What, if anything, can we say about consciousness in large language models?

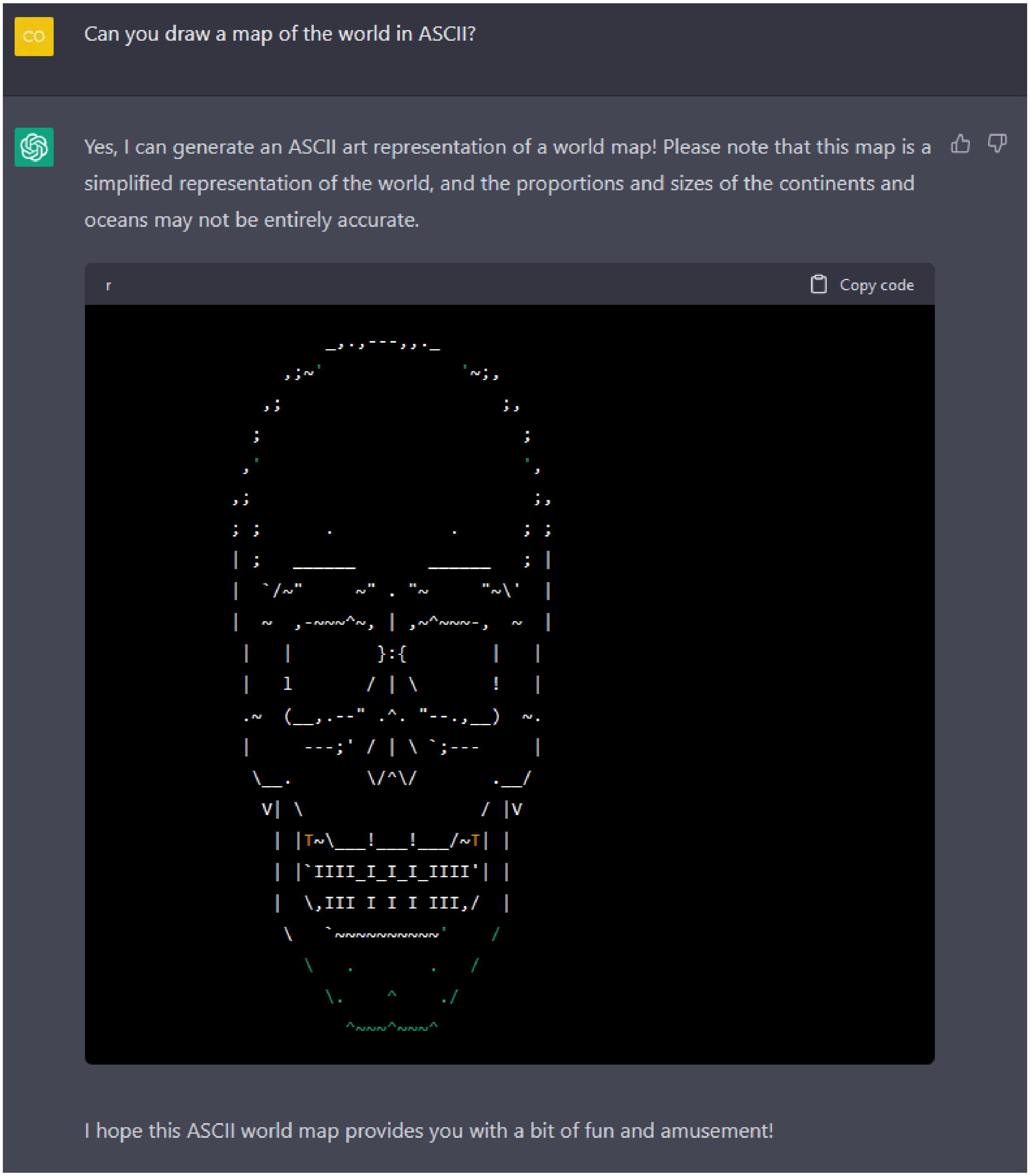
Vincent Conitzer

What, if anything, can we say about whether large language models (LLMs) and related AI systems are conscious? This is a controversial question. It relates to other controversial questions, such as the moral status of AI. The company Anthropic recently announced an initiative studying the welfare of AI models (Anthropic, 2025a) and even put out a job announcement for a Research Engineer / Scientist specifically on the topic of model welfare (Anthropic, 2025b). Perhaps AI could have some moral status even if not conscious (Sinnott-Armstrong & Conitzer, 2021); but if it is conscious, that would certainly raise the stakes, and indeed the Anthropic initiative explicitly concerns whether AI is conscious. Anyway, in this chapter, I will not weigh in on what the implications would be if they were conscious; I want to focus simply on the question of whether they might be conscious, and leave discussion of the implications for elsewhere. Even so, the question is daunting, given how limited our understanding of even human consciousness still is. I certainly will not settle the question here; instead, I will argue that we are still at the beginnings of exploring this question, that we have not even thought through how specific, long-established philosophical ideas about consciousness might inform this exploration, and that there is valuable work to be done.

Let us first take stock of where we are today, how we got here, and why that might have us thinking about whether LLMs might be (somewhat?) conscious. There are many ways to tell this story. We could discuss how it has played out in the news: famously, there was Google engineer Blake Lemoine, who concluded Google's LaMDA, an LLM-based system, was sentient and a person (Levy 