when there is a single, objective metric, you can generate synthetic data via RL.



AlphaZero has a single, objective metric for performance.

Screenshot from "Nobel Prize lecture: Demis Hassabis, Nobel Prize in Chemistry 2024"

 **Denny Zhou** 
@denny_zhou

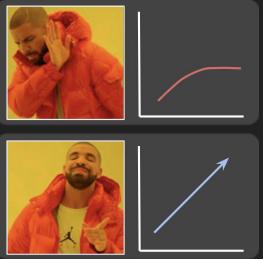
Follow  ...

any benchmark—including ARC-AGI—can be rapidly solved, as long as the task provides a clear evaluation metric that can be used as a reward signal during fine-tuning.

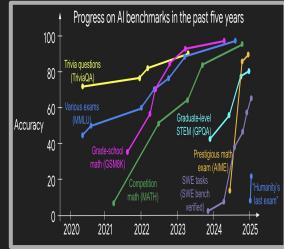
4:58 PM · Dec 20, 2024 · 109.7K Views

65 94 1.1K 242

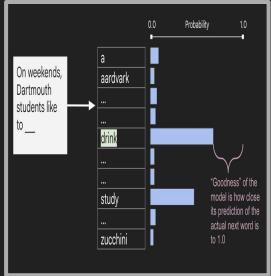
	<u>Difficulty for humans</u>	<u>Digital</u>	<u>Easy to get or create</u>	<u>Will AI be able to do it?</u>
Translate to top 50 languages	easy	yes	yes	~2018
Help debug basic code	medium	yes	yes	2023
Competition math	hard	yes	yes	2024
Conduct AI research	hard	yes	maybe	2026?
Conduct chemistry research	hard	no	maybe	later?
Make a movie	very hard	yes	yes	2027?
Stock market prediction	very hard	yes	yes	??
Translation to Tlingit	easy	yes	no	no
Fix your plumbing	medium	no	yes	?
Hairdressing	medium	no	yes	?
Traditional Uzbekistani carpet making	hard	no	no	no
Take girlfriend a date she is happy with	impossible	no	no	no



Scaling is an underlying driver of progress

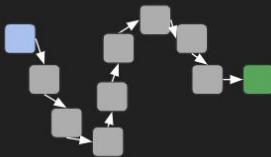


AI research will push frontier capabilities and bring AI to the world, emphasizing RL and evals

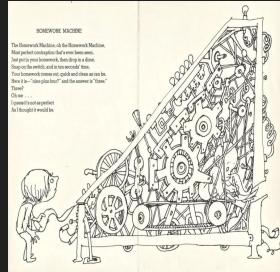


Predicting the next word (pre-training) taught AI about our world

Chain of thought



Test-time compute
via RL teaches AI to
do long-horizon
tasks



AI will first transform industries that are digital, easy for humans, and data-abundant

Thanks

X / Twitter: @_jasonwei

Feedback? <https://tinyurl.com/jasonwei>

Happy to take questions*

A chain of thought from OpenAI o1

First, let's understand what is being asked.

A

So both NH_4^+ and F^- can react with

W

V

of

F

N

V

Gi

b

th

Ka(

E

s

Gi

s

N

S

W

Ka\right)

$$F \quad 10^{-2}) = -1.5800$$

V

Le

p

Then:

pH = 7 + 0.5 × (-1.5800) = 7 -

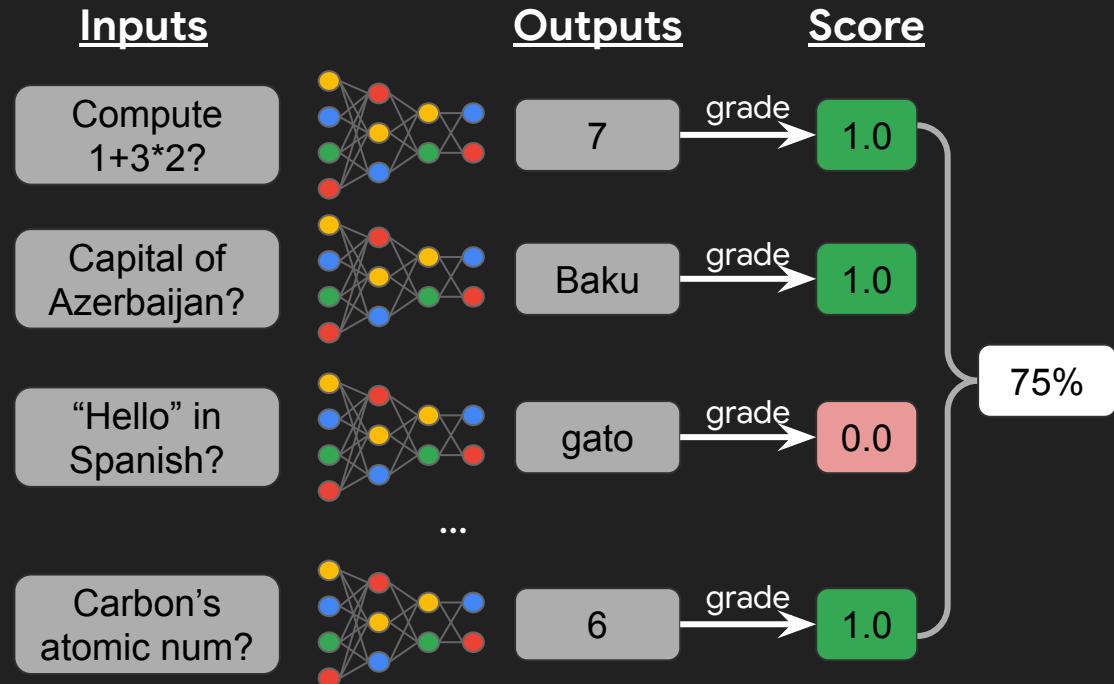
$$0.79 = 6.21$$

So

Therefore, the pH is approximately 6.21.

Evals are the measuring sticks in AI

Incentives that set the direction of the AI research community



Scaling was once non-obvious



Jerry Tworek  @MillionInt

Ilya Sutskever  @ilyasut

It's not a religion if it's true

9:28 PM · May 4, 2023 · 326.3K Views

87 120 829

This is a screenshot of a video thumbnail from a platform like Twitter or X. It features a video frame at the top showing two women speaking at a conference. Below the video are two Twitter-style cards. The first card is for Jerry Tworek (@MillionInt) and the second is for Ilya Sutskever (@ilyasut). Both cards include a profile picture, the name, the handle with a blue checkmark, and a short bio or quote. At the bottom of the thumbnail, there is a timestamp, the date, the view count (326.3K), and engagement metrics (87 comments, 120 retweets, and 829 likes).



Yann LeCun: Machine learning is great. But the idea that somehow we're going to just scale up the tec...
No, he said, we can't
Beyond 2030
Gary Marcus  @GaryMarcus

Warned everyone in 2022 that scaling would run out.
garymarcus.substack.com Joined December 2010

6,767 Following 169.7K Followers

This is a screenshot of a Substack profile for Gary Marcus (@GaryMarcus). The profile includes a photo of him smiling, a bio quote from Yann LeCun, and the title "Beyond 2030". Below the bio, it says "Warned everyone in 2022 that scaling would run out." and provides a link to his Substack page. At the bottom, it shows his following and follower counts.

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

...

