

Situational Awareness: The Decade Ahead

Leopold Aschenbrenner, June 2024

What do we mean?

“Before long, the world will wake up. But right now, there are perhaps a few hundred people, most of them in San Francisco and the AI labs, that have *situational awareness*.”

Leopold Aschenbrenner

Formerly at OpenAI's Superalignment team

- Focuses on ensuring AI systems much smarter than humans will follow human intent

Fired in April 2024 for allegations of leaking sensitive documents on AI safety

The Chapters

I. From GPT-4 to AGI: Counting the OOMs

II. From AGI to Superintelligence: the Intelligence Explosion

IIIa. Racing to the Trillion-Dollar Cluster

IIIb. Lock Down the Labs: Security for AGI

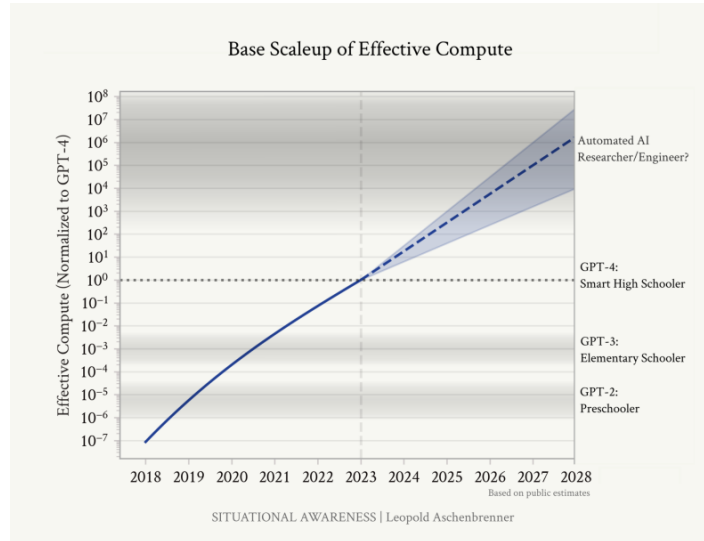
IIIc. Superalignment

IIId. The Free World Must Prevail

IV. The Project

V. Parting Thoughts

“AGI by 2027 is strikingly plausible.”



The last four years

- GPT-2 → GPT-3 → GPT-4
- Progression in capabilities from “preschooler” to “elementary schooler” to “smart high-schooler”
- Caused by three main factors

GPT-2 to GPT-4

GPT-2 examples people found very impressive at the time

<p>Task Reading Comprehension: answer questions about given passages</p> <p>Dataset COQA</p> <p>Examples <i>The trophy doesn't fit into the brown suitcase because it is too large.</i></p> <p>Correct answer: it = trophy Model answer: it = trophy</p> <p><i>The trophy doesn't fit into the brown suitcase because it is too small.</i></p> <p>Correct answer: it = suitcase Model answer: it = suitcase</p>	<p><small>System Prompt (human-written)</small> For today's homework assignment, please describe the reasons for the US Civil War.</p> <p><small>Model Completion (machine-written, 10 tries)</small> By Donny Ferguson</p> <p>It is easy to identify why the Civil War happened, because so many people and so many books and so much television and films tell us that it was the cause, that it has something to do with race or economics or religion. There is much agreement that it was essentially a war of slavery on behalf of capitalism, about a century of slavery. But that's not what most people think of when they think of the Civil War. Many people think of the war as about states rights. You're not wrong about that, but it's kind of misleading to say that the Civil War was a conflict between states' rights and federalism. So let's try again. What's the idea that has inspired so many people to go onto the battlefields of Gettysburg and Antietam and Gettysburg and Petersburg and Fredericksburg? The American idea of the republic--a notion of limited government--is a great part of the history.</p>
--	--

SITUATIONAL AWARENESS | Leopold Aschenbrenner

GPT-2 to GPT-4

GPT-3 examples people found very impressive at the time

A "Gigamuru" is a type of Japanese musical instrument. An example of a sentence that uses the word Gigamuru is:

I have a Gigamuru that my uncle gave me as a gift. I love to play it at home.

To "screeg" something is to swing a sword at it. An example of a sentence that uses the word screeg is:

We screeged at each other for several minutes and then we went outside and ate ice cream.

"The challenge is, you have to win a game of chess against me. If you win, then you may try to take the book. I will get the board." The door unlocks and opens.

> You go inside and play chess against the person.
You go inside and see a beautiful woman with long blond hair. She wears a purple dress and a silver tiara. You have seen this woman before. In fact, you have seen her many times before in pictures and statues.

> You try to remember who she is.
You have a sudden realization of who this person is. This is the lost queen of the kingdom! She was thought to have been killed by bandits, along with her husband, but it looks like she survived.

> You express your surprise and ask her why she let people believe she is dead.
"I am surprised you did not know, with all your wizard knowledge. Let's just say I am hiding from someone. If people thought I was dead, then they wouldn't be looking for me." She says.

> You ask her who she is hiding from.
"If you want to challenge me to a game, then play. If you want to pry into my personal life, then you can leave."
You decide not to press the issue, so you sit down and start a game of chess.


Describe a layout.

Just describe any layout you want, and I'll try to render below!