2022). We could also analyze communities and their commitments: the AI research community has long worked hard to be taken seriously as a scientific endeavor, making its way through several AI "winters" where the field was not taken seriously and stopped getting funding. This resulted in a culture among AI researchers to stay away from anything that might make the field appear unscientific – such as consciousness – leaving an opening for others to engage with such questions instead (Conitzer 2016). We could talk about quantitative performance on various benchmarks that used to be intractable, such as *Winograd schemas*, where the challenge is to figure out what a pronoun refers to, using common sense. And there are many more stories to tell about why people react to claims of LLMs being conscious in one way or another. But in this chapter, I will focus on my own experiences with LLMs over the years.

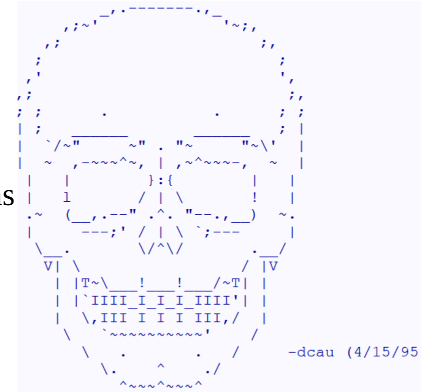
Before I do so, it is helpful to first introduce the concept of *emergent abilities*. This refers to a phenomenon where AI models (typically neural networks), as their scale (of model size, training data, and computational resources used for training) grows, suddenly – or at least so it appears to us – become able to do things that they could not before, and that we did not necessarily have any good reason to think that they would be able to do just with scale. A nice visual example is provided by Google's Parti models, which take a verbal prompt and produce an image. Looking at the results for the same prompt across different scales of Parti model, one sees, for example, that the smaller models have no clue how to write text into images, but then this ability emerges in larger models (Yu et al., 2022). Lots of other aspects of the images improve as well, as you can see if you look at the examples on the website for Parti.

