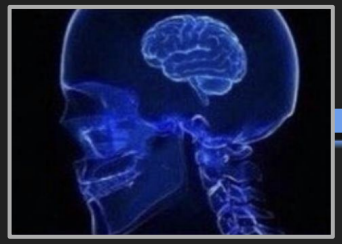
Driving forces in AI in 2025

Jason Wei
Research Scientist
OpenAI

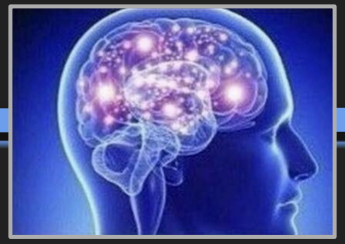
(Opinions are my own and do not reflect my employer.)

2020



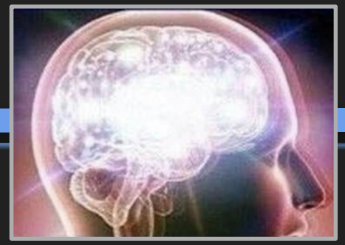
- Can barely write a coherent paragraph
- Can't do any reasoning

2025



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

2030



?

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

Talk outline

Motivate the driving forces in AI in 2025:

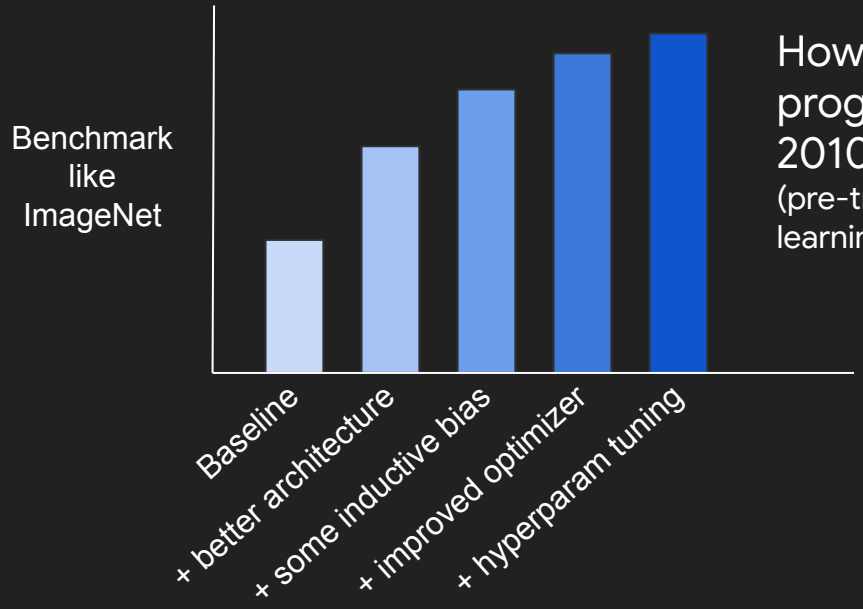
0. Scaling
1. Pre-training
2. Test-time compute via RL

Try to extrapolate into the next few years:

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

Scaling

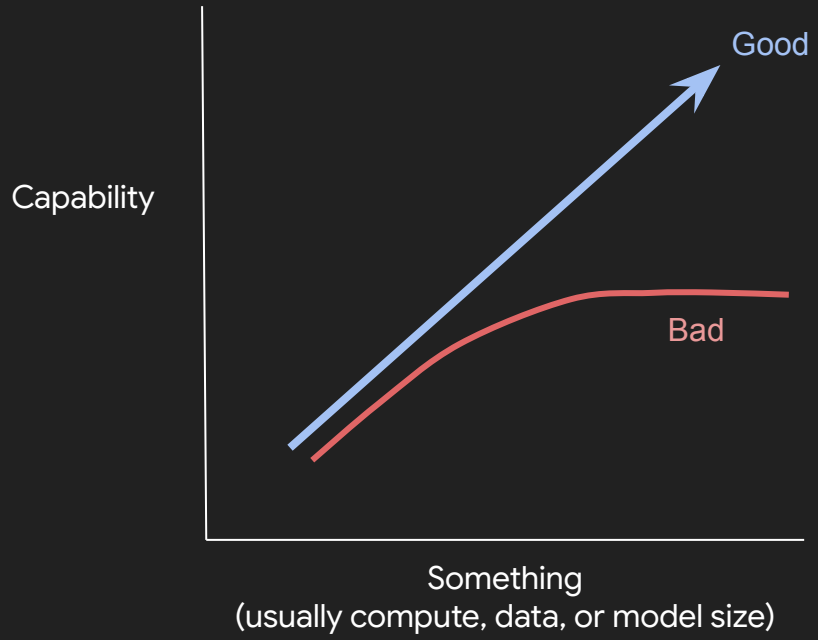
“Studying the past tells you what’s special about the current moment.”

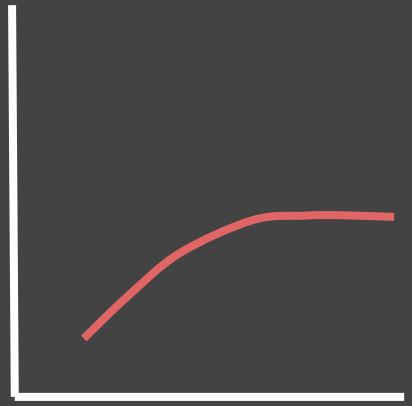


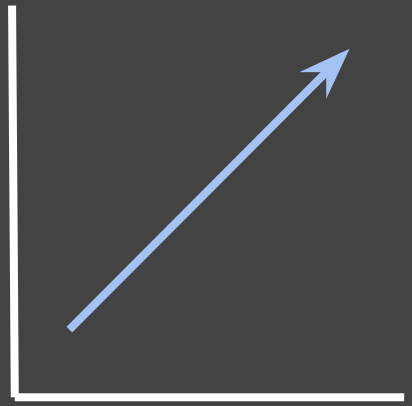
How we made progress, early 2010s to 2017 (pre-transformer deep learning)

What is scaling?

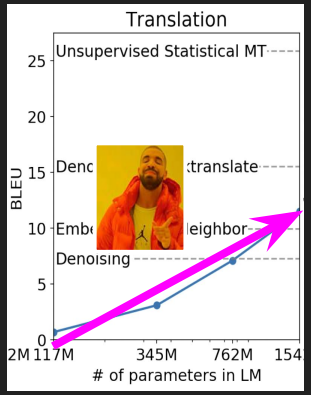
Scaling is when you put yourself in a situation where you move along a continuous axis and expect sustained improvement.



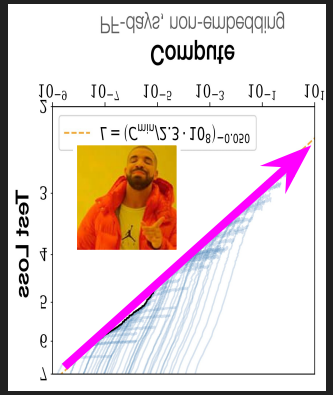




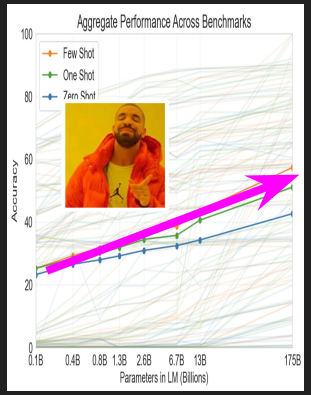
Scaling is everywhere



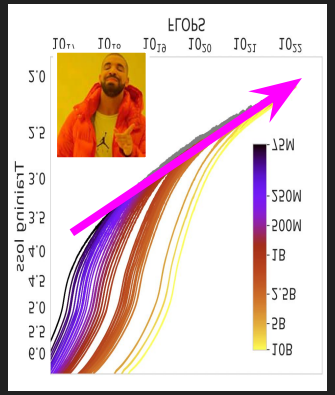
GPT-2 (2019)



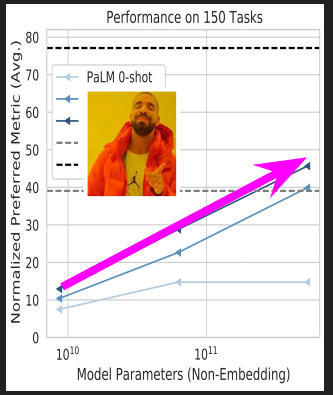
Scaling laws (2020)



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

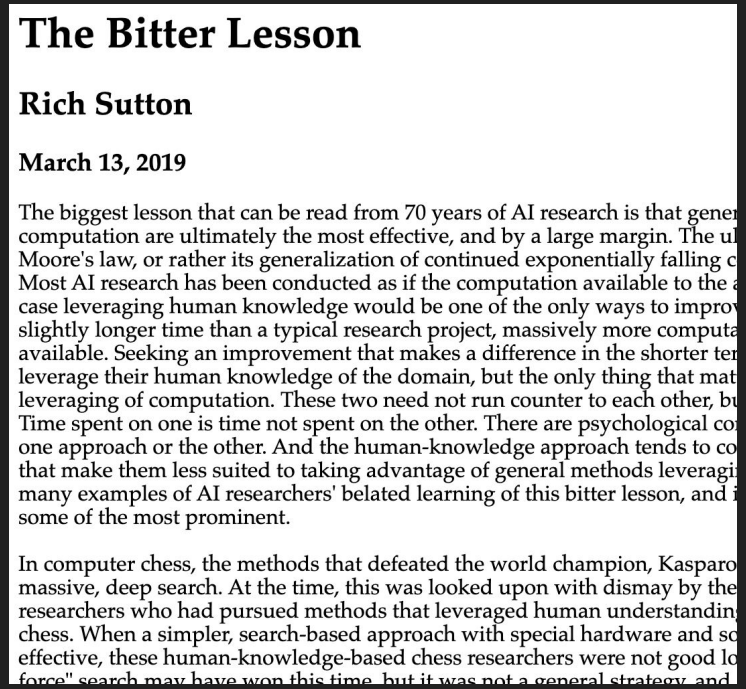
You can reliably improve capability (even if it's expensive)

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 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 understandin
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