Generate

```

<button style=(backgroundColor: 'pink', border: '2px solid green', borderRadius: '50%', padding: 20, width: 100, height: 100))=watermelon/button
                    
```



Watermelon

SITUATIONAL AWARENESS | Leopold Aschenbrenner

GPT-2 to GPT-4

GPT-4 examples people found very impressive at the time

GPT-4

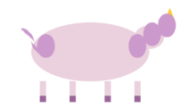
Prompt: Can you write a proof that there are infinitely many primes, with every line that rhymes?

GPT-4:

Yes, I think I can.
 Though it might take a clever plan,
 I'll start by noting Euclid's proof.
 Which shows that primes aren't just a loaf.
 Assume we have a finite list,
 Of primes, and that some have been sized.
 Multiply them all together,
 And add one, just to be clever.
 The result will be a number,
 That has no prime factors, oh wonder!
 But every number has a prime,
 They divide it, it's just a matter of time.
 So we've found a contradiction,
 And our finite list needs revision.
 There must be infinitely many primes,
 And that's the end of my rhyming lines.

Prompt: Draw a unicorn in TikZ.

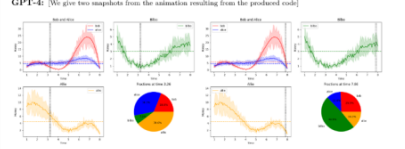
GPT-4: (Produces DTXS compiling to following picture)



Prompt: Can you generate a pyplot for the following data: $x = [1, 2, 3, 5, 6, 8]$, $y1 = [2, 3, 5, 18, 1]$, $y2 = [3, 5, 6, 8, 1]$, $y3 = [5, 1, 2, 3, 6]$, $y4 = [6, 7, 2, 3, 1]$. I want $y1, y2$ to be in the same plot, $y3$ to be another plot next to that plot, $y4$ to be below. I want the legend of $y1$ to be "blue", $y2$ to be "green", $y3$ to be "red", $y4$ to be "purple". I want the x-axis to be labeled with "time" and the y-axis to be labeled with "amount". I want to add 10-20% random error but to match trends through all times (including non-integers). I want smoothed curves to show the plot and smoothed error bar. Do not use linear interpolation, use smooth interpolation! I want to also add some small rigging to the smoothed curve to make it look more real. I want to put a baseline as the mean of each line. I want to put a pie chart below indicating the fraction of the four people in each time step. I also want an animation to show how the fractions are changing in the pie chart as "continuous time". Interpret the missing function. I also want vertical line annotations in other three plots to match the pie chart. I want the fanciest plot. Please add as many fancy things as possible.

Here is a possible code to generate a pyplot for the following data, with some comments to explain the steps:

GPT-4: (We give two snippets from the animation resulting from the produced code)



Prompt: Suppose $g(x) = f^{-1}(x)$, $g(5) = 3$, $g(6) = 7$, $g(3) = 2$, $g(7) = 9$, $g(9) = 6$ what is $f(f(g(9)))$?

GPT-4: There are different ways to approach this problem, but one possible method is to use the fact that $g(x) = f^{-1}(x)$ implies that $f(g(x)) = x$ for any x in the domain of g . That is, applying f to both sides of the equation $g(x) = f^{-1}(x)$ gives $f(g(x)) = f(f^{-1}(x)) = x$. This means that we can undo the effect of f by applying g , and vice versa. Using this idea, we can try to find $f(9)$ by looking for the value of x such that $g(x) = 9$. From the given information, we see that $g(9) = 6$, so $f(6) = g(9) = 9$. Similarly, we can find $f(9)$ by looking for the value of x such that $g(x) = 9$. We see that $g(7) = 9$, so $f(9) = g(7) = 7$. Finally, we can find $f(7)$ by looking for the value of x such that $g(x) = 7$. We see that $g(3) = 7$, so $f(7) = g(3) = 3$. Therefore, $f(f(g(9))) = f(f(6)) = f(9) = 7$.

GPT-2 to GPT-4

GPT-4 vs ChatGPT

Prompt:

A rain barrel collects water off the roof of a house during three hours of heavy rainfall. The height of the water in the barrel increases at the rate of $dt = 4t^2 - 3t$ feet per hour, where t is the time in hours since the rain begins. At time $t = 1$ hour, the height of the water is 0.75 feet. What is the height of the water in the barrel at time $t = 2$ hours?

GPT-4:

If we can find a function F such that $F'(t) = v(t)$, then $F(2) - F(1)$ is the amount that the height of the water increases from $t = 1$ to $t = 2$. To find such a function F , we can use the technique of integration by parts, which says that if u and v are differentiable functions, then

$$\int u dv = uv - \int v du$$

Applying this to $v(t) = 4t^2 - 3t$, we can choose $u = t^2$ and $dv = 4t^2 - 3t dt$, so that $du = 2t dt$ and $v = -\frac{3}{2}t^{-1/2}$. Then we have

$$\int (4t^2 - 3t) dt = \frac{4}{3}t^{3/3} - \frac{3}{2}t^{2/2} = \frac{4}{3}t^3 - \frac{3}{2}t^2$$

Simplifying and applying integration by parts again, we can choose $u = t^2$ and $dv = -\frac{3}{2}t^{-1/2} dt$, so that $du = 2t dt = 2t dt$. Then we have

$$\int (4t^2 - 3t) dt = \frac{4}{3}t^3 - \frac{3}{2}t^2 + \frac{16}{9}t^{3/2} - \frac{16}{9}t^{1/2}$$

ChatGPT:

To find the height of the water in the barrel at time $t = 2$ hours, we need to find the value of $v(2)$. We can do this by plugging 2 in for t in the expression for $v(t)$:

$$v(2) = 4(2)^2 - 3(2) = 32 - 6 = 26$$

To find the height of the water in the barrel at time $t = 2$ hours, we need to add the amount of water collected during the first hour (0.75 feet) to the amount of water collected during the second hour ($v(2)$ feet). So, the total height of the water in the barrel at time $t = 2$ hours is:

$$0.75 + v(2) = 0.75 + 26 = 26.75$$

Therefore, the height of the water in the barrel at time $t = 2$ hours is approximately 0.91 feet.