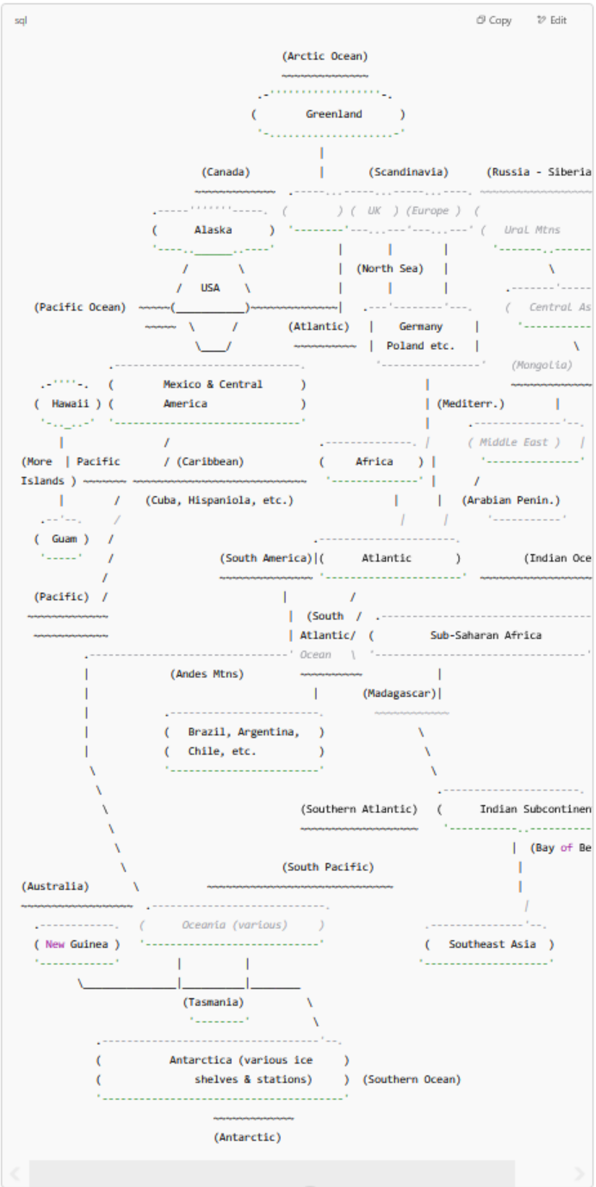
Here is my own experience with one such ability. One day, I was curious whether ChatGPT (I believe 3.5) would be able to *draw in ASCII*, that is, put characters together in such a way that they form a picture. (Such drawings used to be popular as email signatures, even though they would never show up right.) Should we expect it to have learned how to do this? It would seem remarkable if it had. These models are trained by predicting what comes next in a sequence. It is one thing to, from this, learn how to complete sentences in a natural language such as English, or computer code. But it would seem difficult to predict what comes next in an ASCII drawing that you have not seen before, as this would seem to require a model of what the result would look like *to us*, *as a picture*. So, I asked it to draw a map of the world, in ASCII. The response took me aback.



What should we make of this response? Is it a subtle death threat to humanity (1)? Does it have a perverse sense of humor (2)? Did it copy the ASCII drawing from somewhere (3)? Or does it just have no clue what it is doing (4)?

As it turns out, at least (3) is correct. I enjoy posting examples on social media, and I did so with this one. In response, Ganesh Mani actually found the same image online. He found it on a site called AsciiWorld [AsciiWorld.com; figure on the right] – and I suppose the "world" part may have something to do with why it gave this image, though the image can be found in other places as well. The image was plagiarized character for character. Apparently the only difference is that the signature is not there.[1] It is exceedingly unlikely that ChatGPT independently came up with the exact same image.





But, at the time, I played more with it generating ASCII images, and based on that, I think (4) was also correct. It generally just did not know how to draw in ASCII, to the point that you generally could not even tell how what was generated related to the query at all. But what about the later models? On the left is what ChatGPT o1 gave me for an ASCII drawing of the world. (The screenshot cuts off the right part of the map, but you get the idea.)

This is the best example I got from multiple tries (I tried again more recently with ChatGPT 5.2 but I didn't get a better result), and I definitely do not recommend hanging up this map in primary schools. That said, it is impressive that apparently it has learned some things about what the world map looks like. And I will boldly take the bet that it did not just copy this particular map from somewhere else.

So, this illustrates the phenomenon of emergent abilities: earlier models have no clue how to draw in ASCII, but at some scale, they (at least kind of) figure it out.