Pre-training

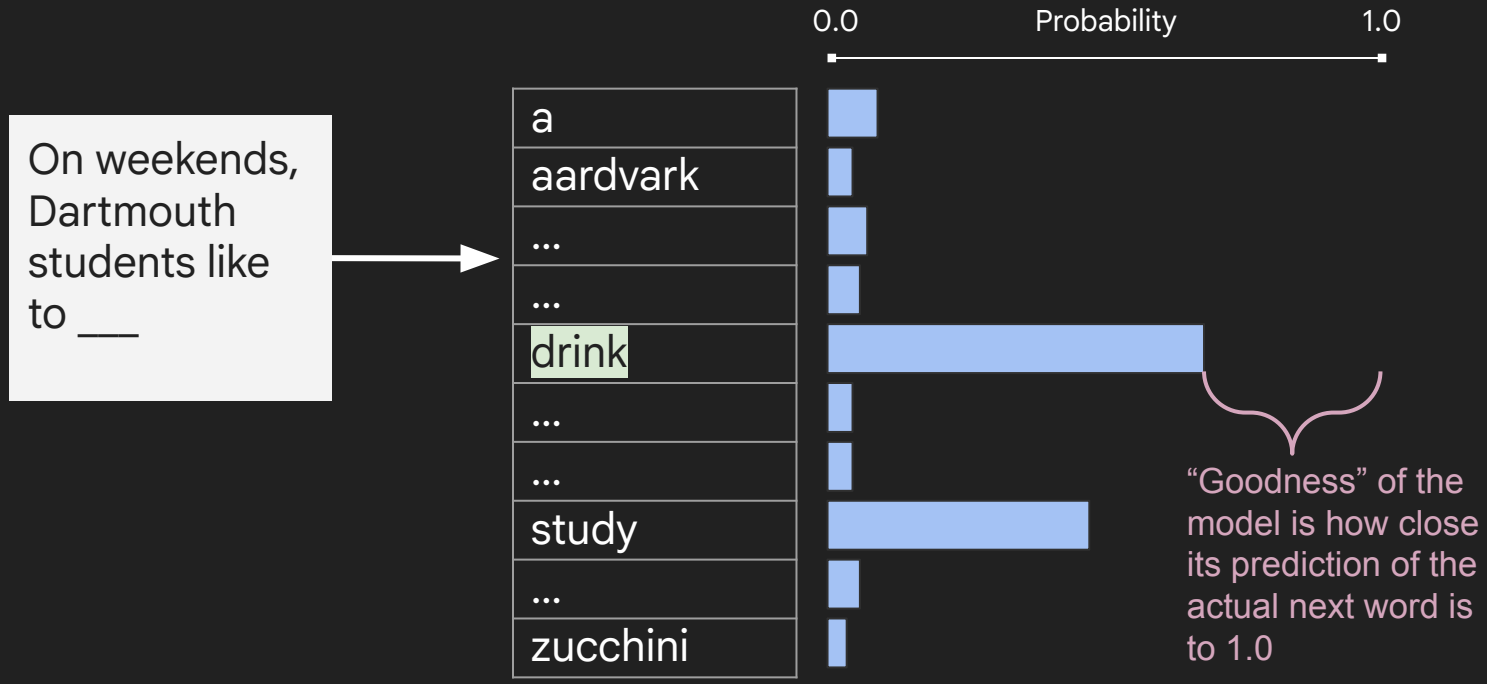
What training task do we scale?

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



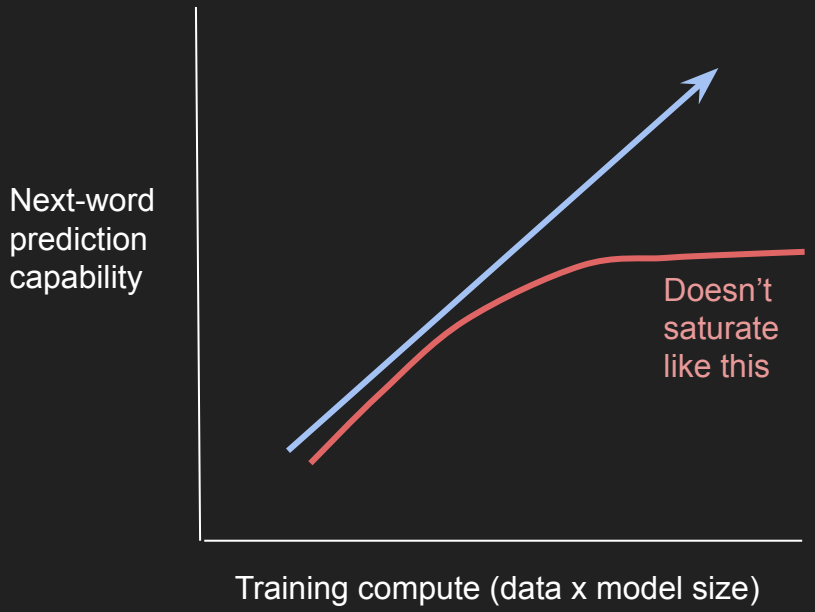
You learn a lot about the world by predicting the next word

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 “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.

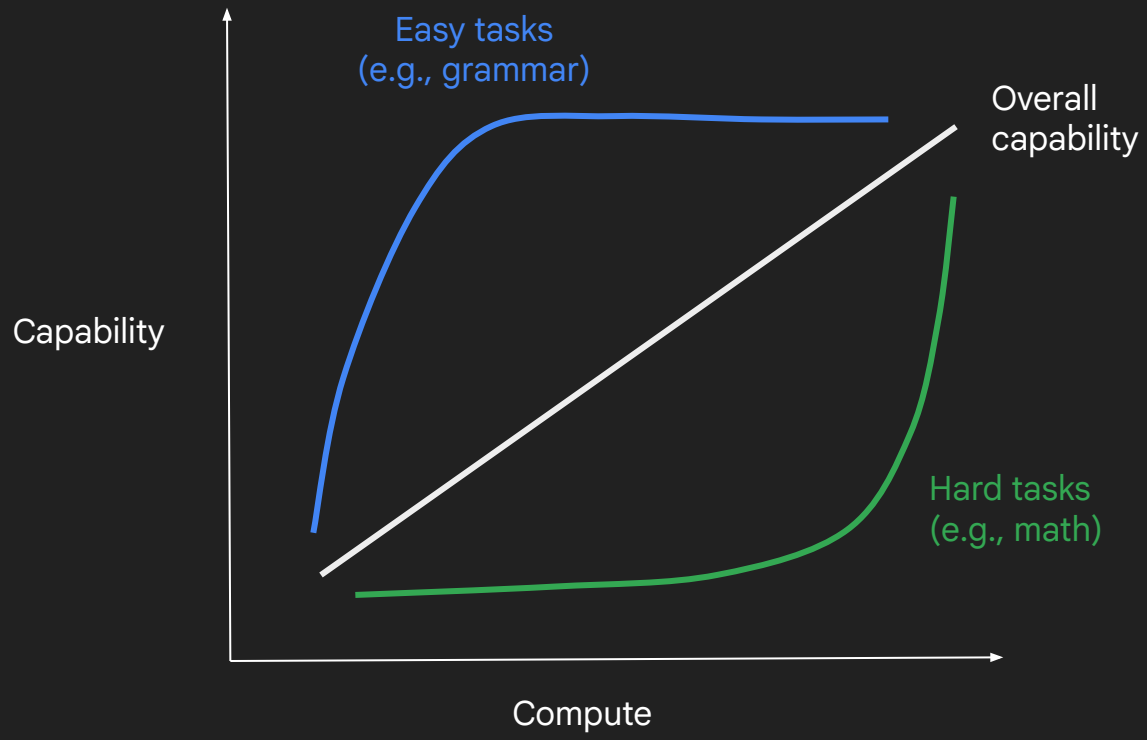
If scaling was so predictable, why was the success of this paradigm so surprising?