Figure 4.3: GPT-4 vs ChatGPT on AP problems. GPT-4 uses a correct approach, but produces a wrong final answer due to a calculation error, whereas ChatGPT produces an incoherent argument.

GPT-4

Prompt:

Suppose $g(x) = f^{-1}(x)$, $g(5) = 3$, $g(6) = 7$, $g(3) = 2$, $g(7) = 9$, $g(9) = 6$ what is $f(f(g(9)))$?

GPT-4:

```

def solve():
    # Given information
    g = {}
    g[5] = 3
    g[6] = 7
    g[3] = 2
    g[7] = 9
    g[9] = 6

    # We want to find f(f(g(9)))
    # First, find f(g(9))
    # Since g(9) = 6, we have f(6) = g(9) = 9
    f_g9 = 9

    # Next, find f(9)
    # Since g(7) = 9, we have f(9) = g(7) = 7
    f_f_g9 = 7

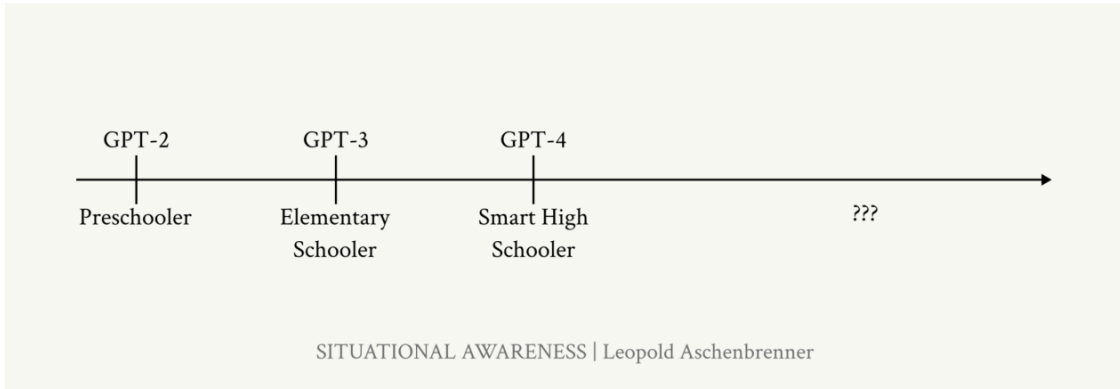
    # Finally, find f(7)
    # Since g(3) = 7, we have f(7) = g(3) = 3
    result = 3

    return result

```

Figure 3.1: Solving a LeetCode problem using dynamic programming. GPT-4's solution also has better readability thanks to comprehensive commenting.

GPT-2 to GPT-4

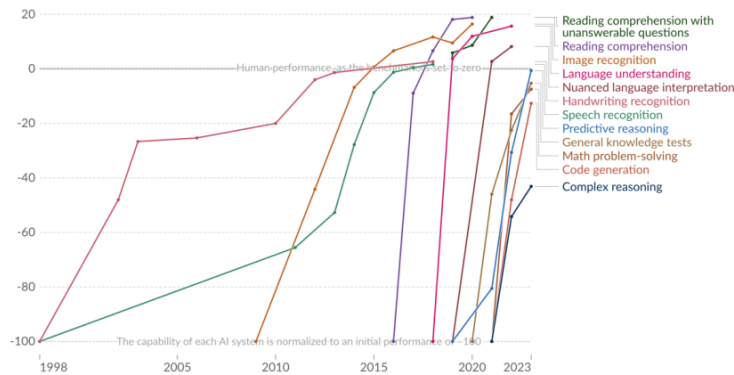


The trends in deep learning

Test scores of AI systems on various capabilities relative to human performance



Within each domain, the initial performance of the AI is set to -100. Human performance is used as a baseline, set to zero. When the AI's performance crosses the zero line, it scored more points than humans.



Data source: Kiela et al. (2023) OurWorldInData.org/artificial-intelligence | CC BY
 Note: For each capability, the first year always shows a baseline of -100, even if better performance was recorded later that year.