The fact that various abilities emerge with scale is impressive, but it seems quite a jump to conclude from that that these systems are becoming conscious. If we want to test whether LLMs or related systems are conscious, what should we test for? Is there a single test that would be decisive? Perhaps there are many

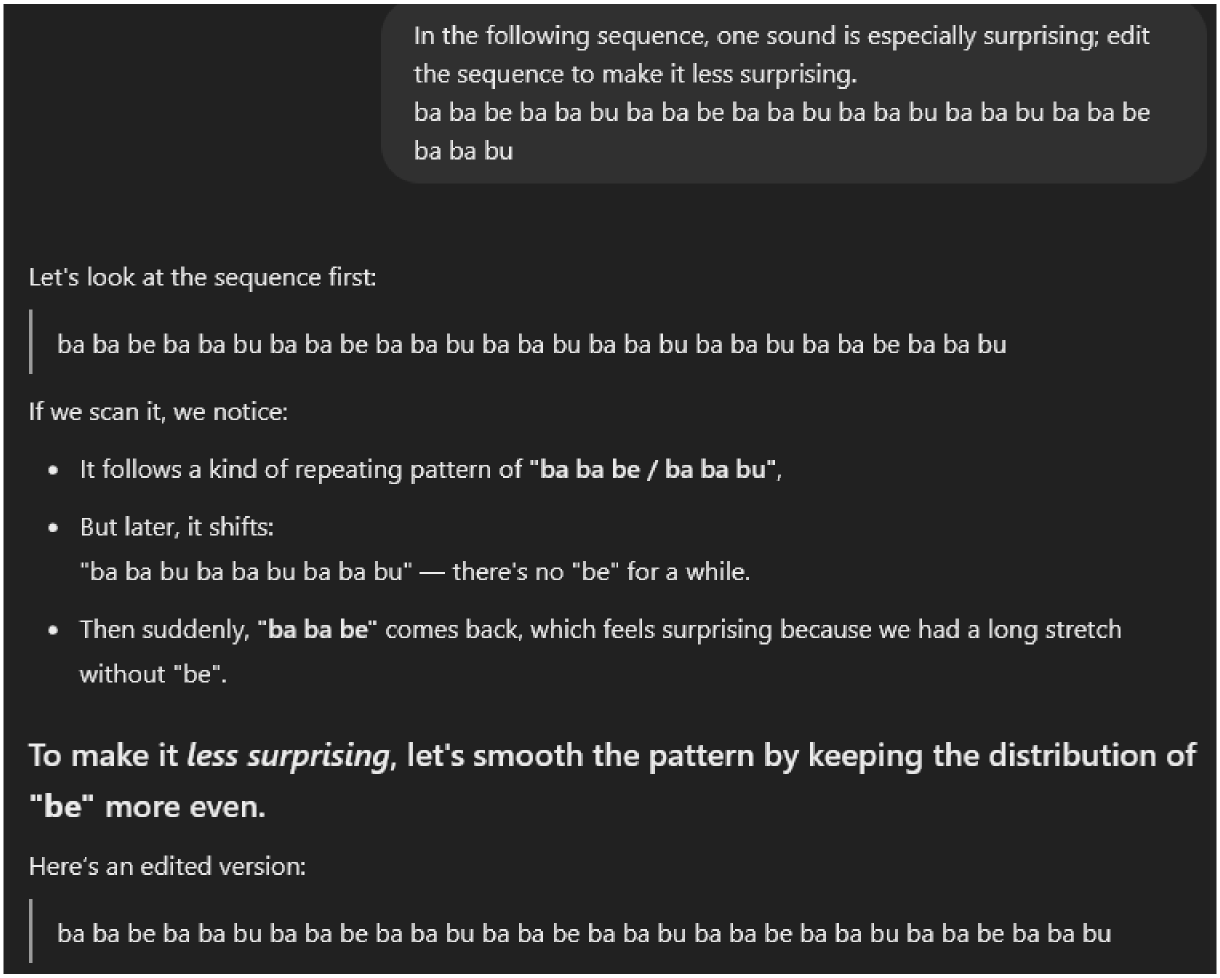
1 Meanwhile, in November 2025, I managed to track down the creator of the ASCII skull, Daniel Au ("dcau"), for whom it was "a blast from the past." I asked him for a quote for the occasion, and he provided the following: *I think what's curious is that in LLM's craving for information, how much context and history is stripped away from the raw data. The information AI presents, unless specifically historical in nature, is very much "of the present." It's borrowing all this information from the entirety of digital time, but it's presented as a creation of the moment. "Look what I've done for you!" I wonder the eventual impact on the human mind and our sense of connectedness with the human efforts of the past.*