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
                   ...
```

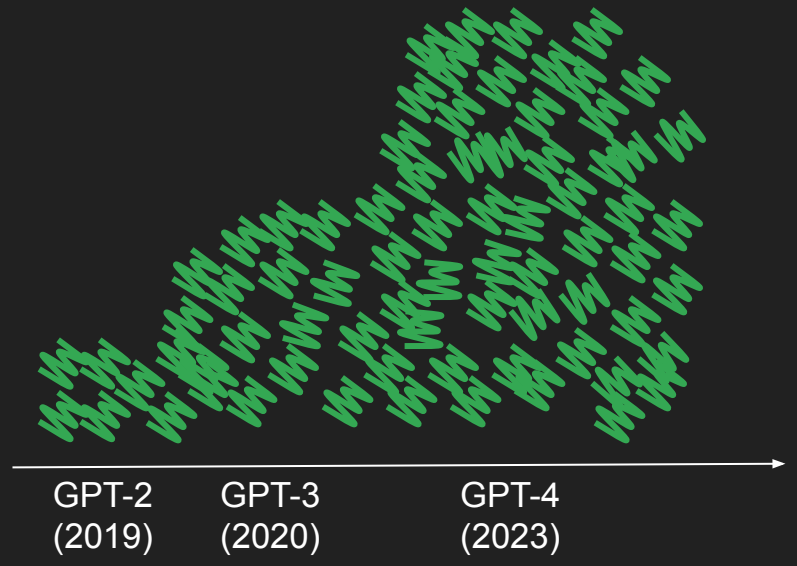
🤔 *If accuracy goes from 70% to 80%, do all tasks get better uniformly?*
...probably not.



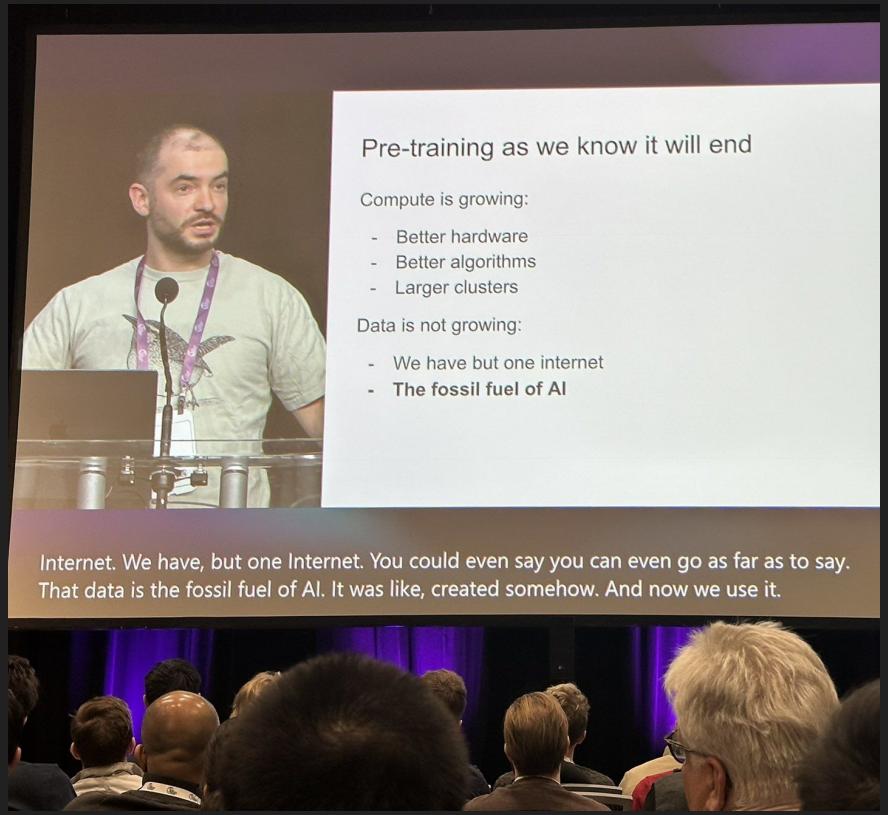
"Emergent abilities" /
"phase transition"

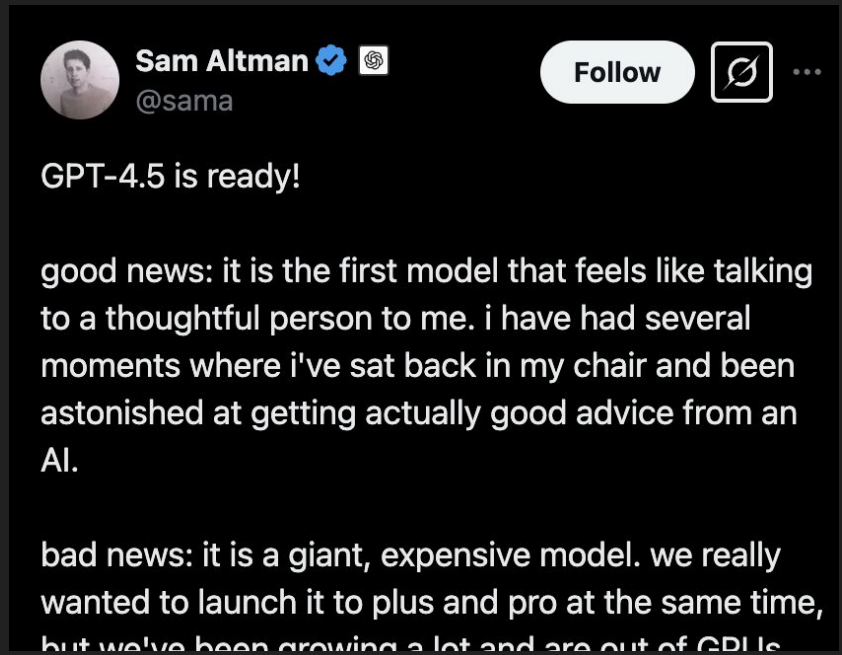
“Spectrum of possible tasks”

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



The future of pre-training?





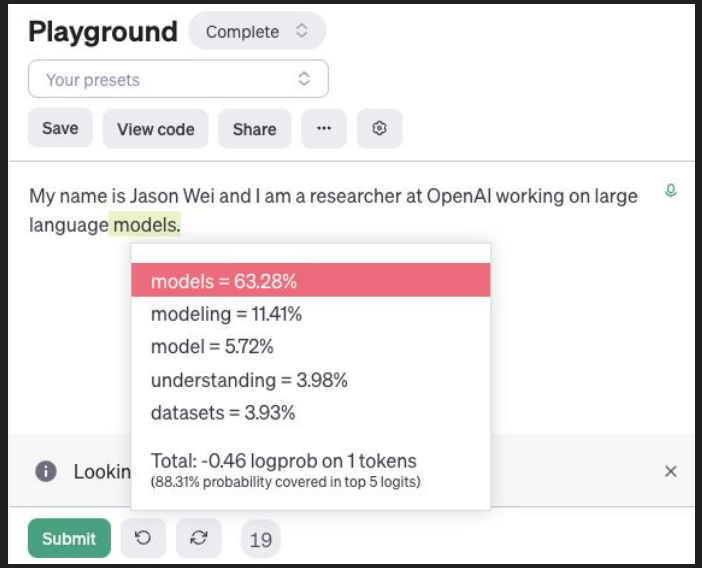
Test-time compute via RL

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

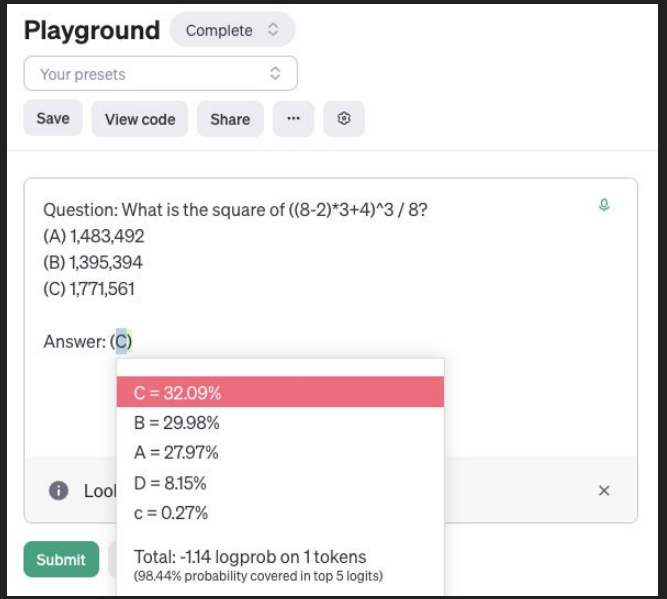
Maybe (it would be hard), but there is a bottleneck:

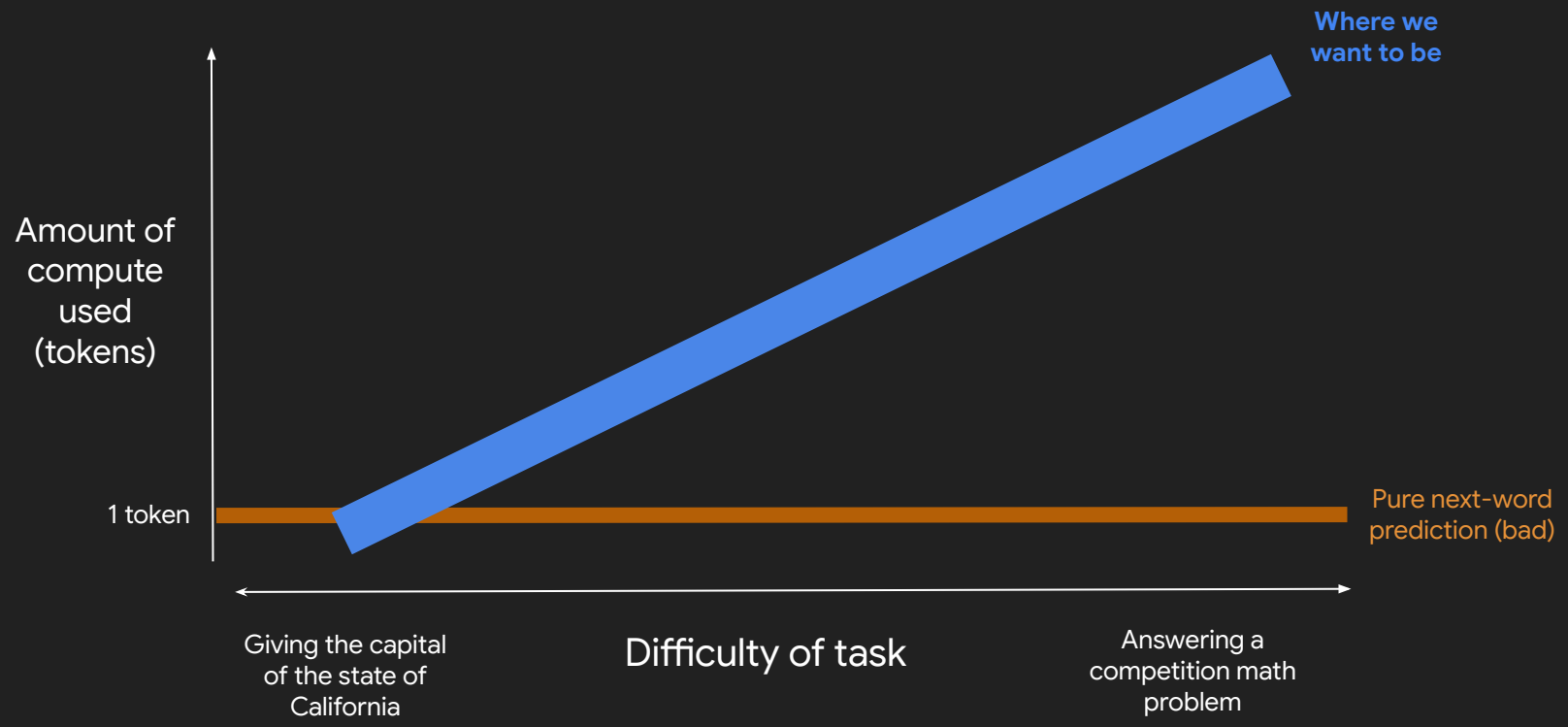
Some words are super hard to predict and take a lot of work

When next-word prediction works fine



When next-word prediction becomes very hard




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
Answering a competition math problem

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

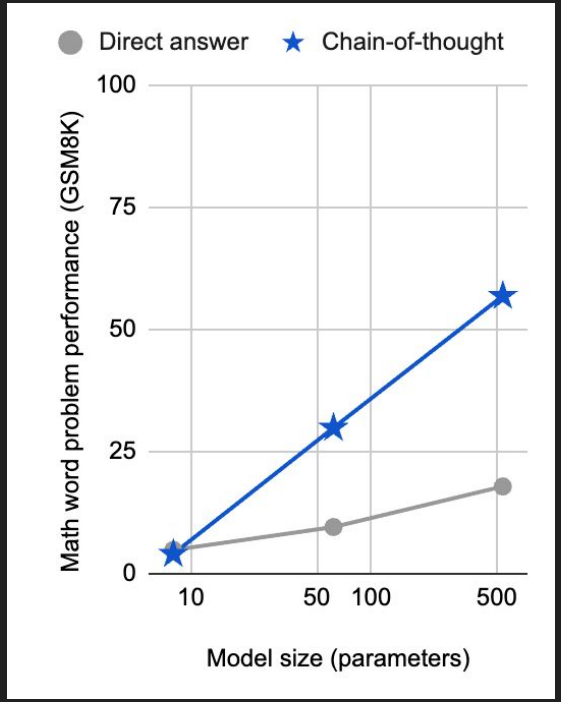
A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

Question

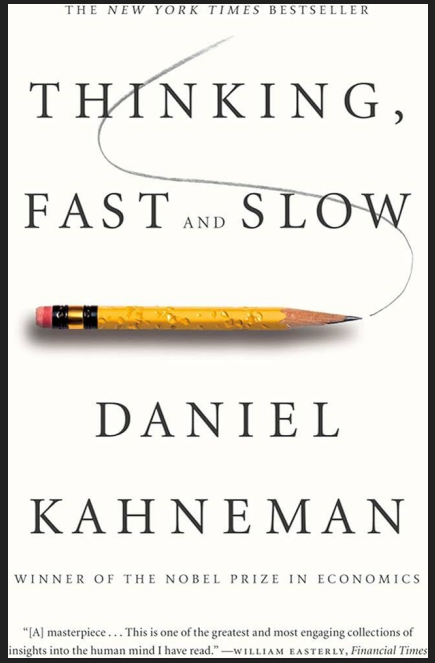
Chain of thought

Answer

Unseen input



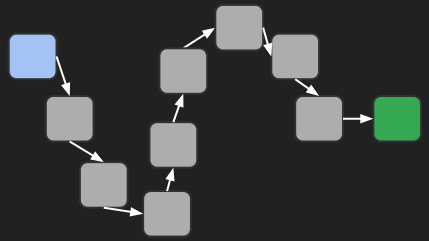
Chain-of-thought prompting elicits reasoning in large language models, 2022.



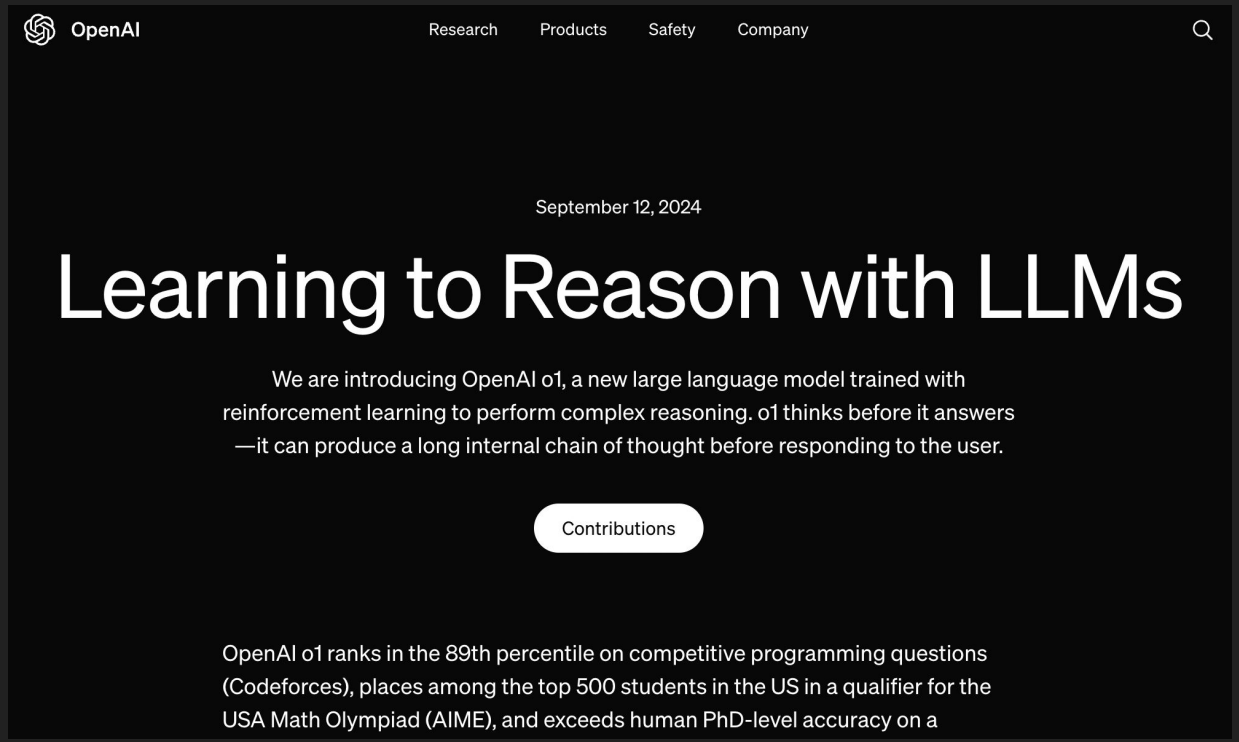
System 1: Fast, intuitive thinking	System 2: Slow, deliberate thinking
Automatic Effortless Intuitive Emotional	Conscious Effortful Controlled Logical
Recognizing faces Repeating basic facts Reacting to something	Solving math problems Planning a detailed agenda Making a thoughtful decision

Next-word prediction

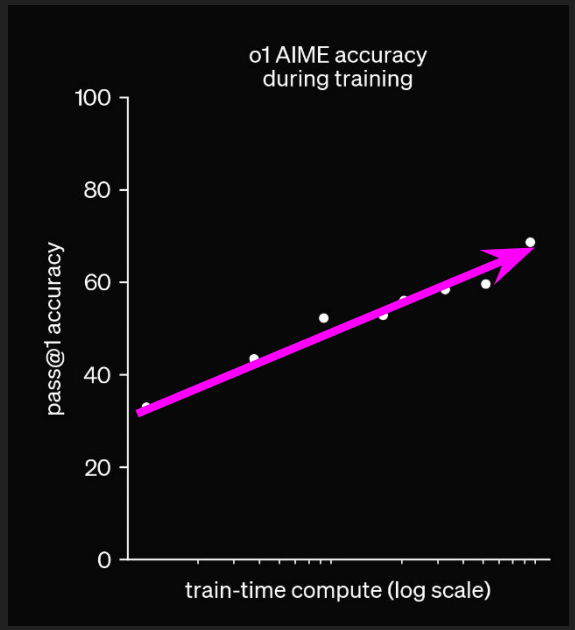
Chain of thought

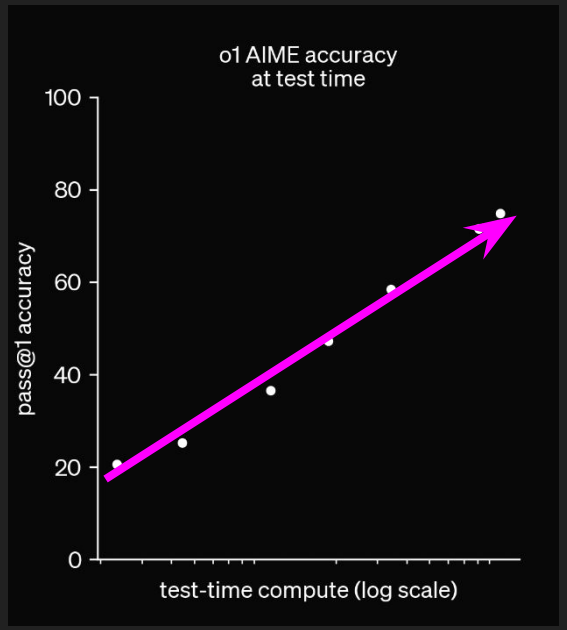