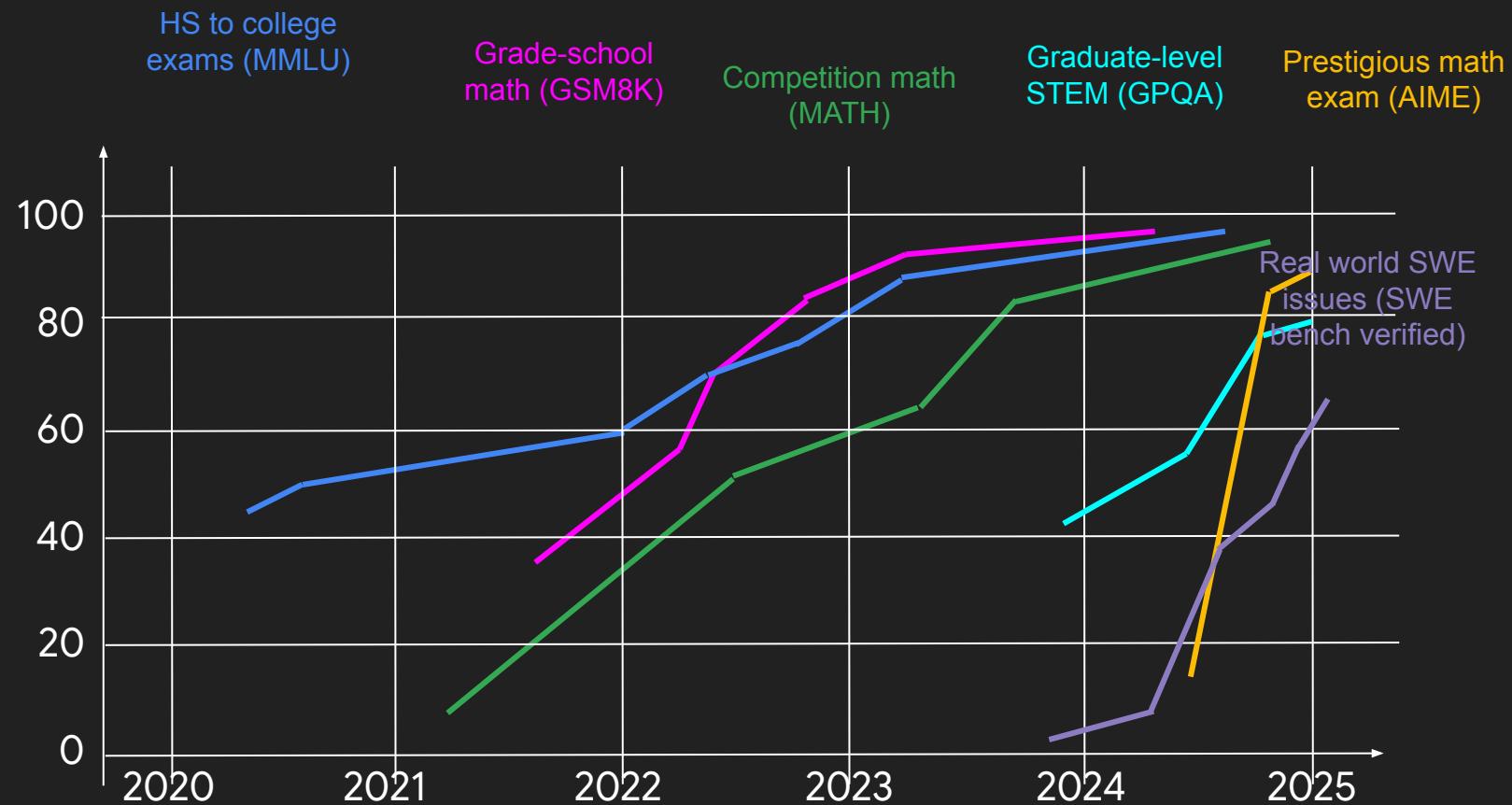
Tough right?

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

Across:			
Solve the		1. Eva	Now let's look at Down clues.
		2. On	1 Down: _____ car (station wagon) (6 letters)
		3. Mc	Possible words:
PlainT		4. Init	ESCAF
1 +	5. Na		- ESTATE car (6 letters)
2	6. Mi	AVOID	In British English, 'Estate car' is a term for station wagon.
3 +		DODG	Since 'station wagon' is called 'estate car' in the UK.
4	Down:		Therefore 'ESTATE' fits.
5 +	1. ---	ELUDE	Also aligns with ESCAPE as Across 1.
6	2. Au		
7 +	3. Pro	Maybe	
8	4. Sy		
9 +	5. An	But let	
10			
11 +			
12			
13 +			
	6. Deletes		



AI will almost certainly be good when all 4 criteria are met.
AI will struggle to automate when two or more are missing



Andrej Karpathy
@karpathy

[X] ...

I don't have too too much to add on top of this earlier post on V3 and I think it applies to R1 too (which is the more recent, thinking equivalent).

I will say that Deep Learning has a legendary ravenous appetite for compute, like no other algorithm that has ever been developed in AI. You may not always be utilizing it fully but I would never bet against compute as the upper bound for achievable intelligence in the long run. Not just for an individual final training run, but also for the entire innovation / experimentation engine that silently underlies all the algorithmic innovations.