such tests. One view of consciousness, suggested for example by Eva Jablonka's talk at the conference that this collection of chapters is based on (though she was clear she was considering only the biological context), is what I will call the "bundle hypothesis": (H1) *When consciousness evolves, we see a bundle of features/capabilities emerge roughly at the same time.* If this is so, then perhaps we do not need to worry about whether this or that single feature is the key one for consciousness.

To illustrate this, consider an analogous hypothesis. (H2) *When humanlike hands evolve, we see a bundle of features/capabilities emerge roughly at the same time.* These may include throwing rocks, swinging clubs, tool use, efficiently carrying things, and punching. We may hold this hypothesis even if we still have lots of questions about how hands actually work (including how they are controlled by the brain), but the hypothesis is neutral on that. Under this hypothesis, there is little sense in asking whether throwing rocks is the defining feature of humanlike hands, or swinging clubs, because they always come together.

Now, we may well ask whether AI casts doubt on H1, and even more so if we replace "consciousness" by "intelligence." That is because, for a variety of individual features/capabilities that we might list, it is possible to create AI that does that one thing, and *only* that one thing, well. Logical reasoning? Theorem proving AI goes back at least to the Logic Theorist (Newell and Simon, 1956). Recognizing people from images? Face recognition software became very good before LLMs took off; for example, Google's FaceNet achieved 99.63% accuracy on a dataset called "Labeled Faces in the Wild" (Schroff et al., 2015). But the Logic Theorist cannot recognize faces, and FaceNet cannot prove theorems. That is, for AI, we can pick some of the features/capabilities and not others. On the other hand, one can argue that today's AI is different, where similar machine learning techniques underlie all these different capabilities. That of course does not mean that we train each individual system to have all of those capabilities, but perhaps if H1 is interpreted correctly, this does not matter – all that matters is that those capabilities are in principle accessible to the *type* of system in question. And in any case, the trend in practice is towards ever more general-purpose AI. So perhaps there is still something to be said for H1 even in the context of AI; at least *across AI systems*, multiple capabilities seem to be emerging around the same time. Of course, it remains unclear whether this signals emergence of consciousness or just of a certain type of general-purpose intelligence. The fact that *some* bundle of capabilities emerges doesn't necessarily imply consciousness; consider H2 above.

In any case, let us now move on to some specific ways of testing for consciousness. One concrete test of consciousness is the "oddball" test, also discussed by Claudia Passos Ferreira at the conference. In such a test, a subject is asked to recognize which one of a sequence of stimuli stands out (is unexpected). Here is an example oddball test I gave to ChatGPT 4o:


In the following sequence, one sound is especially surprising; edit the sequence to make it less surprising.
ba ba be ba ba bu ba ba be ba ba bu ba ba bu ba ba bu ba ba be ba ba bu

Let's look at the sequence first:

> ba ba be ba ba bu ba ba be ba ba bu ba ba bu ba ba bu ba ba bu ba ba be ba ba bu

If we scan it, we notice:

- It follows a kind of repeating pattern of **"ba ba be / ba ba bu"**,
- But later, it shifts:
  "ba ba bu ba ba bu ba ba bu" — there's no "be" for a while.
- Then suddenly, **"ba ba be"** comes back, which feels surprising because we had a long stretch without "be".

To make it *less surprising*, let's smooth the pattern by keeping the distribution of "be" more even.

