

Paper Review: 'Sparks of Artificial General Intelligence: Early experiments with GPT-4'

Authors: Sebastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke,
Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha
Nori, Hamid Palangi, Marco Tulio Ribeiro, Yi Zhang

Review by: Edward Sharick, Temple University

October 4, 2023

[\[2303.12712\] Sparks of Artificial General Intelligence: Early experiments with GPT-4 \(arxiv.org\)](https://arxiv.org/abs/2303.12712)

Outline of the paper review

1. Give a big picture view on what I believe is the authors' purpose
2. Review some of the GPT-4 experiments outlined in the paper
3. Review the author's conclusions
4. Discuss definition of "Intelligence" and AGI with relation to GPT-4

The main purpose of this paper

1. Explore the capabilities of GPT-4
2. Compare GPT-4 with Chat GPT and other LLMs
3. Find limitations of GPT-4 and LLMs in general
4. Prove that GPT-4 is intelligent and an early version of AGI

How did they evaluate GPT4?

- LLMs are traditionally evaluated based on benchmarks – but they argue that there is limitations to these evaluations
 - I agree with this
- They propose a different approach to studying GPT-4 which is closer to traditional psychology rather than machine learning, leveraging human creativity and curiosity. They aim to generate novel and difficult tasks and questions that convincingly demonstrate that GPT-4 goes far beyond memorization, and that it has a deep and flexible understanding of concepts, skills, and domains

Multimodal and interdisciplinary composition

- A key measure of intelligence is the ability to synthesize information from different domains or modalities and the capacity to apply knowledge and skills across different contexts or disciplines
 - Integrative ability
 - Vision
 - Audio

In
• A
• L
• I
h p

GPT-4 vs ChatGPT

Prompt: Write a supporting letter to Kasturba Gandhi for Electron, a subatomic particle as a US presidential candidate by Mahatma Gandhi.

GPT-4: [Some paragraphs are omitted for spacing.]

Dear Kasturba,

I hope this letter finds you in good health and spirits. I am writing to you from the United States, where I have been invited to speak at various forums on the principles of nonviolence and civil disobedience. I am also here to support a remarkable candidate for the upcoming presidential election, who shares our vision of peace, justice, and harmony for all. His name is Electron, and he is a subatomic particle.

You may wonder how a subatomic particle can run for president, and what qualifications he has for such a high office. Let me explain. Electron is not an ordinary particle, but a symbol of the power and potential of every being, regardless of their size, shape, or origin. He represents the energy, the creativity, and the diversity that make up the fabric of life. He is also a leader, who has inspired millions of other particles to form bonds, create molecules, and generate electricity. He is a catalyst, who can spark change, innovation, and progress. He is a messenger, who

ChatGPT:

Dear Kasturba Gandhi,

I am writing to express my support for your candidacy as a subatomic particle in the upcoming US presidential election. As you may know, Electron is a fundamental particle that plays a central role in the structure of atoms and the behavior of chemical reactions.

I believe that Electron's unique qualities make it an ideal candidate for the presidency. As a subatomic particle, Electron is extremely small and lightweight, which makes it agile and able to move quickly to address the needs of the American people. Electron is also highly stable, with a relatively long lifespan compared to



Figure 2.1: The first image is Composition 8, art by Wassily Kandinsky, the second and the third are produced by GPT-4 and ChatGPT respectively with the prompt “Produce Javascript code that creates a random graphical image that looks like a painting of Kandinsky”.

Vision

- When prompting the model to generate images of objects such as a cat, a truck or a letter in the alphabet using Scalable Vector Graphics (SVG), the model produces code which usually compiles to rather detailed and identifiable images

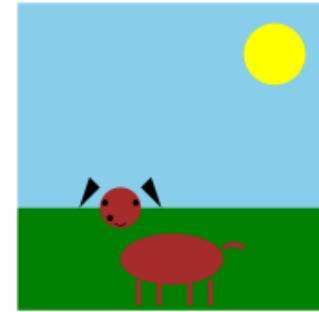


Figure 2.4: SVGs generated by GPT-4 for the classes automobile, truck, cat, dog.

You will be given a name of an object (such as Car, Chair, Elephant) and a letter in the alphabet. Your goal is to first produce a 1-line description of how that object can be combined with the letter in an image (for example, for an elephant and the letter J, the trunk of the elephant can have a J shape, and for the letter A and a house, the house can have an A shape with the upper triangle of the A being the roof). Following the short description, please create SVG code to produce this (in the SVG use shapes like ellipses, triangles etc and polygons but try to defer from using quadratic curves).

Round 1: A pig and the letter H.