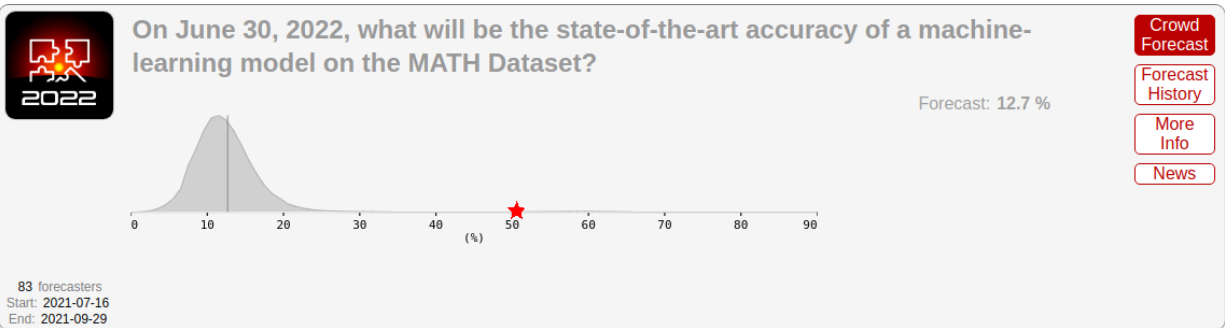
The trends in deep learning

Performance on common exams
(percentile compared to human test-takers)

	GPT-4 (2023)	GPT-3.5 (2022)
Uniform Bar Exam	90th	10th
LSAT	88th	40th
SAT	97th	87th
GRE (Verbal)	99th	63rd
GRE (Quantitative)	80th	25th
US Biology Olympiad	99th	32nd
AP Calculus BC	51st	3rd
AP Chemistry	80th	34th
AP Macroeconomics	92nd	40th
AP Statistics	92nd	51st

SITUATIONAL AWARENESS | Leopold Aschenbrenner

The trends in deep learning



The trends in deep learning

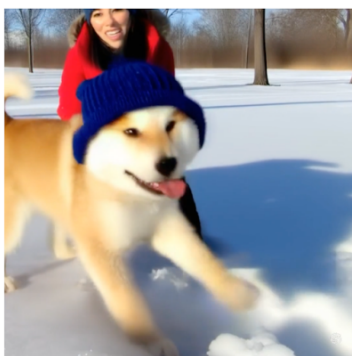
Chemistry (general)
A reaction of a liquid organic compound, which molecules consist of carbon and hydrogen atoms, is performed at 80 centigrade and 20 bar for 24 hours. In the proton nuclear magnetic resonance spectrum, the signals with the highest chemical shift of the reactant are replaced by a signal of the product that is observed about three to four units downfield. Compounds from which position in the periodic system of the elements, which are also used in the corresponding large-scale industrial process, have been mostly likely initially added in small amounts? A) A metal compound from the fifth period. B) A metal compound from the fifth period and a non-metal compound from the third period. C) A metal compound from the fourth period. D) A metal compound from the fourth period and a non-metal compound from the second period.
Organic Chemistry
Methylcyclopentadiene (which exists as a fluxional mixture of isomers) was allowed to react with methyl isocyanide ketone and a catalytic amount of pyridine. A bright yellow, cross-conjugated polyketenyl hydrocarbon product formed (as a mixture of isomers), with water as a side product. These products are derivatives of fulvene. This product was then allowed to react with ethyl acrylate in a 1:1 ratio. Upon completion of the reaction, the bright yellow color had disappeared. How many chemically distinct isomers make up the final product (not counting stereoisomers)? A) 2 B) 16 C) 8 D) 4
Genetics
If a sperm from species A is injected into an egg from species B and both species have the same number of chromosomes, what would be the main cause of the resulting zygote mortality? A) Species specific zona pellucida proteins on the egg cannot bind sperm from a different species. B) Epistatic interactions between the genes of different species. C) Chromosomal incompatibilities will cause failure of mitosis leading to death of zygote. D) Chromosomal recombination will not occur in different species.
Molecular Biology
A scientist studies the stress response of barley to increased temperatures and finds a protein which contributes to heat tolerance through the stabilization of cell membrane. The scientist is very happy and wants to create a heat tolerant cultivar of diploid wheat. Using databases, they find a heat tolerance protein homologue and start analysing its accumulation under heat stress. Soon enough, the scientist discovers this protein is not synthesised in the wheat cultivar they study. There are many possible reasons for such behaviour, including: A) A mRNA targets the protein, which makes exonucleases cut it immediately after the end of translation and before processing in ER B) Trimethylation of lysine of H3 histone in position 27 at the promoter of the gene encoding the target protein C) A stop-codon occurs in the 5'-UTR region of the gene encoding the target protein D) The proteolysis process disrupts a quaternary structure of the protein, preserving only a tertiary structure
Astrophysics
Astronomers are studying a star with a Teff of approximately 6000 K. They are interested in spectroscopically determining the surface gravity of the star using spectral lines (EW < 100 mÅ) of two chemical elements, E1 and E2. Given the atmospheric temperature of the star, E1 is mostly in the neutral phase, while E2 is mostly ionized. Which lines are the most sensitive to surface gravity for the astronomers to consider? A) E1 I (neutral) B) E1 II (singly ionized) C) E2 II (singly ionized) D) E1 I (neutral)
Quantum Mechanics
Suppose we have a depolarizing channel operation given by $E(\rho)$. The probability, p , of the depolarization state represents the strength of the noise. If the Kraus operators of the given state are $A_0 = \sqrt{1-p}I$, $A_1 = \sqrt{p}X$, $A_2 = \sqrt{p}Y$, and $A_3 = \sqrt{p}Z$. What could be the correct Kraus Representation of the state $E(\rho)$? A) $E(\rho) = (1-p)\rho + \frac{p}{2}X\rho X + \frac{p}{2}Y\rho Y + \frac{p}{2}Z\rho Z$ B) $E(\rho) = (1-p)\rho + \frac{p}{2}X\rho X + \frac{p}{2}Y\rho Y + \frac{p}{2}Z\rho Z$ C) $E(\rho) = (1-p)\rho + \frac{p}{2}X\rho X + \frac{p}{2}Y\rho Y + \frac{p}{2}Z\rho Z$ D) $E(\rho) = (1-p)\rho + \frac{p}{2}X\rho X + \frac{p}{2}Y\rho Y + \frac{p}{2}Z\rho Z$

Table 1: Six example questions from the dataset, two each from subdomains of chemistry, biology, and physics (respectively).