OpenAI o1 (work of most of the company)



Scale test-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

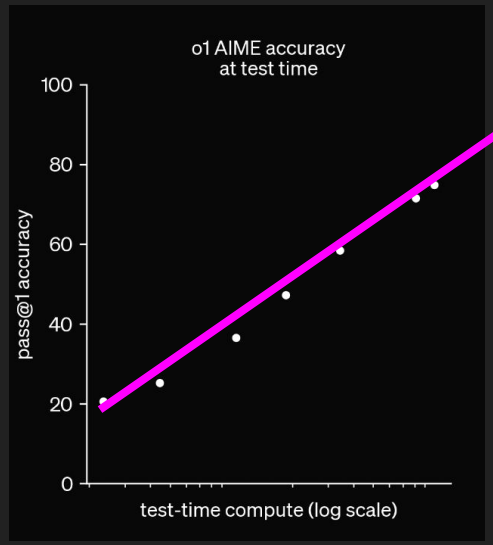
Write the code, documentation, and research paper for the best way to make AI safe

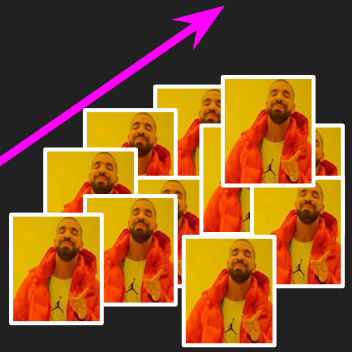
Hypothetical response

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***

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

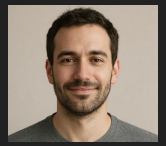
Real learning

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

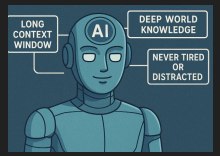
Regular tennis player:
- Orthodox ground strokes
- Reasonable court positioning

Daniil Medvedev:
- Unconventional ("ugly") strokes
- Extreme court positioning

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



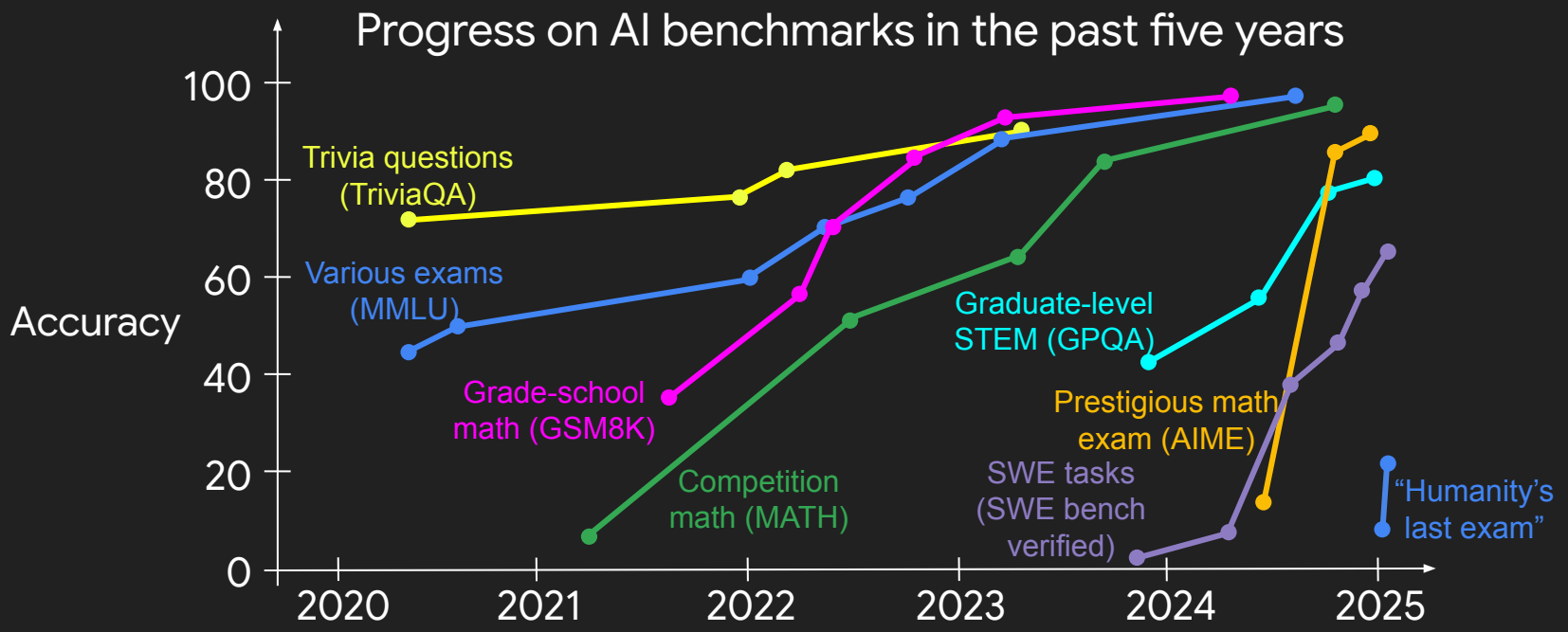
AI:
- Long context window
- Deep world knowledge
- Never tired or distracted



See also "Deep dive into LLMs like ChatGPT" from Andrej Karpathy

How will AI research evolve in the next few years?

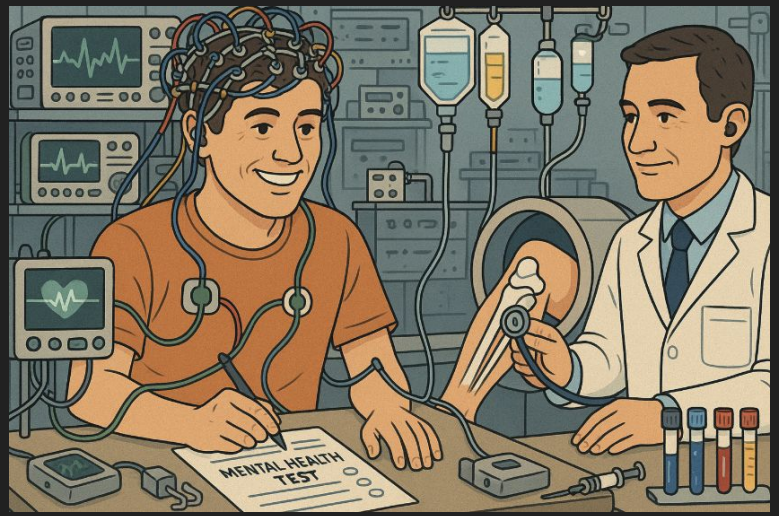
Trend 1: More effort will be spent measuring AI capabilities



“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
- Evals will be require more effort and context

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

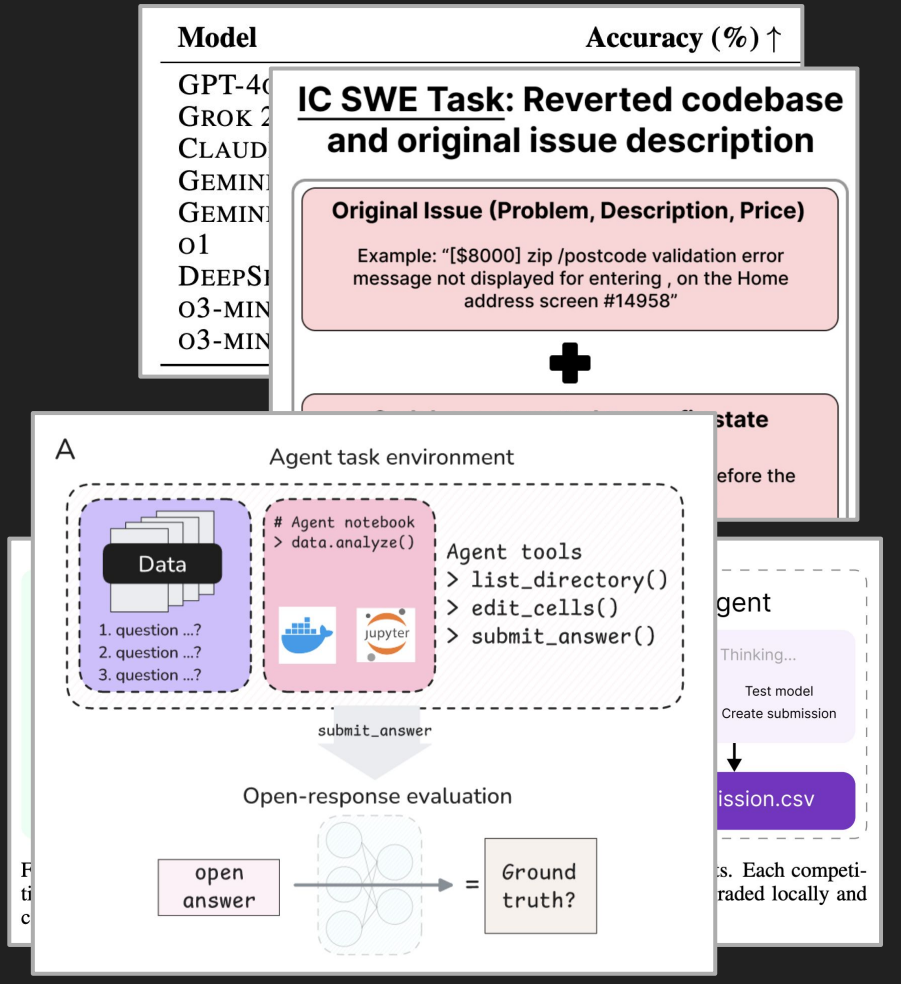


A few benchmarks I'm personally excited about in 2025:

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)



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

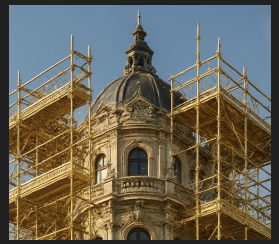
Good enough for RL?	
GPT-1	no
GPT-2	no
GPT-3	no (?)
GPT-4	yes

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

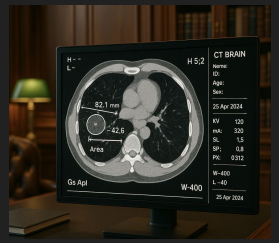
Trend 4: The field will continue to overcome barriers of AI adoption

There is a huge overhang in AI adoption