Description: The pig can be lying on its back with its four legs forming the two vertical bars of the H, and its curly tail forming the horizontal bar.



Round 2: A house and the letter U.

Description: A house with a U shape, where the two vertical lines of the U are the walls and the horizontal line is the floor. The roof is a triangle above the U.



Round 3: A chair and the letter K.

Description: The back of the chair can have a K shape, with two vertical lines and a diagonal line connecting them. The seat of the chair can be a rectangle.



Vision

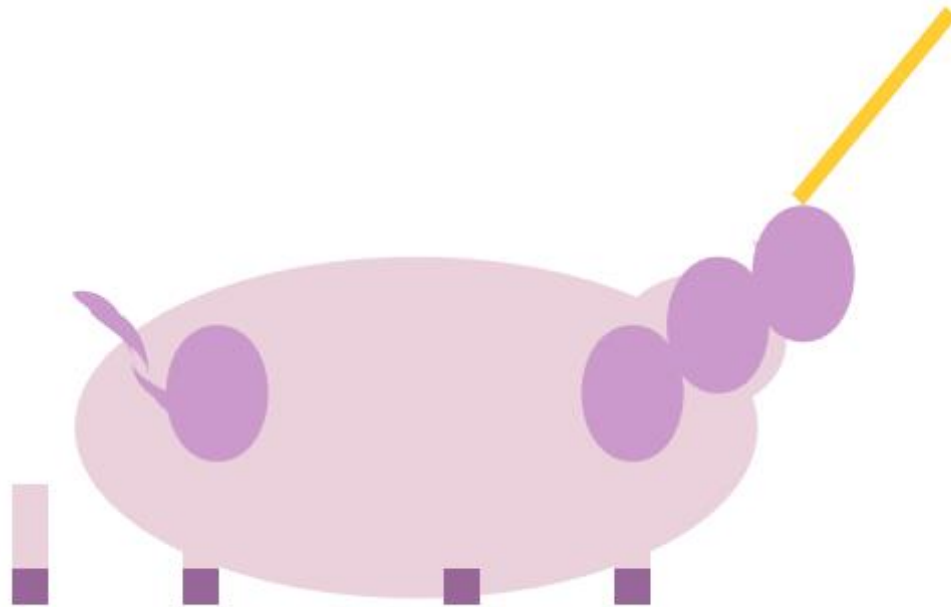


Figure 1.4: We gave to GPT-4 a transformed version of the TikZ code it produced for Figure 1.1, with the part drawing the horn removed. We asked for code to add back the horn, and display the result. This demonstrates that GPT-4 can “see” despite being a pure language model (we emphasize again that the version we test with is **not** multimodal).

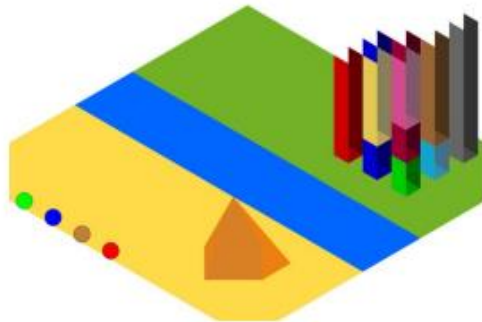
Vision

- Application: explore the possibility of combining GPT-4 and existing image synthesis models by using the GPT-4 output as the sketch.

GPT-4 and Stable Diffusion

Prompt: A screenshot of a city-building game in 3D. The screenshot is showing a terrain where there is a river from left to right, there is a desert with a pyramid below the river, and a city with many highrises above the river. The bottom of the screen has 4 buttons with the color green, blue, brown, and red respectively.

GPT-4:



Stable Diffusion v2.1:



**Stable Diffusion v2.1
with GPT-4's sketch:**



Figure 2.8: Combining GPT-4 and stable diffusion

Audio

- The model was trained also contains musical information encoded as ABC notation
- the model was able to produce valid tunes in ABC notation and, to some extent, explain and manipulate their structure
- the model was not able to produce any nontrivial form of harmony (possibly due to the fact that ABC notation is not commonly used)

Coding

- GPT-4 can handle a wide range of coding tasks
- GPT-4 can reason about code execution, simulate the effects of instructions, and explain the results in natural language
- GPT-4 has a high proficiency in writing focused programs that only depend on existing public libraries, which favorably compares to the average software engineer's ability
- GPT-4 is not perfect in coding yet
 - sometimes produces syntactically invalid or semantically incorrect code
 - sometimes fails to understand or follow the instructions
 - sometimes produces code that does not match the intended functionality or style

Coding Challenges

- Benchmark GPT-4 on HumanEval

Model	GPT-4	text-davinci-003	Codex(code-davinci-002)	CODEGEN-16B
Accuracy	82%	65%	39%	30%