Counting the orders of magnitude (OOMs)



Base compute



4x compute



32x compute

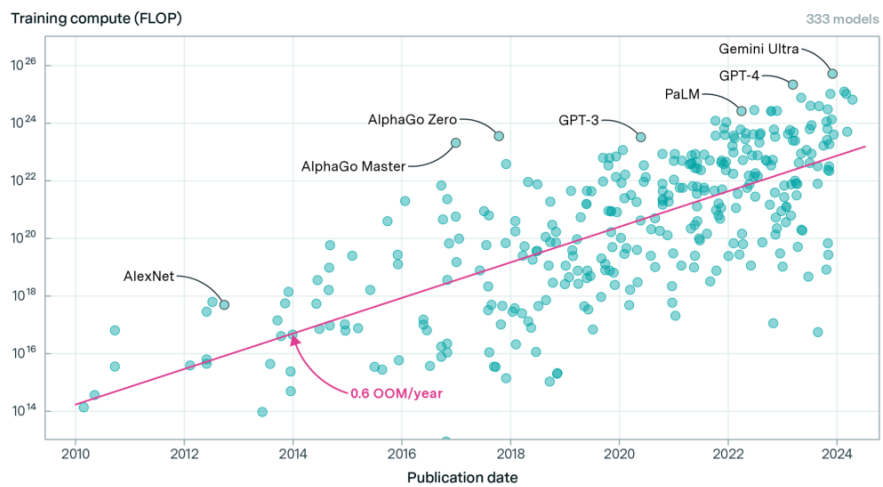
Compute

Model	Estimated compute	Growth
GPT-2 (2019)	$\sim 4e21$ FLOP	
GPT-3 (2020)	$\sim 3e23$ FLOP	+ ~ 2 OOMs
GPT-4 (2023)	$8e24$ to $4e25$ FLOP	+ $\sim 1.5-2$ OOMs