Data has historically been seen as a separate category from compute, but even data is downstream of compute to a large extent - you can spend compute to create data. Tons of it. You've heard this called synthetic data generation, but less obviously, there is a very deep connection (equivalence even) between "synthetic data generation" and "reinforcement learning". In the trial-and-error learning process in RL, the "trial" is model generating (synthetic) data, which it then learns from based on the "error" (/reward). Conversely, when you generate synthetic data and then rank or filter it in any way, your filter is straight up equivalent to a 0-1 advantage function - congrats you're doing crappy RL.

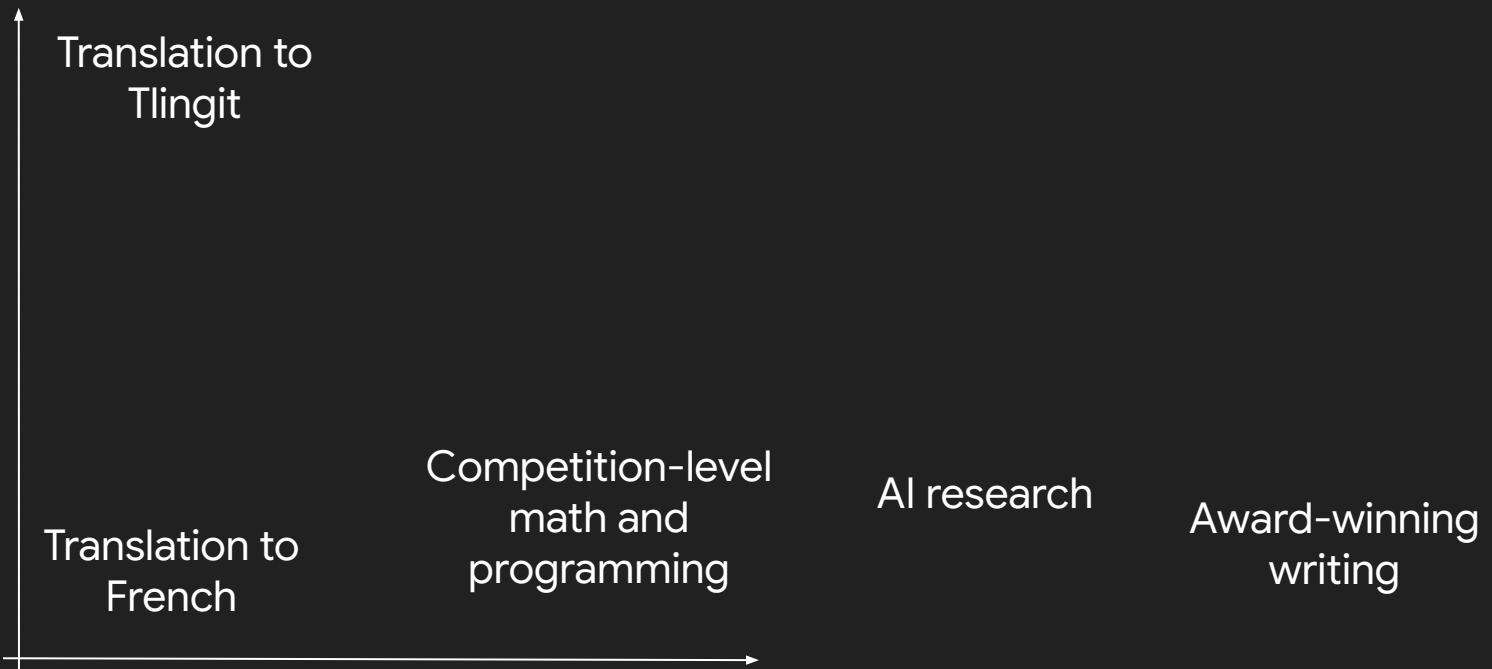
Last thought. Not sure if this is obvious. There are two major types of learning, in both children and in deep learning. There is 1) imitation learning (watch and repeat, i.e. pretraining, supervised finetuning), and 2) trial-and-error learning (reinforcement learning). My favorite simple example is AlphaGo - 1) is learning by imitating expert players, 2) is reinforcement learning to win the game. Almost every single shocking result of deep learning, and the source of all "magic" is always 2. 2 is significantly significantly more powerful. 2 is what surprises you. 2 is when the paddle learns to hit the ball behind the blocks in Breakout. 2 is when AlphaGo beats even Lee Sedol. And 2 is the "aha moment" when the DeepSeek (or of etc.) discovers that it works well to re-evaluate your assumptions, backtrack, try something else, etc. It's the solving strategies you see this model use in its chain of thought. It's how it goes back and forth thinking to itself. These thoughts are "emergent" (!!!) and this is actually seriously incredible, impressive and new (as in publicly available and documented etc.). The model could never learn this with 1 (by imitation), because the cognition of the model and the cognition of the human labeler is different. The human would never know to correctly annotate these kinds of solving strategies and what they should even look like. They have to be discovered during reinforcement learning as empirically and statistically useful towards a final outcome.