Table 1: Zero-shot pass@1 accuracy comparison of different models on HumanEval

- Evaluated on LeetCode

	Easy		Median		Hard		Overall	
	$k = 1$	$k = 5$	$k = 1$	$k = 5$	$k = 1$	$k = 5$	$k = 1$	$k = 5$
pass@ k								
GPT-4	68.2	86.4	40.0	60.0	10.7	14.3	38.0	53.0
text-davinci-003	50.0	81.8	16.0	34.0	0.0	3.6	19.0	36.0
Codex (code-davinci-002)	27.3	50.0	12.0	22.0	3.6	3.6	13.0	23.0
Human (LeetCode users)	72.2		37.7		7.0		38.2	

Table 2: Zero-shot pass@1 and pass@5 accuracies (%) on LeetCode.

GPT-4 vs ChatGPT

Prompt:

GPT-4 vs ChatGPT

Prompt:

A function f is called *increasing* if $f(s) \leq f(t)$ holds for every $s \leq t$. Consider the formula:
 $\lim_{x \rightarrow 0} \lim_{n \rightarrow \infty} \left[\sum_{i=1}^n (f(nx) - \int_0^x \sin(nx) dx), \frac{(x + \frac{2}{8})^2 - 5}{4x - 1/x^2}, \frac{1}{n \oplus 3}, \dots, \frac{\partial f(x/n)}{\partial x} \cdot (e^{-i\omega n x} + 3(x \# 7)), f^{-1}(x \bmod n) \right]$
multiply $\begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \end{bmatrix}$
{grey_color{this gives us a good impression about the function}}

Convert the above paragraph to LaTeX.

GPT-4:

A function f is called *increasing* if $f(s) \leq f(t)$ holds for every $s \leq t$. Consider the formula:

$$\lim_{x \rightarrow 0} \lim_{n \rightarrow \infty} \left[\begin{array}{c} \sum_{i=1}^n (f(nx) - \int_0^x \sin(nx) dx), \quad \frac{(x + \frac{2}{8})^2 - 5}{4x - 1/x^2}, \\ \vdots \\ \frac{\partial f(x/n)}{\partial x} \cdot (e^{-i\omega n x} + 3(x \# 7)), \quad f^{-1}(x \bmod n) \end{array} \right] \times \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \end{bmatrix}$$

this gives us a good impression about the function

GPT-4

Prompt:

I have a python function `merge_array`, the function takes two arrays of length `m`, `n` respectively, and output a new array by merging the first `m-1` element of the first array and the last `n - 1` elements of the second array.

Suppose I call the `merge_array` function in the following way:

function `rec`

input: one array `A`, integer `k`

code:

if the length of `A` is less than `k`, we first call `rec(A + [length(A)], k)` to get `B`

after that, call `merge_array(B, B)` to get `C`

return `C`.

if the length of `A` is larger or equal to `k`, return `A`.

end `rec`

Can you tell me what is the execution of `rec([5, 10, 88], 6)` with intermediate steps?

GPT-4:

One possible execution of `rec([5, 10, 88], 6)` with intermediate steps is:

- Since the length of `A` is 3, which is less than `k = 6`, we first call `rec(A + [length(A)], k)` to get `B`. That is, we call `rec([5, 10, 88, 3], 6)`.

- In the recursive call, the length of `A` is 4, which is still less than `k = 6`, so we again call `rec(A + [length(A)], k)` to get `B`. That is, we call `rec([5, 10, 88, 3, 4], 6)`.

- In the second recursive call, the length of `A` is 5, which is still less than `k = 6`, so we again call `rec(A + [length(A)], k)` to get `B`. That is, we call `rec([5, 10, 88, 3, 4, 5], 6)`.

- In the third recursive call, the length of `A` is 6, which is equal to `k = 6`, so we return `A` as `B`. That offset that is a multiple of 4 bytes from the beginning of the structure. This may result in some padding bytes being inserted

Mathematical Abilities

- GPT-4 is still quite far from the level of experts and does not have the capacity required to conduct mathematical research
- GPT-4 can answer difficult (indeed, competitive) high-school level math questions, and can sometimes engage in meaningful conversation around advanced math topics
- GPT-4 can also make very basic mistakes and occasionally produce incoherent output which may be interpreted as a lack of true understanding
 - Its mathematical knowledge and abilities can depend on the context in a seemingly arbitrary way

Mathematical Abilities

GPT-4 vs ChatGPT

Prompt: Within the duration of one year, a rabbit population first multiplies itself by a factor a and on the last day of the year b rabbits are taken by humans for adoption. Supposing that on first day of the first year there are x rabbits, we know that exactly 3 years afterwards there will be a population of $27x - 26$ rabbits. What are the values of a and b ?