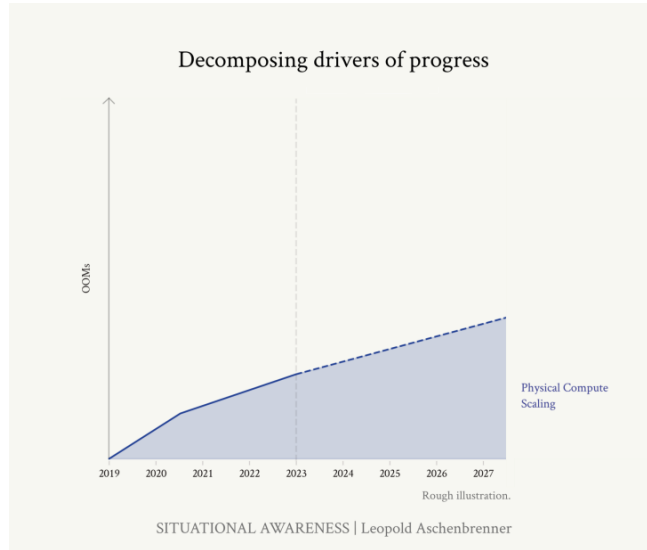
Compute

Training compute of notable models

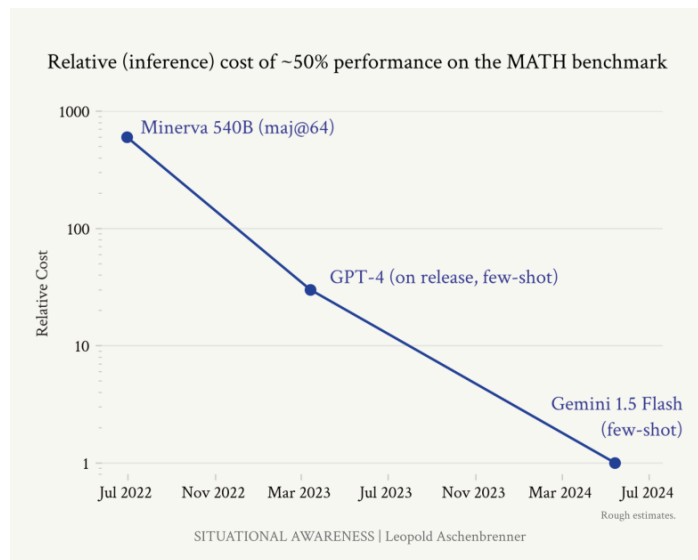
EPOCH AI



Compute



Algorithmic efficiencies



Algorithmic efficiencies

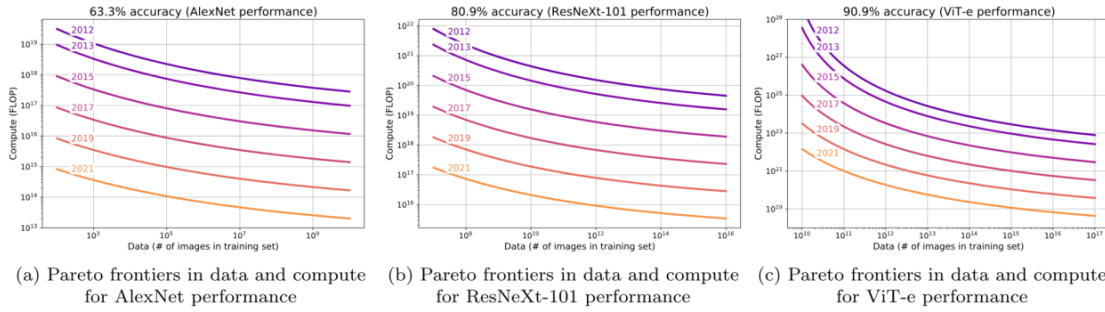
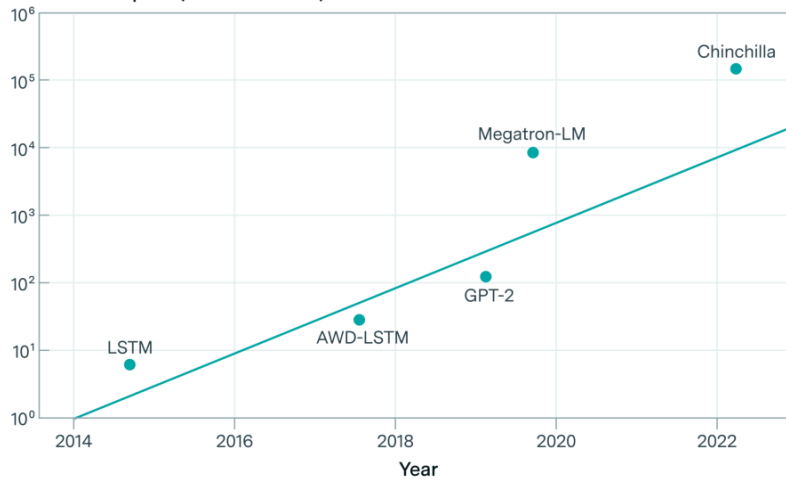


Figure 1. Pareto frontiers for training models to achieve performance of well-known models over time.

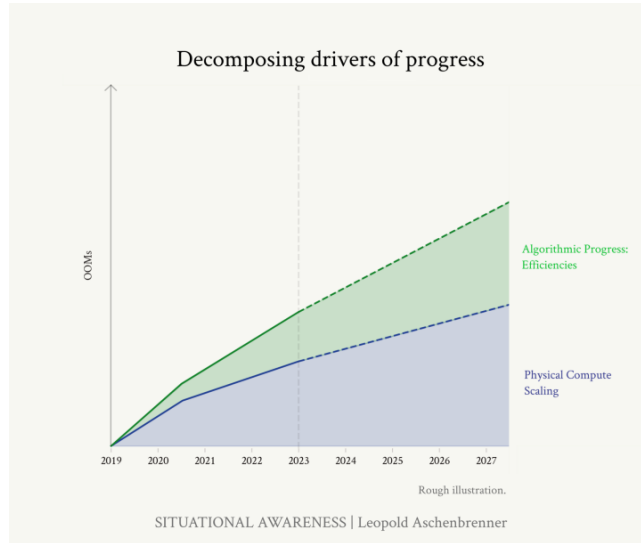
Algorithmic efficiencies

Efficiency doubles roughly every 8 months

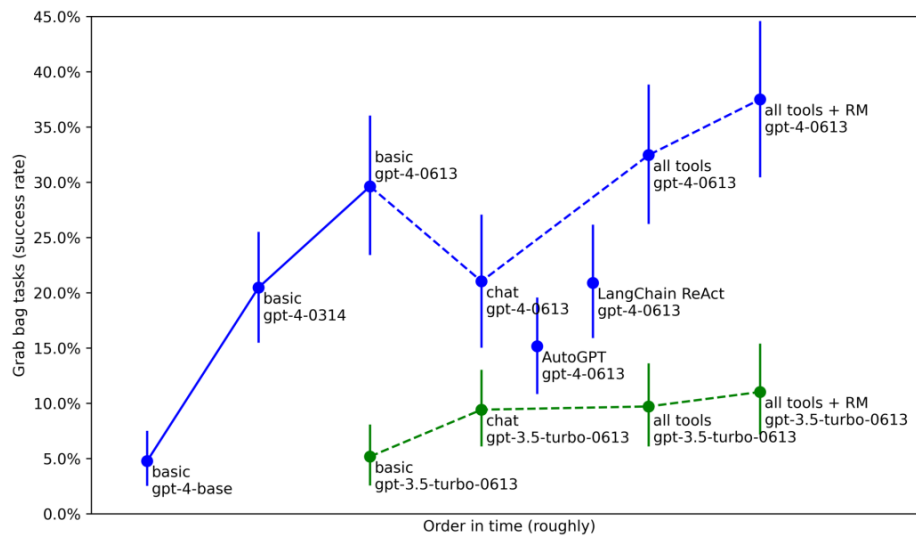
Effective compute (relative to 2014)