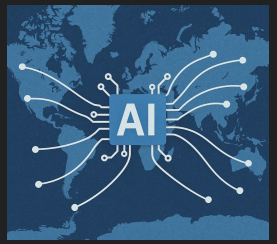


Infrastructure and tooling

New interfaces



Reduced friction and easier to use



Awareness

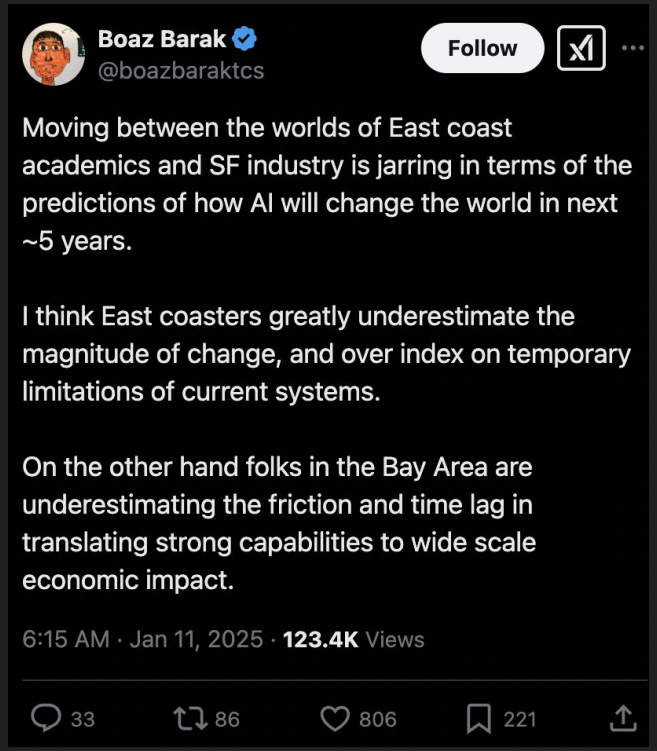
What tasks will AI be able to automate in our world?

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

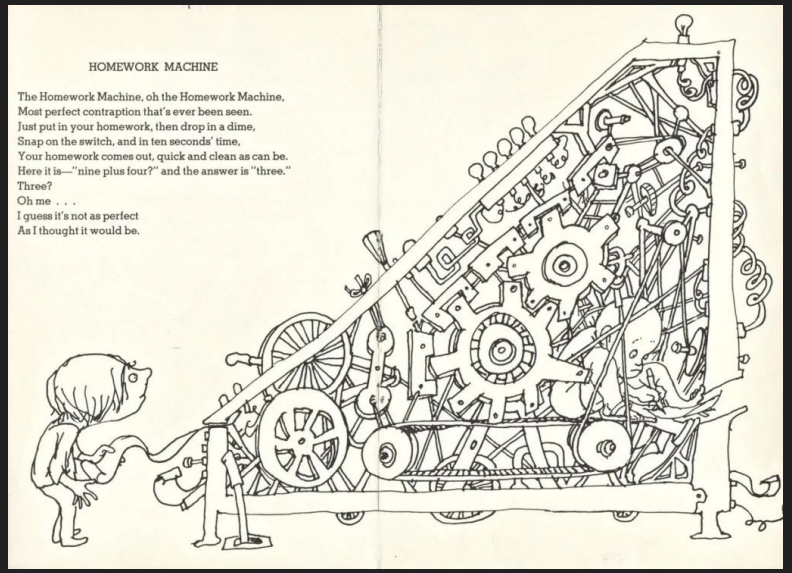
“I think we have 2-3 more years of working as AI researchers before AI takes our own jobs” - AI researcher at another lab

How AI will change the world



AI is good at digital tasks

I, robot (2004) ?



"Homework machine", Shel Silverstein (1981)

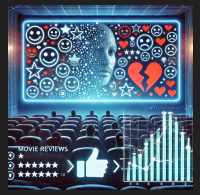
Tasks that are easier for humans tend to be easier for AI (as expected)

easy / fast

hard / time-consuming

?

Identify sentiment of movie review



Competition math problems

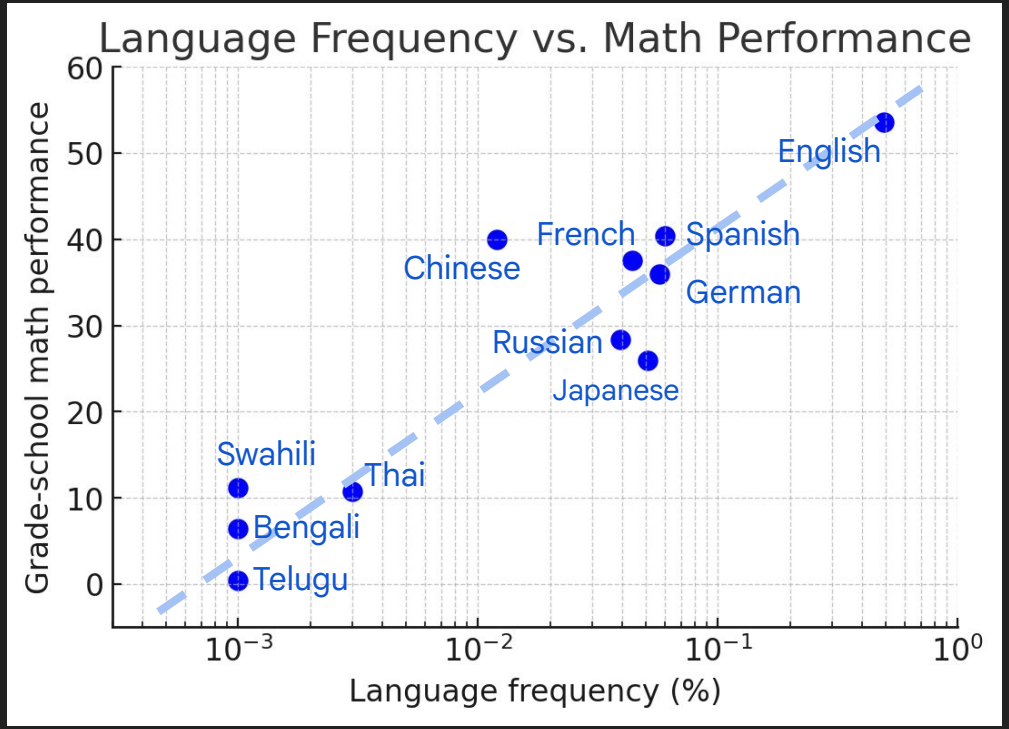
Write a novel

Predict recurrence of breast cancer

Solve open scientific problems

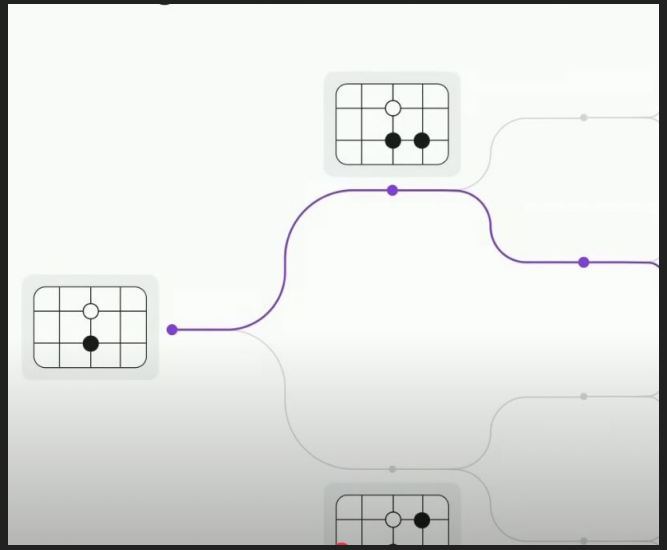
AI thrives when data is abundant

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



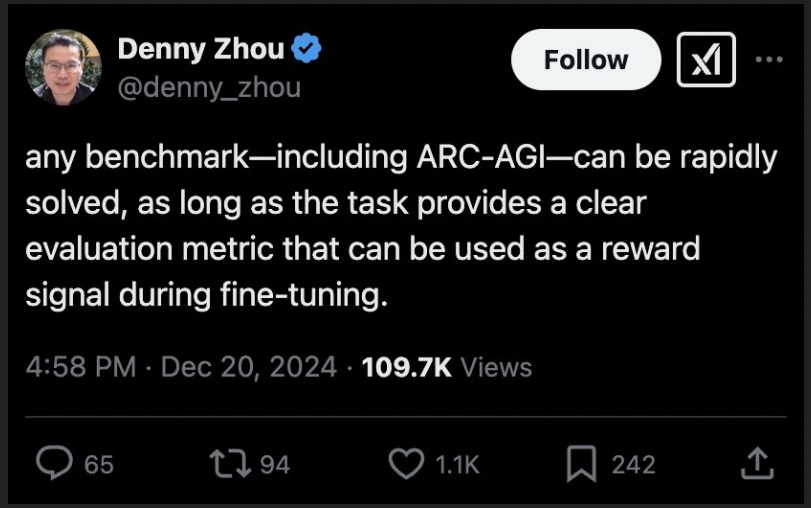
Language models are multilingual chain-of-thought reasoners, 2022.
Language frequency from W3Techs.

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

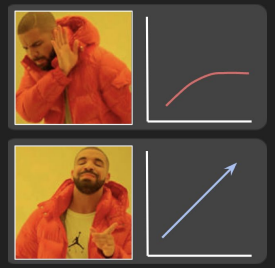


AlphaZero has a single, objective metric for performance.

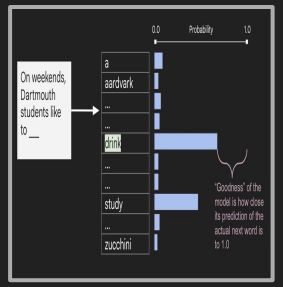
Screenshot from “Nobel Prize lecture: Demis Hassabis, Nobel Prize in Chemistry 2024”



	Difficulty for humans	Digital	Easy to get or create data	Will AI be able to do it?
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

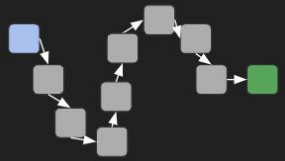


Scaling is an underlying driver of progress

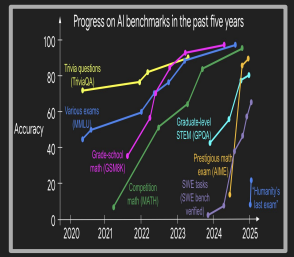


Predicting the next word (pre-training) taught AI about our world