(Last last thought/reference this time for real is that RL is powerful but RLHF is not. RLHF is not RL. I have a separate rant on that in an earlier tweet
x.com/karpathy/status...)

Why does scaling work?

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

<u>Small language model</u>	<u>Large language model</u>
Memorization is costly	More generous with memorizing tail knowledge
First-order correlations	Complex heuristics





Andrej Karpathy

@karpathy



...

I don't have too too much to add on top of this earlier post on V3 and I think it applies to R1 too (which is the more recent, thinking equivalent).

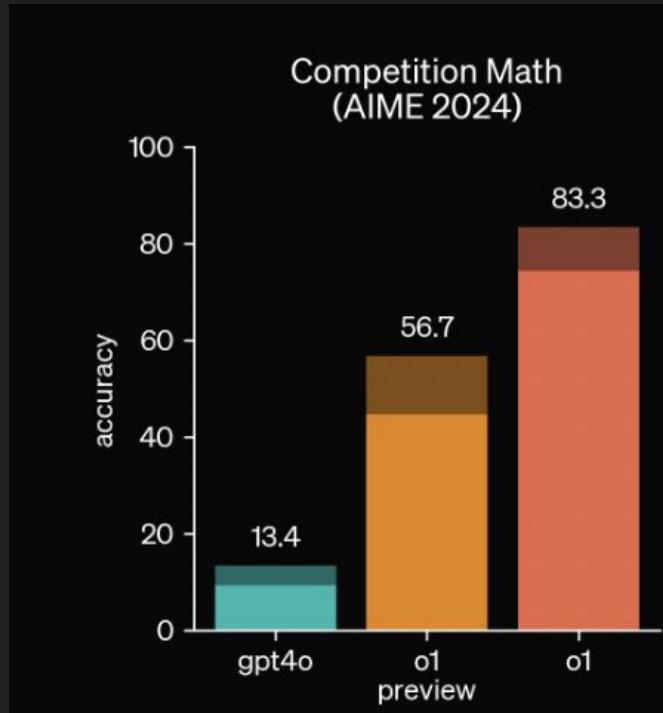
I will say that Deep Learning has a legendary ravenous appetite for compute, like no other algorithm that has ever been developed in AI. You may not always be utilizing it fully but I would never bet against compute as the upper bound for achievable intelligence in the long run. Not just for an individual final training run, but also for the entire innovation / experimentation engine that silently underlies all the algorithmic innovations.

Data has historically been seen as a separate category from compute, but even data is downstream of compute to a large extent - you can spend compute to create data. Tons of it. You've heard this called synthetic data generation, but less obviously, there is a very deep connection (equivalence even) between "synthetic data generation" and "reinforcement learning". In the trial-and-error learning process in RL, the "trial" is model generating (synthetic) data, which it then learns from based on the "error" (/reward). Conversely, when you generate synthetic data and then rank or filter it in any way, your filter is straight up equivalent to a 0-1 advantage function - congrats you're doing crappy RL.