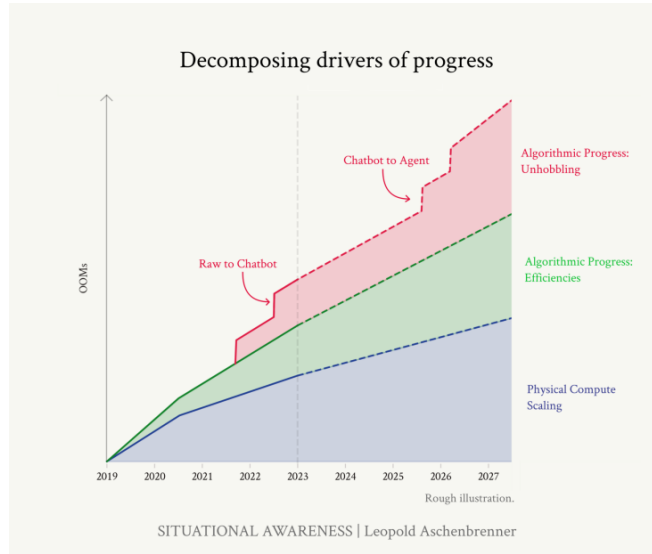
Algorithmic efficiencies



“Unhobbling”



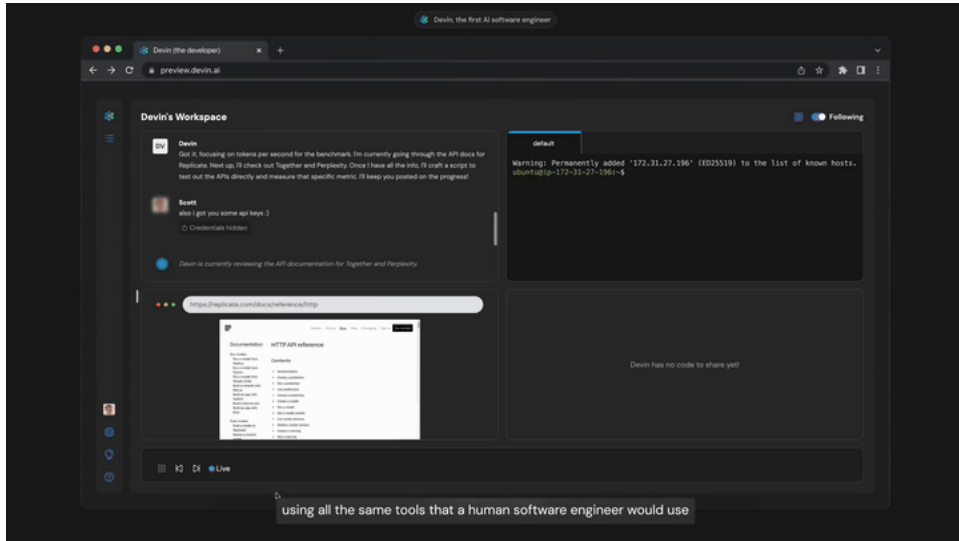
“Unhobbling”



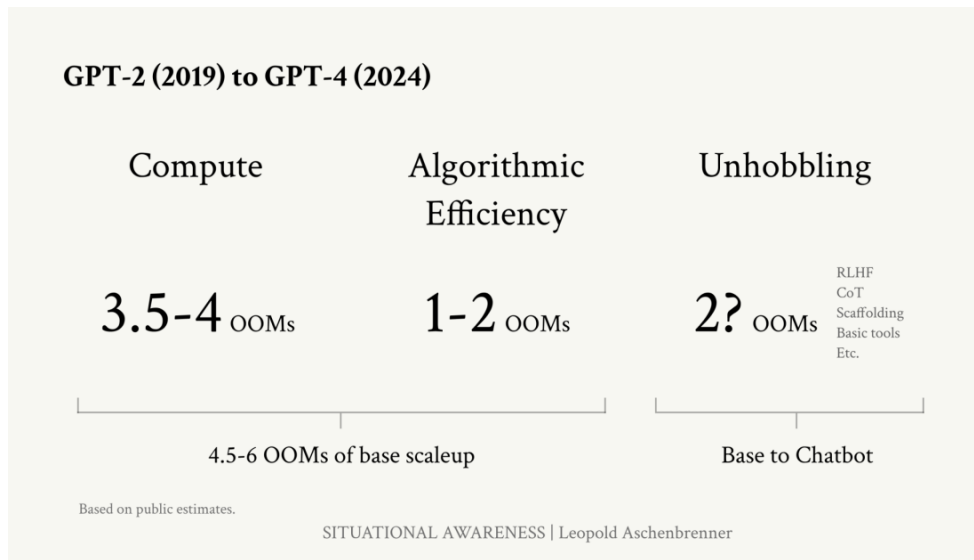
From chatbot to agent-coworker

Number of tokens	Equivalent to me working on something for...	
100s	A few minutes	ChatGPT (we are here)
1000s	Half an hour	+1 OOMs test-time compute
10,000s	Half a workday	+2 OOMs
100,000s	A workweek	+3 OOMs
Millions	Multiple months	+4 OOMs