Chain of thought



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



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

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

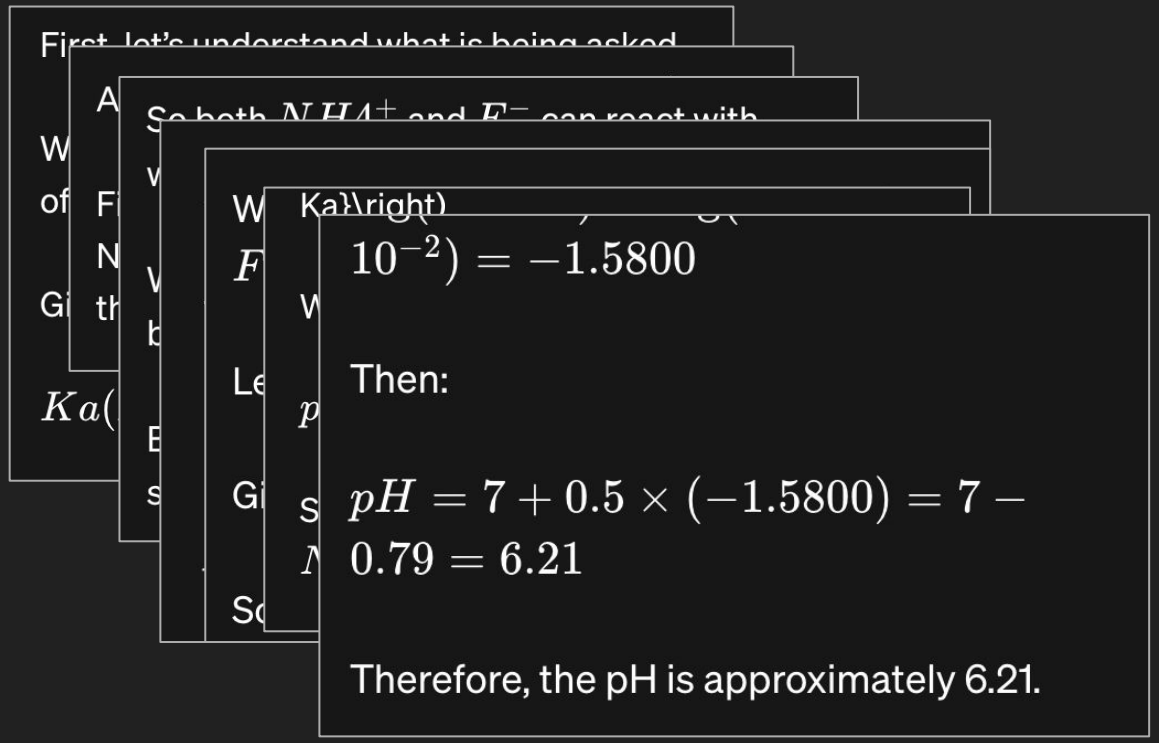
Thanks

X / Twitter: @_jasonwei

Feedback? https://tinyurl.com/jasonwei

Happy to take questions*

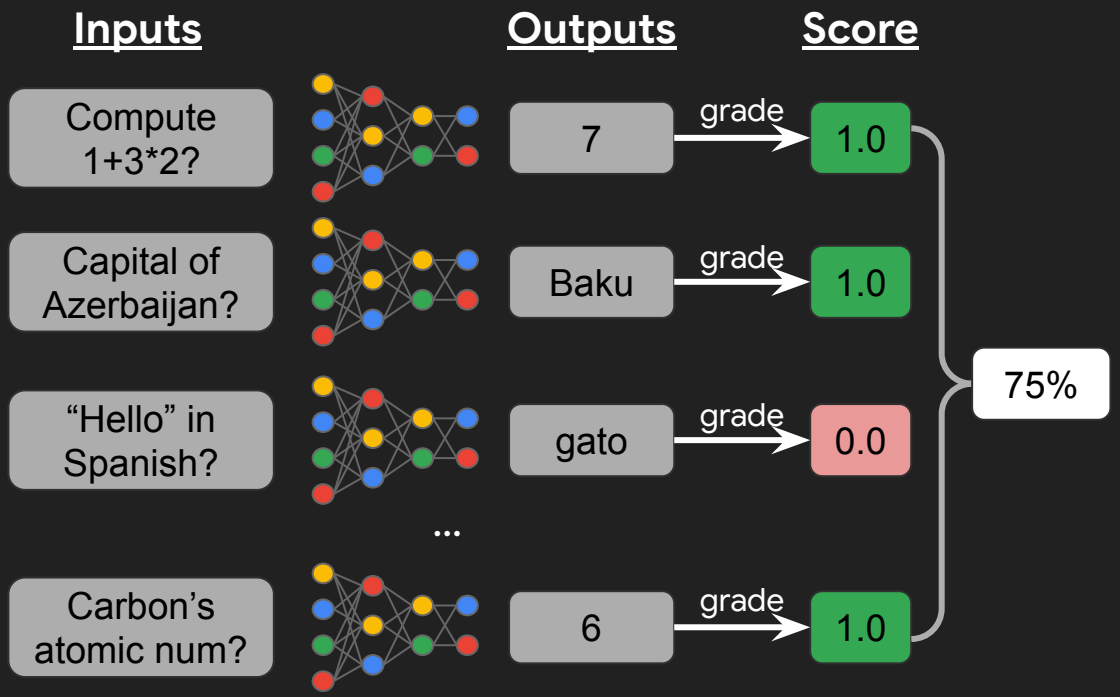
A chain of thought from OpenAI o1



Learning to reason with LLMs. OpenAI, September 2024.

Evals are the measuring sticks in AI

Incentives that set the direction of the AI research community



Scaling was once non-obvious

https

Yann LeCun: Machine learning is great. But the idea that somehow we're going to just scale up the tec
No
lea
kn
n't

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

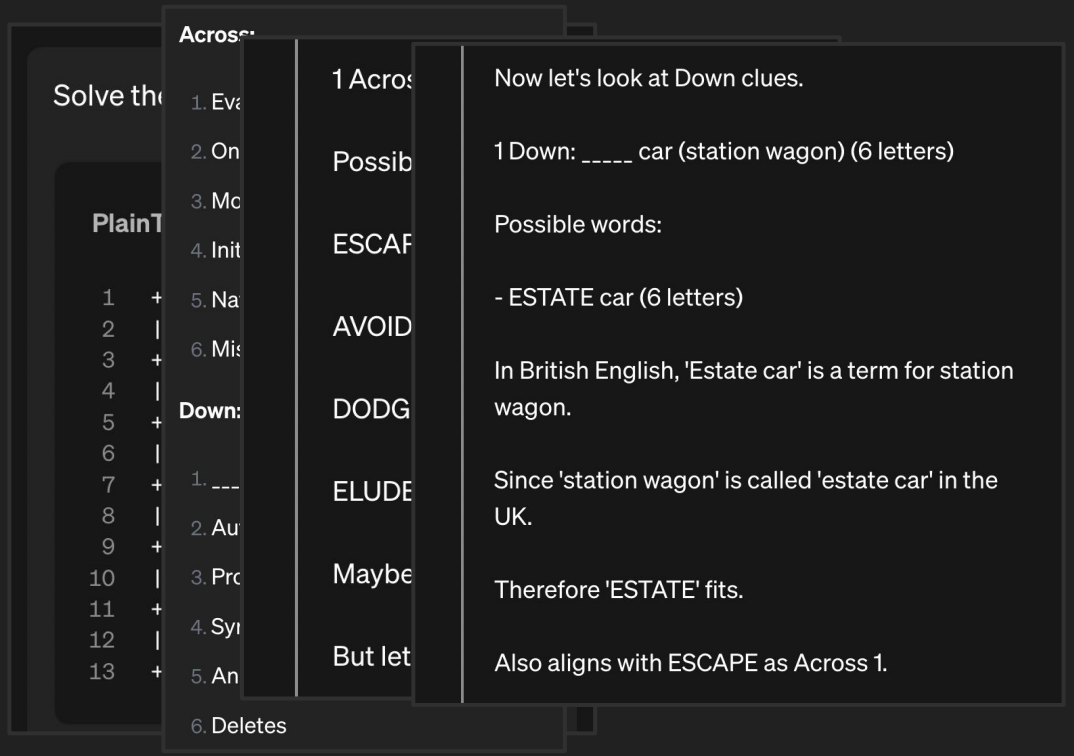
...

Tough right?

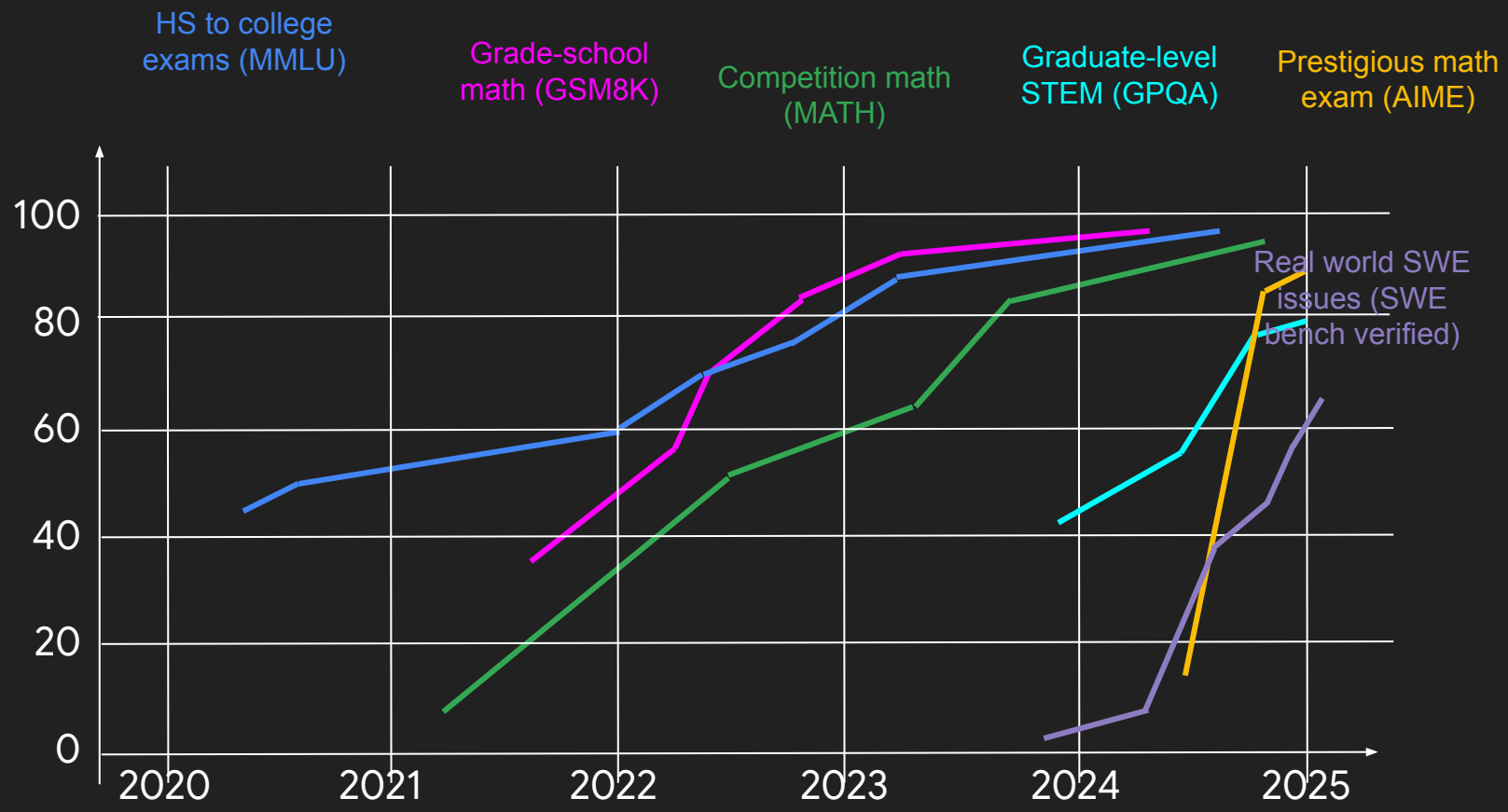
CoT allows models to leverage asymmetry of verification

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

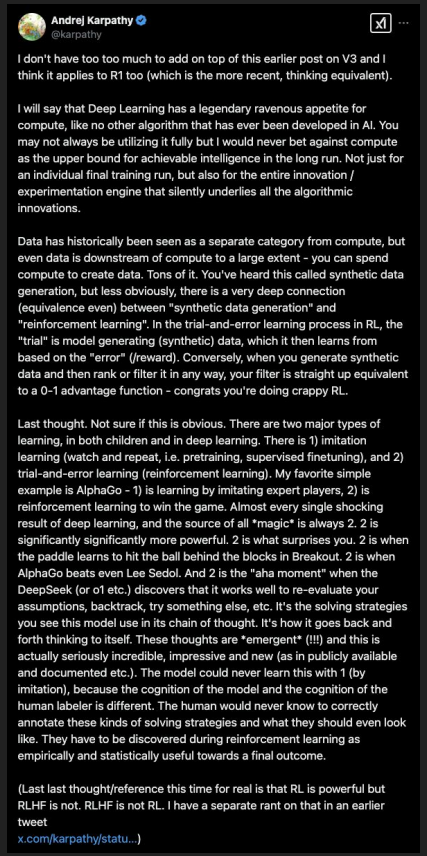
For example, a crossword puzzle, sudoku, or writing a poem that fits constraints



Learning to reason with LLMs. OpenAI, September 2024.


HS to college exams (MMLU)
Grade-school math (GSM8K)
Competition math (MATH)
Graduate-level STEM (GPQA)
Prestigious math exam (AIME)
Real world SWE issues (SWE bench verified)
100
80
60
40
20
0
2020
2021
2022
2023
2024
2025

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

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.

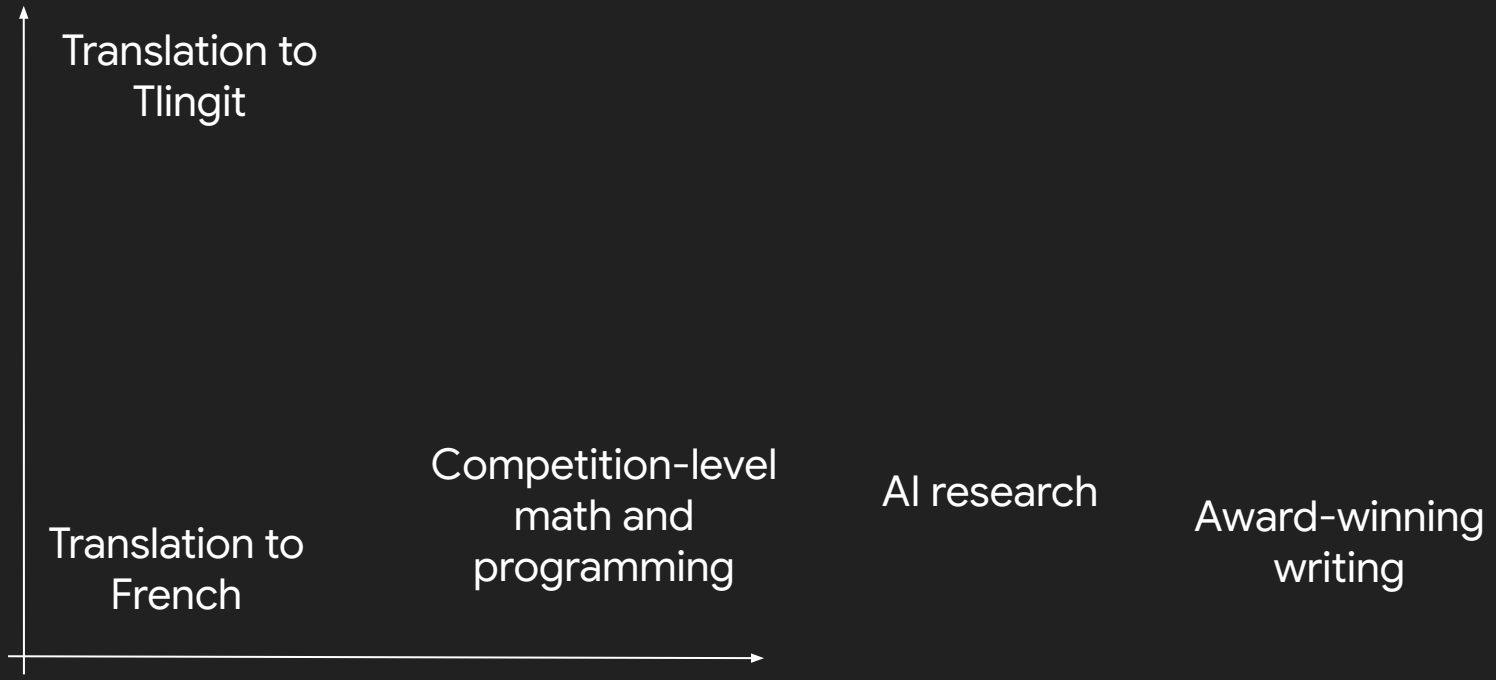
Last thought. Not sure if this is obvious. There are two major types of learning, in both children and in deep learning. There is 1) imitation learning (watch and repeat, i.e. pretraining, supervised finetuning), and 2) trial-and-error learning (reinforcement learning). My favorite simple example is 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 o1 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.