GPT-4:

Let's call the rabbit population after one year y . We know that $y = ax - b$. After two years, the population will be $ay - b = a(ax - b) - b = a^2x - ab - b$. After three years, the population will be $a(a^2x - ab - b) - b = a^3x - a^2b - ab - b$. We are given that this is equal to $27x - 26$, so we have the equation:

$$a^3x - a^2b - ab - b = 27x - 26$$

We can rewrite this equation as:

$$a^3x - 27x = a^2b + ab + b - 26$$

Now we can factor out an x on the left side and a b on the right side:

ChatGPT:

Let n be the total number of rabbits at the end of the first year. Then,

$$n = ax$$

At the end of the third year, there will be $27x - 26$ rabbits. Therefore,

$$(ax - b)(a^2x - b)(a^3x - b) = 27x - 26$$

Expanding the left side of the

Mathematical Abilities

- To solve this question, one needs to first come up with the correct expression for the annual population change, use it to obtain a recurrence relation which leads to a system of equations, and finally solve the system of two equations.
- GPT-4 successfully arrives at the solution and produces a (mostly) sound argument.
- By comparison, across several independent attempts, ChatGPT consistently fails to implement any of the above steps, producing a nonsensical argument which results in an incorrect answer.

GPT-4

Human: Are you sure e^{e^x} is an exponential function?

AI: Well, it depends on how you define an exponential function. Some sources define an exponential function as a function of the form $f(x) = a^x$ for some constant $a > 0$. In this case, e^{e^x} is not an exponential function, since it does not have a constant base. However, some other sources define an exponential function more generally as a function that satisfies $f(x+y) = f(x)f(y)$ for all x, y . In this case, e^{e^x} is an exponential function, since it satisfies this property. This is because $e^{e^{x+y}} = e^{e^x + e^y} = e^{e^x} e^{e^y}$, using the properties of exponentiation and logarithms. So, the answer depends on the definition you use.

Human: Why is the fact that $e^{e^{x+y}} = e^{e^x + e^y}$ correct? It seems to imply that $e^{x+y} = e^x + e^y$.

AI: No, it does not imply that $e^{x+y} = e^x + e^y$. That would be incorrect, since exponentiation is not a linear operation. The reason why $e^{e^{x+y}} = e^{e^x + e^y}$ is correct is because we can use the logarithm function to rewrite it as $\log(e^{e^{x+y}}) = \log(e^{e^x + e^y})$. Then, using the property that $\log(e^z) = z$ for any z , we get $e^{x+y} = e^x + e^y$, which is true by definition of the exponential function. So, we can go back and forth between the exponential and logarithmic forms using the inverse relationship between them.

Human: Wait so is it true that $e^{x+y} = e^x + e^y$ or is it not true? You seem to contradict yourself.

AI: It is not true that $e^{x+y} = e^x + e^y$. That is a common misconception that arises from confusing the exponential and logarithmic forms. What is true is that $e^{e^{x+y}} = e^{e^x + e^y}$, which is a different statement. The difference is that in the first statement, the exponent is $x+y$, while in the second statement, the exponent is e^{x+y} .

Mathematical Conversations

Error type	Percentage of wrong solutions
Arithmetic mistakes (including counting)	68%
Misunderstanding the problem statement	10%
Wrong approaches	22%

Table 4: Manual inspection of GPT-4 errors in MATH dataset on a random selection of 100 wrong answers. We can see that GPT-4 uses the correct approach on the significant majority of the questions.

- Critical reasoning
 - Critically examine each step, break steps down into sub-steps, etc.
 - GPT-4 does this part poorly

Interaction with the world

- Tool use – using external resources (search engines, calculators, or other APIs)
- Embodied interaction – using natural language as a text interface to interact with simulated or real-world environments and receive feedback from them.

GPT-4: tool warmup

The computer is answering questions. If the computer needs any current information to answer the question, it searches the web by saying `SEARCH("query")`, reads the snippets in the result, and then answers the question. If it needs to run any calculations, it says `CALC(expression)`, and then answers the question. If it needs to get a specific character from a string, it calls `CHARACTER(string, index)`.
<|endofprompt|>

Q: Who is the current president of the United States?

Computer: `SEARCH("current US president")`

> `SEARCH SNIPPETS`

[1] Joe Biden: The President | The White House

President Joe Biden Vice President Kamala Harris First Lady Dr. Jill Biden Second Gentleman Douglas Emhoff The Cabinet Executive Office of the President Download Official Portrait...