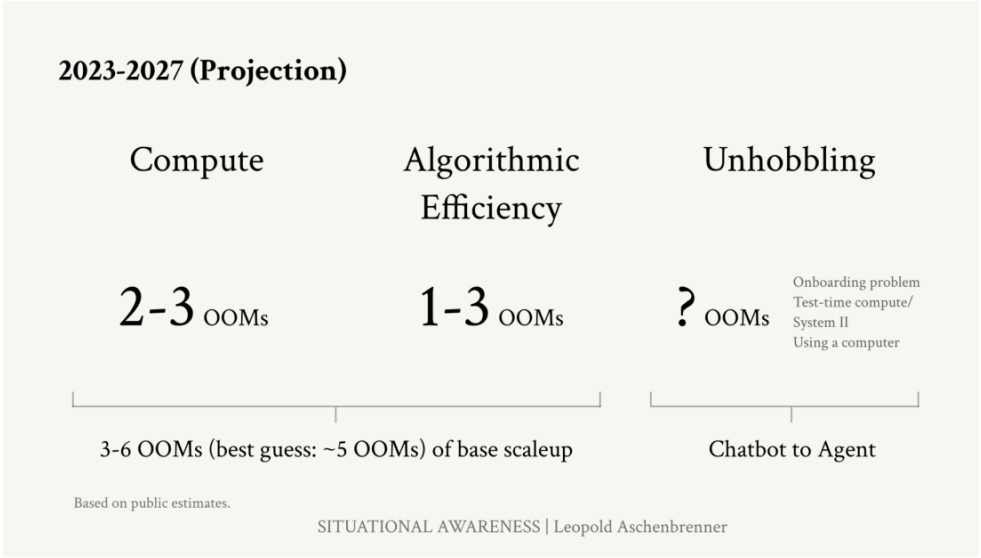
From chatbot to agent-coworker



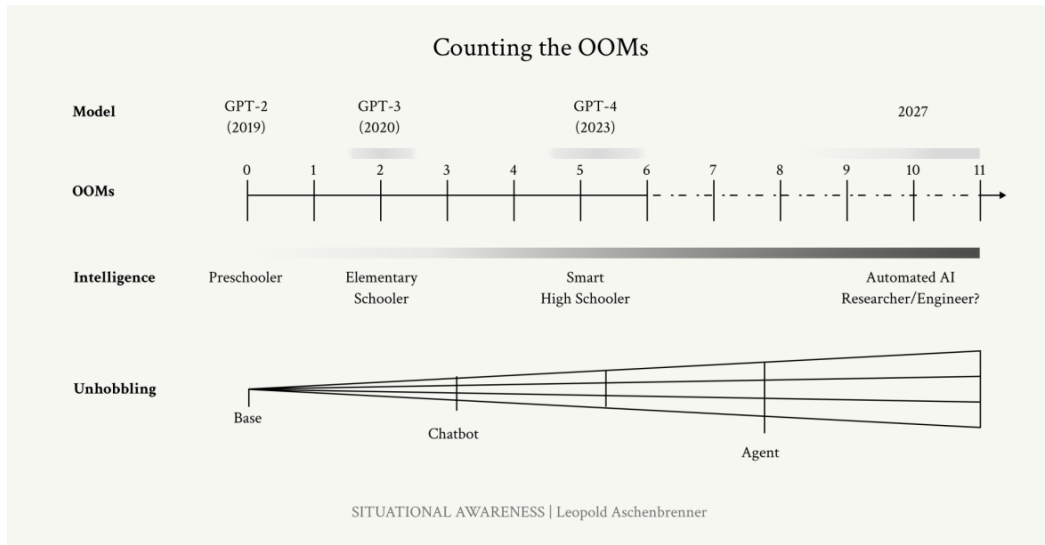
The next four years



The next four years



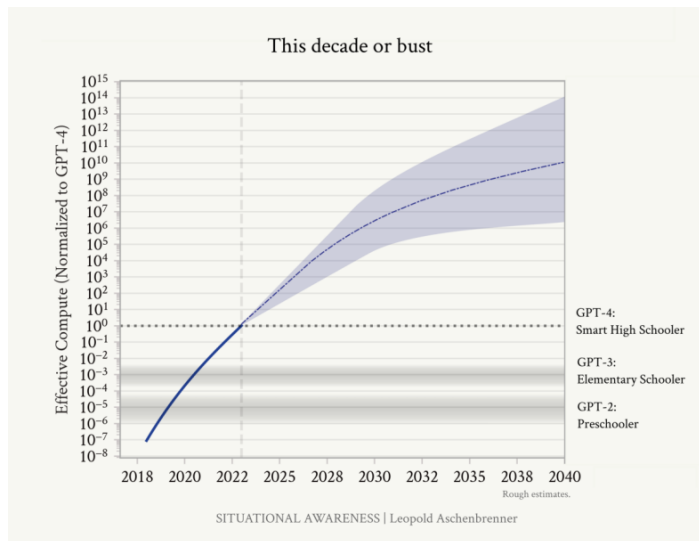
The next four years



The next four years



Racing through the OOMs: It's this decade or bust



Racing through the OOMs: It's this decade or bust



Matthew Barnett
@MatthewJBar

...

My own basic calculations suggest that, given the potential for increased investment and hardware progress, we could very soon move through a large fraction of the remaining compute gap between the current frontier models and the literal amount of computation used by evolution.



7:54 PM · Mar 26, 2024 · 3,968 Views

Contributions, rebuttal, and discussion