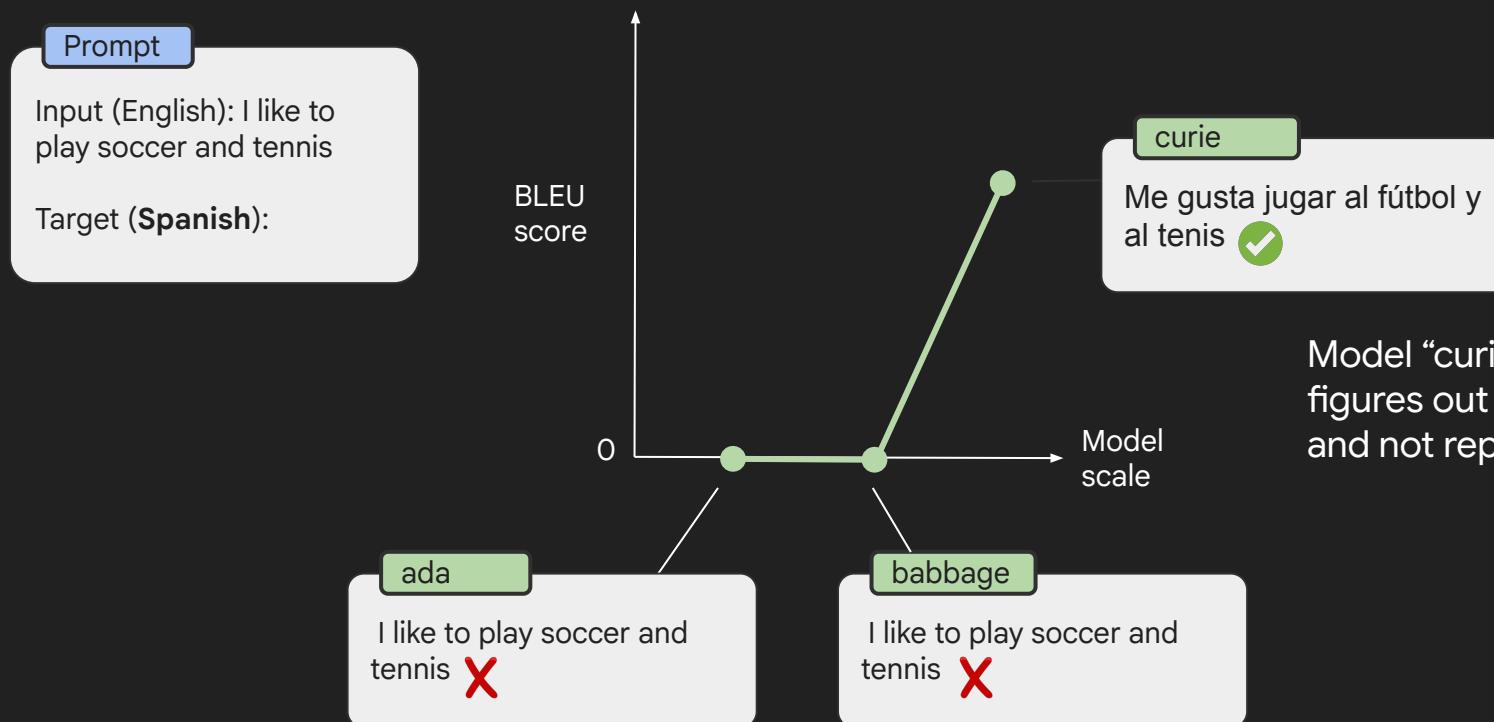
Last thought. Not sure if this is obvious. There are two major types of learning, in both children and in deep learning. There is 1) imitation learning (watch and repeat, i.e. pretraining, supervised finetuning), and 2) trial-and-error learning (reinforcement learning). My favorite simple example is AlphaGo - 1) is learning by imitating expert players, 2) is reinforcement learning to win the game. Almost every single shocking result of deep learning, and the source of all "magic" is always 2. 2 is significantly significantly more powerful. 2 is what surprises you. 2 is when the paddle learns to hit the ball behind the blocks in Breakout. 2 is when AlphaGo beats even Lee Sedol. And 2 is the "aha moment" when the DeepSeek (or of etc.) discovers that it works well to re-evaluate your assumptions, backtrack, try something else, etc. It's the solving strategies you see this model use in its chain of thought. It's how it goes back and forth thinking to itself. These thoughts are "emergent" (!!!) and this is actually seriously incredible, impressive and new (as in publicly available and documented etc.). The model could never learn this with 1 (by imitation), because the cognition of the model and the cognition of the human labeler is different. The human would never know to correctly annotate these kinds of solving strategies and what they should even look like. They have to be discovered during reinforcement learning as empirically and statistically useful towards a final outcome.

(Last last thought/reference this time for real is that RL is powerful but RLHF is not. RLHF is not RL. I have a separate rant on that in an earlier tweet
x.com/karpathy/status...)

Scale RL on chain-of-thought



Emergence ability example



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-ESTIMATION parameter ϵ must use $\Omega(\min\{m, n, \epsilon^{-1}\})$ space. In particular, in order to get ϵ -accuracy with $m = n$, the space required must use $\Omega(\min\{m, n\})$ space.