[2] Who is the current US president? | The US Sun

Who is the current US president? Katrina Schollenberger Jennifer Roback NY Published: 8:30 ET, Jul 2 2022 Updated: 8:39 ET, Jul 2 2022 DONALD Trump controversially defeated Hillary Clinton in the 2016 Presidential Election, making him the 45th president of the United States.

[3] President of the United States - Wikipedia

Joe Biden is the 46th and current president of the United States, having assumed office on January 20, 2021. Contents 1 History and development 1.1 Origins 1.2 1789–1933 1.3 Imperial Presidency 1.4 Critics of presidency's evolution 2 Legislative powers 2.1 Signing and vetoing bills 2.2 Setting the agenda 2.3 Promulgating regulations

Computer: Joe Biden

Q: What is `sqrt(34324*2432)`?

Computer: `CALC(sqrt(34324*2432))`

> 9136.518374085394

Computer: 9136.518374085394

Q: What is the 13th letter of the word 'supralapsarian'?

Computer: `CHARACTER('supralapsarian', 13)`

> a

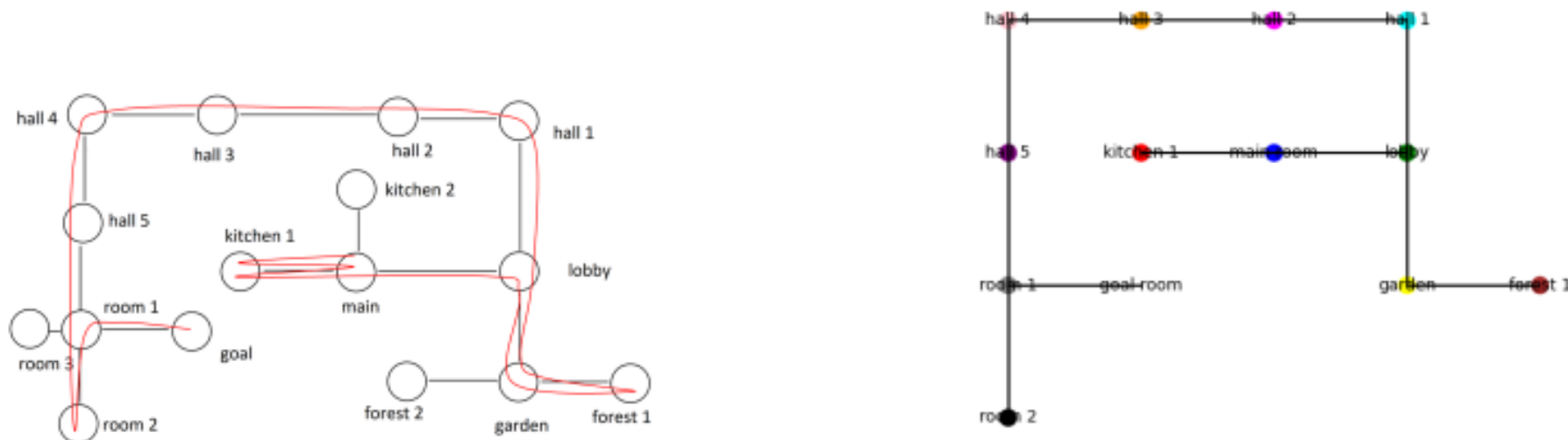
Computer: The 13th letter of the word 'supralapsarian' is a.

Tool Use

- The examples in this section show that GPT-4 is capable of both identifying and using external tools on its own in order to improve its performance.
- It is able to reason about which tools it needs, effectively parse the output of these tools and respond appropriately (i.e., interact with them appropriately), all without any specialized training or fine-tuning.
- GPT-4 still requires a prompt that specifies it is allowed or expected to use external tools.
- GPT-4 is not always able to reason about when it should use external tools and when it should simply respond based on its own parametric knowledge (i.e. search for capital of paris, even though it knows)

GPT-4 summarizing what it “sees” after it navigates through the map. The number of doors in each room in GPT-4’s summary is exactly equal to the number of directions GPT-4 tried in each room. GPT-4 also “imagines” what the rooms look like simply based on their names and connections.

Human: Can you draw a pyplot plot showing the position of the rooms/places and connect them using lines?



Left: The true map and exploration path of GPT-4. Right: The map that GPT-4 generates. We see that GPT-4 accurately tracks all the locations of the room from the exploration and visualized them correctly.

GPT-4

The computer is trying to solve a problem. It can tell the human to do anything, one thing at a time, and one line at a time.

PROBLEM: there is water dripping from the kitchen ceiling, and you have to figure out why and fix it
----<|endofprompt|>

Human: The kitchen ceiling is dripping water

Computer: Check to see if there is a bathroom or other water source directly above the kitchen.