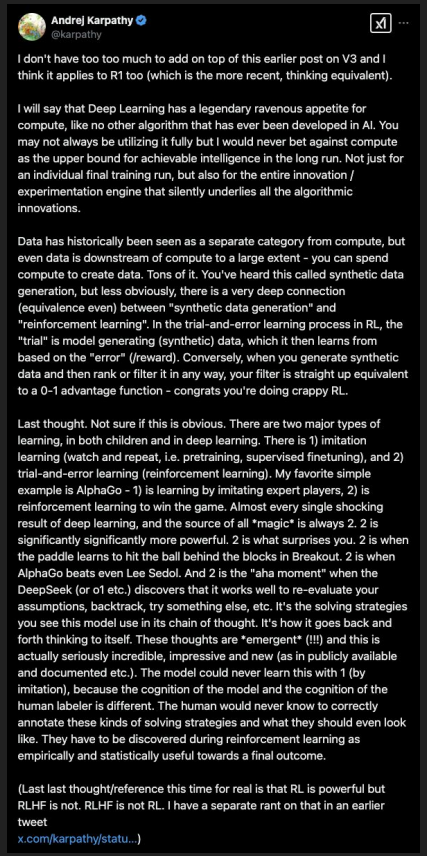
(Last last thought/reference this time for real is that RL is powerful but RLHF is not. RLHF is not RL. I have a separate rant on that in an earlier tweet
x.com/karpathy/statu...)

Why does scaling work?

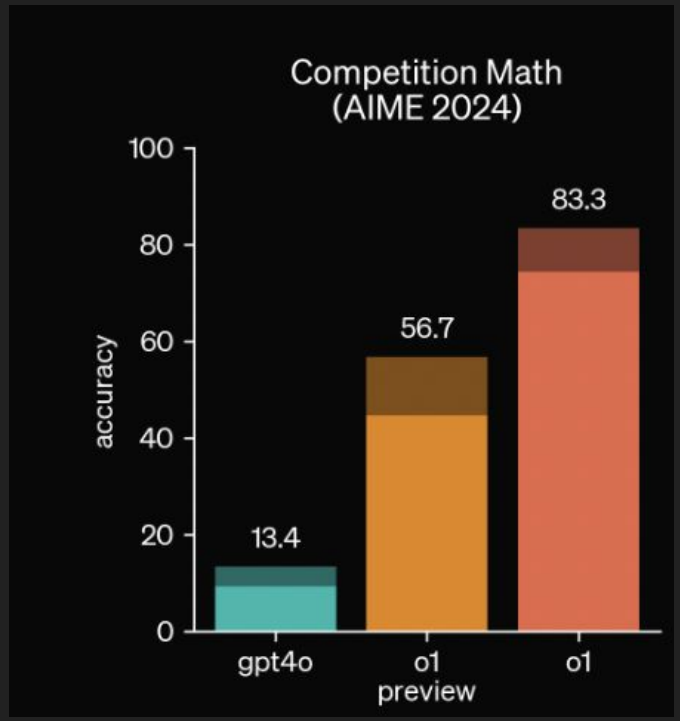
Hard to answer, but here is a hand-wavy explanation

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


Translation to Tlingit
Translation to French
Competition-level math and programming
AI research
Award-winning writing


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 o1 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/statu...)

Scale RL on chain-of-thought



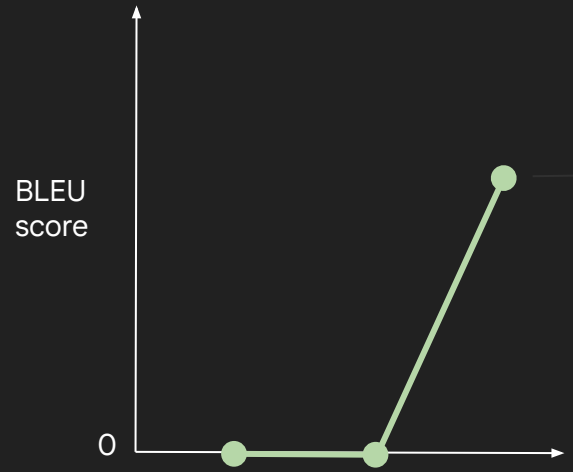
Learning to reason with LLMs. OpenAI, September 2024.

Emergence ability example

Prompt

Input (English): I like to play soccer and tennis

Target (**Spanish**):



curie

Me gusta jugar al fútbol y al tenis ✓

ada

I like to play soccer and tennis ✗

babbage

I like to play soccer and tennis ✗

Model “curie” suddenly figures out to translate and not repeat.

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

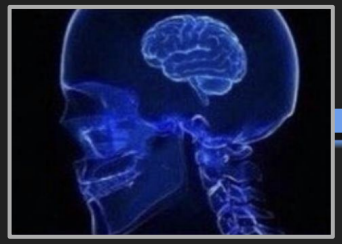
OK I think this is the right answer!

Paradigm 2: Scaling reinforcement learning on chain-of-thought

Train language models to “think” before giving an answer

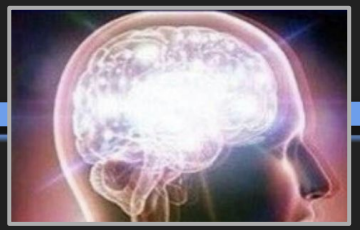
In addition to scaling compute for training, there is a second axis here: scaling how long the language model can think at inference time.

2019



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

2024



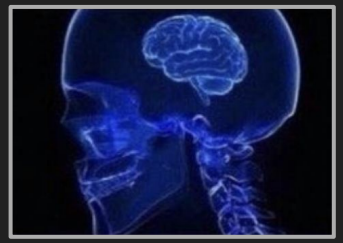