Proof. We will prove the stronger result that the simpler FREQUENCY-ESTIMATION problem which asks whether the input stream contains a token whose frequency is at least ϵn requires space in the deterministic setting. Since the cost of the randomized problem is at most twice that of the deterministic algorithm, this will prove the theorem as a whole.

Let \mathcal{A} be a one-pass S -space deterministic algorithm for FREQUENCY-ESTIMATION. Alice sends a query (x, y) to Bob. Bob runs \mathcal{A} on the input stream (x, y) for the IDX_N . Alice creates a stream $\sigma_1 = (a_1, a_2, \dots, a_N)$ where $a_i = 1$ if $x_i = y$ and 0 otherwise. Bob creates a stream $\sigma_2 = (b, b, \dots, b)$ of length $k - 1$ for $k \geq 2$ where $b = 2y - 1$. Bob runs \mathcal{A} on the combined stream $\sigma_1 \circ \sigma_2$ with parameter k .

The output of $\text{IDX}_N(x, y)$ is 1 iff \mathcal{A} produces b as output. This is so because \mathcal{A} is deterministic and will be the unique entry with $f_b = k \geq k$. Thus Alice and Bob can solve the FREQUENCY-ESTIMATION problem by Alice sending her query to Bob using \mathcal{A} .

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

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

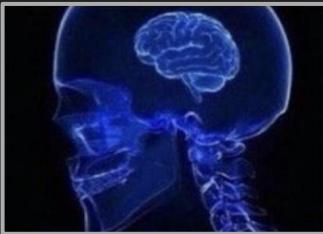
OK I think this is the right answer!

Paradigm 2: Scaling reinforcement learning 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.

2019



2024



2029



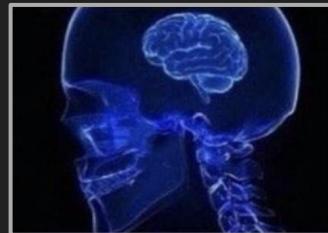
- Can barely write a coherent paragraph
- Can't do any reasoning

- 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.

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?

Scaling is hard and was not obvious at the time

Technical & operational challenges



(1) Distributed training requires a lot of expertise

Image source: HF

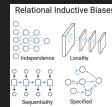


(2) Loss divergences and hardware failures are hurdles



(3) Compute is expensive

Psychological challenges



(1) Researchers like inductive biases

Image source



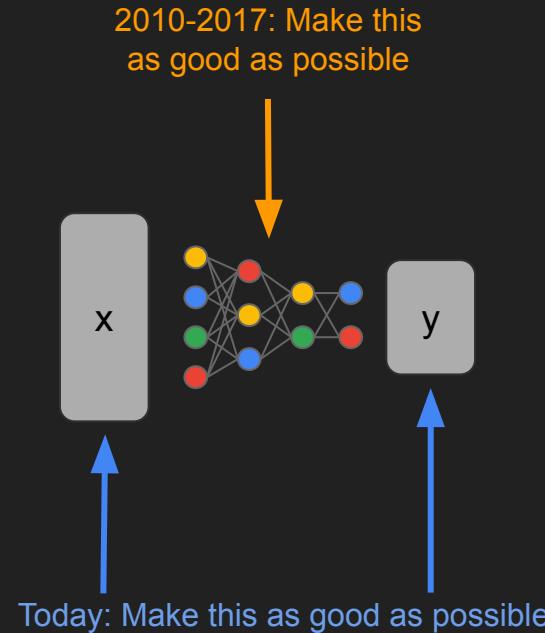
(2) Scaling is different from
human learning