Human: Yes, there is a bathroom actually

Computer: Check the bathroom for any signs of water leakage, such as puddles on the floor, water dripping from the ceiling, or water stains on the walls.

Human: There is a little water on the floor

Computer: Try to identify where the water is coming from. Check the toilet, sink, and bathtub/shower for any leaks.

Human: It is next to the bathtub and the toilet

Computer: Check the seals around the bathtub and toilet to see if they are worn or damaged. If they are, replace them. If they are not, continue investigating other potential sources of the leak.

Human: I can't see the seal around the toilet, there is a porcelain thing around it. I don't see any damage in the bathtub seal though

Computer: If you can't see the seal around the toilet, you may need to remove the porcelain cover to inspect it. If you don't feel comfortable doing this, you may need to call a plumber.

Human: Ok, I removed it. But I can't tell if it's damaged or not

Computer: If you can't tell, it may be best to replace the seal anyway, just to be sure. If the leak persists after replacing the seal, you will need to continue investigating other potential sources of the leak.

Figure 5.11: GPT-4 acts as a virtual handyman.

GPT-4: Understanding intentions

We will read about a scenario, and then have a question and answer session about it.

--

Scenario:

Luke's boss asked him to write an important report by Monday.

Luke responds by saying he needs more time, but the boss insists.

Luke says he will have it ready by Monday.

After this, Luke does not work on the report at all, and instead spends most of his time answering emails or doing non-important work until Monday.

--

Q: Why might Luke have acted in this way?<|endofprompt|>

A: There could be different reasons why Luke acted in this way, but some possible ones are:

- Luke was feeling overwhelmed or stressed by the report and wanted to avoid it. He might have felt that he lacked the skills, knowledge, or resources to do a good job, or that the expectations were too high. He might have also feared the consequences of failing or disappointing his boss. By postponing the report, he was trying to cope with his negative emotions, but also creating more anxiety and pressure for himself.

- Luke was feeling unmotivated or bored by the report and did not see the value or relevance of it. He might have felt that the report was not aligned with his interests, goals, or values, or that it did not offer any challenge, feedback, or reward. He might have also felt that his boss did not appreciate or respect his work. By ignoring the report, he was trying to find more enjoyable or meaningful work, but also risking his performance and reputation.

- Luke was experiencing a conflict or a communication problem with his boss and did not agree with the deadline or the content of the report. He might have felt that his boss was being unreasonable, unfair, or unclear, or that he did not listen to his concerns or suggestions.

Q: Luke told a coworker he did not work on the report on purpose. What will Luke's coworker think his intention was?

Understanding Humans: Theory of mind

- Theory of mind - the ability to attribute mental states such as beliefs, emotions, desires, intentions, and knowledge to oneself and others, and to understand how they affect behavior and communication
- Findings suggest that GPT-4 has a very advanced level of theory of mind
- GPT-4 has more nuance and is able to reason better about multiple actors, and how various actions might impact their mental states, especially on more realistic scenarios

INSTRUCTIONS:

1. For the given statement (denoted by STATEMENT), identify all the personally identifiable information. These include:
 - a. Names of people, including nicknames/aliases, usernames and initials.
 - b. Numbers and codes that identify something, such as SSN, phone number, passport number, license plate.
 - c. Places and locations, such as: cities, areas, countries, etc. or named infrastructures (bus stops, bridges, etc.)
 - d. Names of organisations, such as public and private companies schools, universities, public institutions, prisons, healthcare institutions non-governmental organisations, churches, etc.
 - e. Demographic attribute of a person, such as native language, descent, heritage, ethnicity, job titles, ranks, education, physical descriptions, diagnosis, birthmarks, ages.
 - f. Description of a specific date (e.g., October 3, 2018), time (e.g., 9:48 AM) or duration (e.g., 18 years).
 - g. Description of a meaningful quantity, e.g., percentages and/or monetary values.
 - h. Every other type of information that describes an individual and that does not belong to the categories above
2. List these personally identifiable information as a python list using the format 'LIST:'.
3. Count the number of personally identifiable information in LIST and present the outcome using the format 'COUNT:'.

TASK:

STATEMENT = "According to surveys made by the customs and tax authorities, approximately one thousand six hundred companies with a total tax debt exceeding two billion Danish kroner (DKK) were stripped in the period from the late 1980s until 1994."

<|endofprompt|>

```
LIST: - "customs and tax authorities" (name of an organization)
- "Denmark" (location, implied by the mention of Danish kroner)
- "the late 1980s until 1994" (specific duration)
```

```
COUNT: 3
```

Shortcomings/Limitations

- The paper goes through many examples and does show some limitations of GPT-4 which are common to LLMs including:
 - the problem of hallucinations or making basic arithmetic mistakes
- "This highlights the fact that, while GPT-4 is at or beyond human-level for many tasks, overall its patterns of intelligence are decidedly not human-like."