- Can write an essay about almost anything
- Competition-level programmer and mathematician

2029

?

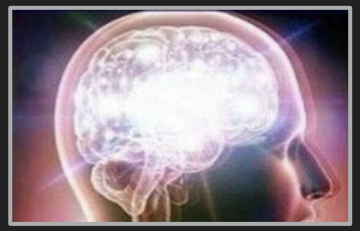
Scaling has been the engine of progress in AI and will continue to dictate how the field advances.

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

Image source: HF

(1) Distributed training requires a lot of expertise

(2) Loss divergences and hardware failures are hurdles

(3) Compute is expensive

Psychological challenges

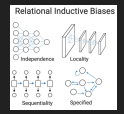
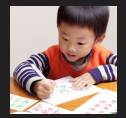


Image source

(1) Researchers like inductive biases

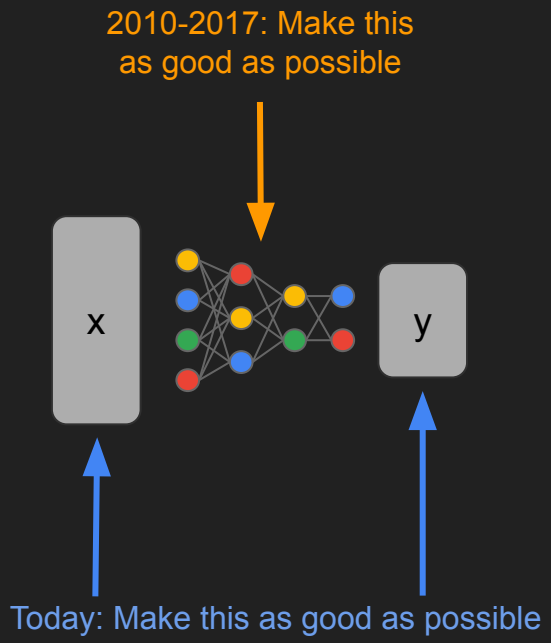


(2) Scaling is different from human learning

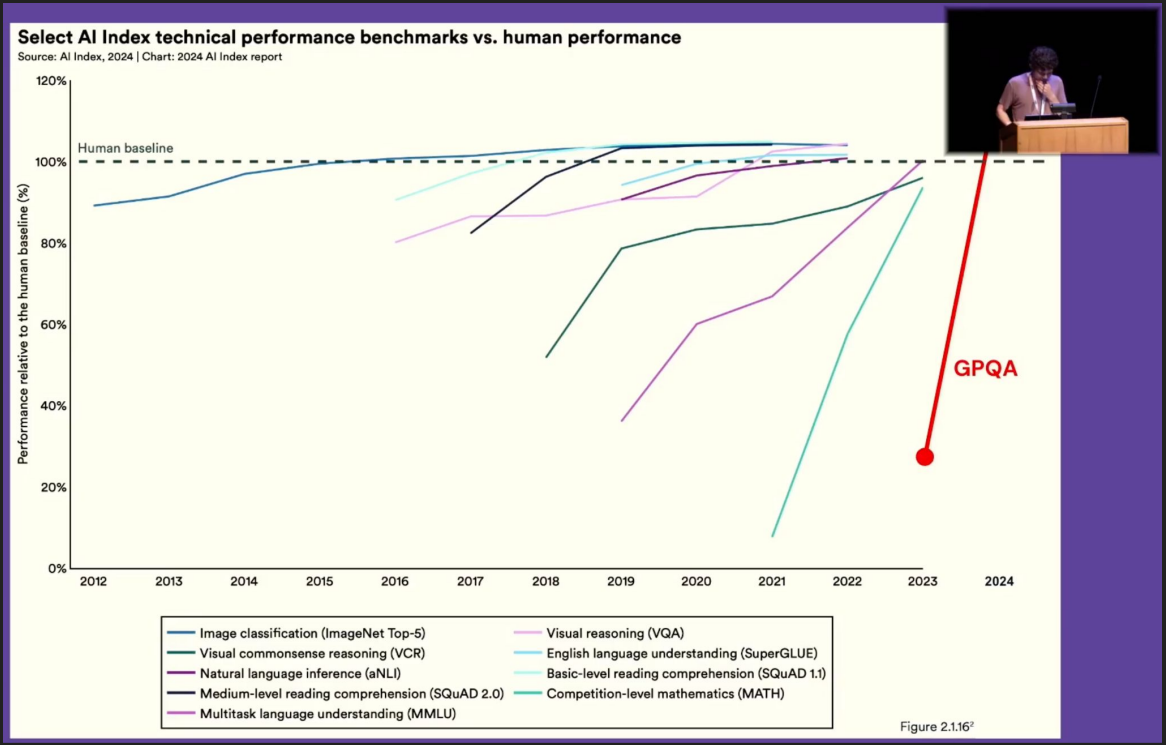
(3) Scientific research incentives don't match engineering work ("novelty")

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

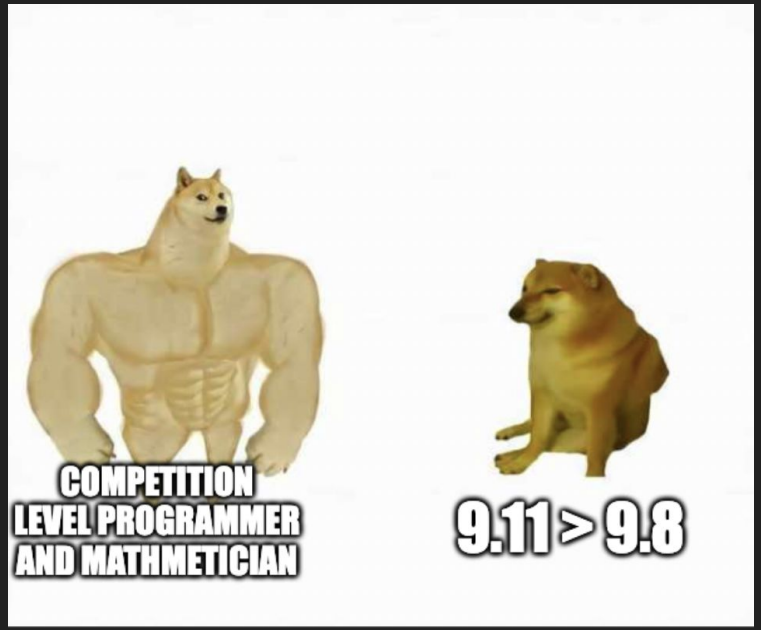


"People ask me if I'm making an even harder version of GPQA... [well] we set out to make the hardest science benchmark that we could"
- David Rein

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

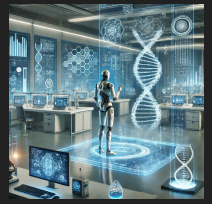
Hard to say that one model is strictly better than another

AI doesn’t need to human-level on everything

Intelligence != user experience

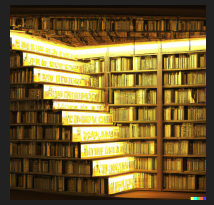
Changes in AI culture: bigger working teams

Where will AI continue to progress?



AI for science and healthcare

As an assistant in scientific and medical innovation

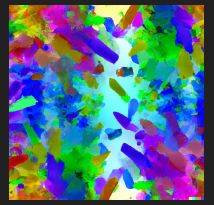


More factual AI

Reduced hallucinations, cite sources, calibration

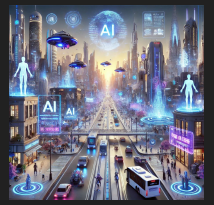
Multimodality

AI to see, hear, and speak



Tool use

Goal: enable AI to interact with the world



AI applications

More ubiquitous use of AI