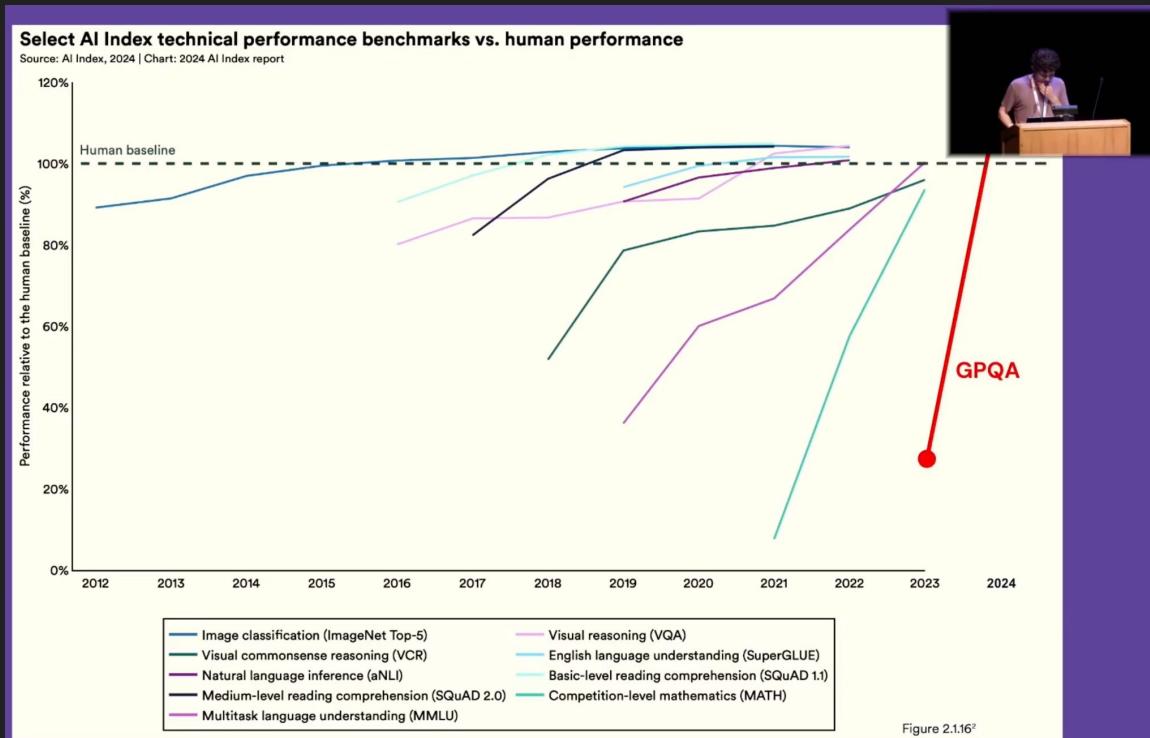
(3) Scientific research incentives
don't match engineering work
("novelty")

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

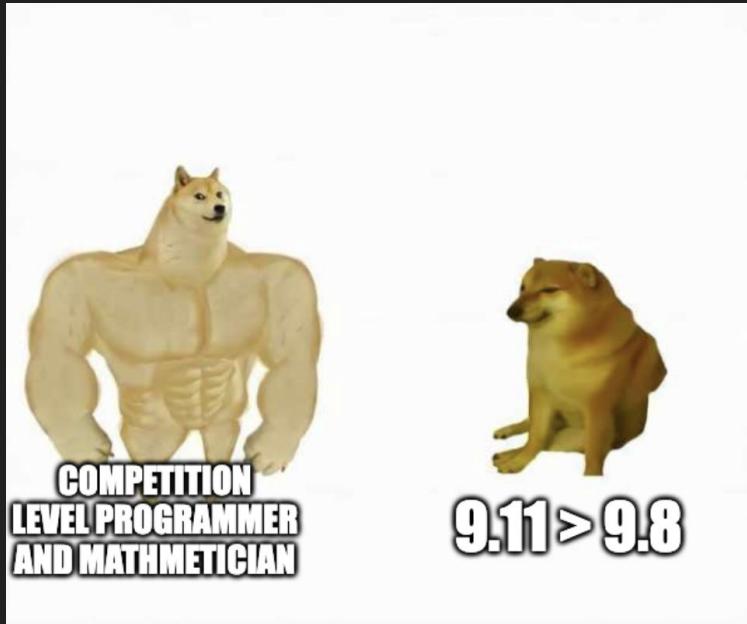


"People ask me if I'm making an even harder version of GPQA... [well] we set out to make the hardest science benchmark that we could"
- David Rein

Academia has had a huge impact!

SQuAD, GLUE/SuperGLUE,
MMLU, MATH, GPQA

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

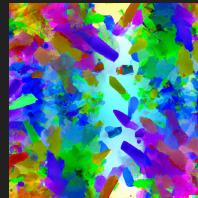
Intelligence != user experience

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