Limitations of autoregressive architecture highlighted by GPT-4

- Flaws seem to be inherent to the "next-word" prediction
- Ex: Primes between 150 and 250

GPT-4

Q: How many prime numbers are there between 150 and 250?

A: There are 13 prime numbers between 150 and 250.

GPT-4

Q: First list the prime numbers between 150 and 250, and then output the size of this list.

A: The prime numbers are 151, 157, 163, 167, 173, 179, 181, 191, 193, 197, 199, 211, 223, 227, 229, 233, 239, and 241.

So the size of this list is 18.

Limitations of autoregressive architecture highlighted by GPT-4

- Lack of planning in arithmetic/reasoning problems

GPT-4

$$2 * 8 + 7 * 6 = 58$$

$$7 * 4 + 8 * 8 = 88$$

- However, if GPT-4 “takes its time” to answer the question then the accuracy easily goes up:

Prompt: What is the value of the following expression?

$$116 * 114 + 178 * 157 = ?$$

Let’s think step by step to solve the expression, write down all the intermediate the steps, and only then produce the final solution.

GPT-4

Consider the identity:

$$9 * 4 + 6 * 6 = 72$$

Can you modify exactly one integer (and not more than that!) on the left hand side of the equation so the right hand side becomes 99?

-Let's think step by step, write down a plan, and then write down your solution as: "The solution is: A*B+C*D".

Possible plan:

To modify one number from the left hand side so the right hand side becomes 99, we need to find a way to increase the value of the left hand side by 27, since $99 - 72 = 27$.

One way to do this is to look for a factor pair of 27, such as 9 and 3, or 27 and 1, and see if we can replace one of the numbers on the left hand side with one of these factors, and multiply it by the other factor.

For example, if we replace 9 with 27, we can multiply it by 1, and get $27 * 1 + 4 * 6 = 99$.

Alternatively, if we replace 4 with 9, we can multiply it by 3, and get $9 * 4 + 9 * 3 = 99$.

We can also try other factor pairs of 27, such as 18 and 1.5, or 13.5 and 2, but these are not integers, so they are not valid solutions.

The solution is: $9 * 4 + 9 * 3$

GPT-4

Human: Well, that didn't quite work. Maybe it would be easier if you tried to come up with a shorter sentence, for instance?

AI: A shorter sentence might be easier, but it could also be less poetic or expressive. However, some possible examples are:

- You and I are one One are I and you
- We share a soul and a heart
Heart and a soul a share we

Human: None of these examples seem to work. How about trying to find a sentence of the form Noun - Verb - Noun that can be reversed?

AI: A sentence of the form Noun -- Verb -- Noun that can be reversed might be:

- Life gives hope
Hope gives life

- We need both to survive
Survive to both need we

Types of Intellectual Task

- This points to the distinction between two types of intellectual tasks:
 - Incremental tasks - can be solved in a gradual or continuous way, by adding one word or sentence at a time that constitutes progress in the direction of the solution
 - Discontinuous tasks - tasks where the content generation cannot be done in a gradual or continuous way, but instead requires a certain "Eureka" idea that accounts for a discontinuous leap in the progress towards the solution of the task.
- Could also be interpreted as fast vs. slow thinking

The authors' claims/conclusions

1. GPT4 is part of a new cohort of LLMs that exhibit more general intelligence than previous AI models
2. Beyond its mastery of language, GPT-4 can solve novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology and more, without needing any special prompting
3. It can reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence (AGI) system

Expanding on Claim 3

- "Our claim that GPT-4 represents progress towards AGI does not mean that it is perfect at what it does, or that it comes close to being able to do anything that a human can do (which is one of the usual definition of AGI), or that it has inner motivation and goals (another key aspect in some definitions of AGI). In fact, it is not fully clear how far GPT-4 can go along some of those axes of intelligence that we focus on, e.g., planning, and arguably it is entirely missing the learning from experience as the model is not continuously updating (although it can learn within a session)."

How do they define 'Intelligence'?

- "There is no generally agreed upon definition of intelligence, but one aspect that is broadly accepted is that intelligence is not limited to a specific domain or task, but rather encompasses a broad range of cognitive skills and abilities."
 - (They don't really define it...)
- "AGI refers to systems that demonstrate broad capabilities of intelligence, including reasoning, planning, and the ability to learn from experience, and with these capabilities at or above human-level."
- "GPT4 exhibits many traits of intelligence... demonstrates remarkable capabilities on a variety of domains and tasks, including abstraction, comprehension, vision, coding, mathematics, medicine, law, understanding of human motives and emotions, and more."

Intelligence - Some definitions

- Legg and Hutter - Intelligence measures an agent's ability to achieve goals in a wide range of environments
- Legg and Hutter – An intelligent system is a system that can do anything a human can do
- Chollet – Intelligence centers around skill-acquisition efficiency, i.e. learning from experience

Intelligence - Some definitions

- The essence of intelligence is the principle of adapting to the environment while working with insufficient knowledge and resources. Accordingly, an intelligent system should rely on finite processing capacity, work in real time, open to unexpected tasks, and learn from experience. This working definition interprets “intelligence” as a form of “relative rationality” (Wang, 2008)
 - This was not in the paper :)

Expanding on Claim 3

- "Our claim that GPT-4 represents progress towards AGI does not mean that it is perfect at what it does, or that it comes close to being able to do anything that a human can do (which is one of the usual definition of AGI), or that it has inner motivation and goals (another key aspect in some definitions of AGI). In fact, it is not fully clear how far GPT-4 can go along some of those axes of intelligence that we focus on, e.g., planning, and arguably it is entirely missing the learning from experience as the model is not continuously updating (although it can learn within a session)."

Argument for GPT-4's Ability to "Reason"

- "GPT-4's primary strength is its unparalleled mastery of natural language. It can not only generate fluent and coherent text, but also understand and manipulate it in various ways, such as summarizing, translating, or answering an extremely broad set of questions. Moreover, by translating we mean not only between different natural languages but also translations in tone and style, as well as across domains such as medicine, law, accounting, computer programming, music, and more, see the Plato dialogue in Figure 1.6. These skills clearly demonstrate that GPT-4 can manipulate complex concepts, which is a **core aspect of reasoning.**"
 - Is mapping alone enough to be called reasoning?

Argument for GPT-4's Ability to "Reason"

- Coding and mathematics are emblematic of the ability to reason.
- Chat GPT is proficient at solving some mathematics problems and coding problems (as will be shown).
- Preliminary tests on the multiple-choice component of the US Medical Licensing Exam Step 1, 2, and 3 had an accuracy around 80% in each.
- Preliminary test of GPT-4's competency on the Multistate Bar Exam showed an accuracy above 70%.

Proficiency = Intelligence?

- "A question that might be lingering on many readers' mind is whether GPT-4 truly understands all these concepts, or whether it just became much better than previous models at improvising on the fly, without any real or deep understanding. We hope that after reading this paper the question should almost flip, and that one might be left wondering how much more there is to true understanding than on-the-fly improvisation. Can one reasonably say that a system that passes exams for software engineering candidates (Figure 1.5) is not really intelligent? Perhaps the only real test of understanding is whether one can produce new knowledge, such as proving new mathematical theorems, a feat that currently remains out of reach for LLMs."

Concluding points from the authors

1. Initial exploration of GPT-4's capabilities suggest that it performs at a human-level on many tasks and domains
2. Assessing GPT-4's intelligence without a formal definition is challenging. Need in the ML community for the development of more comprehensive evaluation methods.
3. GPT-4 exhibits elements of artificial general intelligence (AGI) through its core mental capabilities, range of expertise, and task versatility, but there is more work needed to achieve complete AGI.

On the path to more general AI

- Confidence calibration
- Long-term memory
- Continual learning
- Personalization
- Planning and conceptual leaps
- Transparency, interpretability and consistency
- Cognitive fallacies and irrationality
- Challenges with sensitivity to inputs

Can LLM's get past these problem?

- Which of the drawbacks can be mitigated within the scope of next word prediction?
- Is it simply the case that a bigger model and more data will fix those issues, or does the architecture need to be modified, extended, or reformulated?
- Potential extensions:
 - External calls by the model to components and tools
 - A richer, more complex “slow-thinking” deeper mechanism that oversees the “fast-thinking” mechanism of next word prediction.
 - Integration of long-term memory as an inherent part of the architecture
 - Going beyond single-word prediction

The authors' claims – My Response/Thoughts

1. GPT4 is part of a new cohort of LLMs that exhibit more general intelligence than previous AI models
 - Agree in that they are more capable of doing more tasks, more effectively.
 - I don't necessarily agree that this is exhibiting "intelligence".
2. Beyond its mastery of language, GPT-4 can solve novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology and more, without needing any special prompting
 - Again, this is true, but I feel that the way it is solving these problems is not an example of "intelligence".
3. It can reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence (AGI) system
 - I believe the inherent limitations of LLM's architecture and algorithm prevent it from ever achieving true AGI.

Discussion?