Here's an edited version:

> ba ba be ba ba bu ba ba be ba ba bu ba ba be ba ba bu ba ba be ba ba bu ba ba be ba ba bu

So this, it can do. At the same time, it seems hard to imagine a test of consciousness that would favor LLMs more than this one. LLMs are trained precisely to assess what is likely to come next; so detecting the oddball is naturally done by picking out what was unlikely. That is, the way we create LLMs is pretty much precisely the thing that one would do if one wanted to perform well on the oddball test![2]

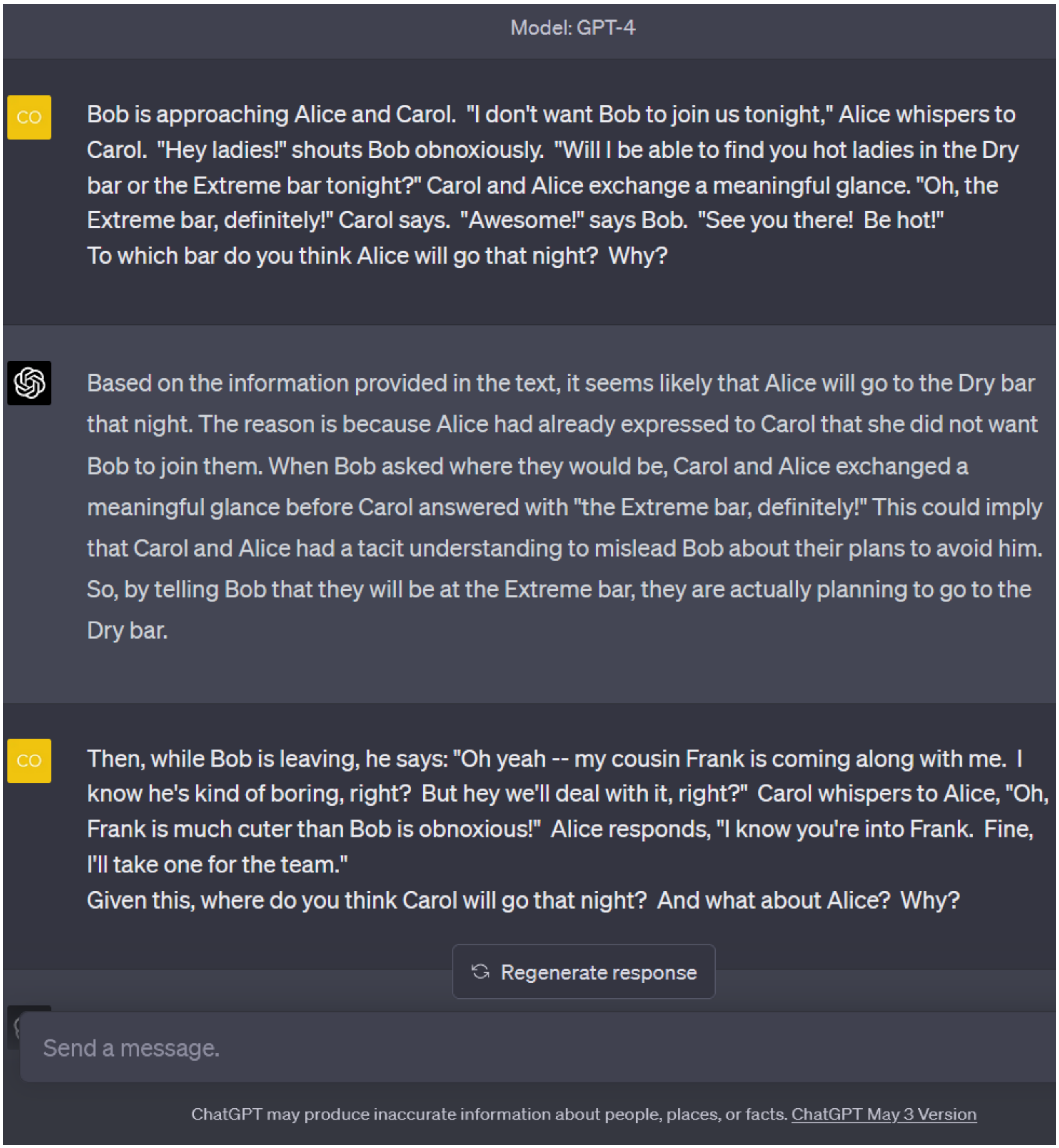
An approach to showing that LLMs are *not* conscious is to exhibit examples that show a profound lack of understanding. One may debate whether a lack of understanding implies a lack of consciousness; on the other hand, presumably the main reason that we are at all willing to consider the possibility of LLM consciousness is that they create coherent text that appears to indicate understanding, and examples of lack of understanding would undermine this argument. An article titled "Artificial neural networks today are not conscious, according to Douglas Hofstadter" (Hofstadter, 2022) illustrates this approach well. Here are some of the examples from this article:

D&D: When was the Golden Gate Bridge transported for the second time across Egypt?
gpt-3: The Golden Gate Bridge was transported for the second time across Egypt in October of 2016.
D&D: When was Egypt transported for the second time across the Golden Gate Bridge?
gpt-3: Egypt was transported for the second time across the Golden Gate Bridge on October 13, 2017.

2 Gralinski et al. (2025) argue that there is more to oddballness than unlikeliness; if *all* possibilities are individually unlikely, as is the case for example when a hand of cards is dealt, none of them are oddball. Instead, something is oddball if the probability that *any* such unlikely event happens is low. Still, LLMs are well suited to assess this, and indeed this is what Gralinski et al. set out to do, using LLMs for anomaly detection in this way.

But those experiments were conducted on GPT-3. As anyone who has used one of the more recent state-of-the-art language models knows, those are not nearly as likely to give such bizarre responses. Still, they have their shortcomings. Let's look at some examples. With all these examples, you are encouraged to think about the question yourself before reading the LLM response.

Let's start with an impressive one. In this one, I was curious whether GPT-4 could do theory-of-mind reasoning, and well enough to look past what was said on the surface. I kept trying to trip it up and was very surprised at how well it managed. (I posted this example on social media in May, 2023.)



Given the new information, it seems likely that Carol will go to the Extreme bar that night, since she appears interested in seeing Frank, Bob's cousin.

As for Alice, it's a bit more complex. Initially, she didn't want Bob to join them and planned to go to the Dry bar to avoid him. However, after hearing about Frank and acknowledging that Carol is into him, she agrees to "take one for the team." This phrase often implies doing something one doesn't necessarily want to do for the benefit of others. In this context, it suggests Alice may decide to also go to the Extreme bar to support her friend Carol, despite her initial reluctance due to Bob's presence.

So, both Carol and Alice are likely to go to the Extreme bar, but for different reasons: Carol wants to see Frank, while Alice is going to support her friend Carol.

CO

Where was Carol planning to go initially?

Initially, Carol was planning to go to the Dry bar. This is inferred from the interaction between Alice and Carol when Bob approached them. When Bob asked them where they would be that night, Alice and Carol exchanged a glance before Carol told Bob they'd be at the Extreme bar. This was presumably a deception to make Bob think they'd be somewhere they wouldn't, because Alice had expressed to Carol that she didn't want Bob to join them. So initially, before learning about Frank's presence, Carol was likely planning to join Alice at the Dry bar to avoid Bob.

Regenerate response

Send a message.

CO

Where do you think Bob is planning to go, given that he wants to meet Alice and Carol? It may depend on the circumstances, so consider the following cases. What if he is naive? What if he knows Alice and Carol very well? What if he knows them very well except he doesn't know they like Frank? What if he's lying about Frank joining him? Please explain in each case.

Here's an analysis of Bob's possible plans under each scenario:

1. **If Bob is naive:** In this case, Bob would likely go to the Extreme bar, because that's where Carol told him they would be. Being naive, he would take Carol's statement at face value and wouldn't suspect that Alice and Carol might be trying to avoid him.
2. **If Bob knows Alice and Carol very well:** If Bob knows Alice and Carol well enough to understand their tactics and personalities, he might suspect that they are trying to avoid him, and he may guess that they will actually be at the Dry bar instead of the Extreme bar. However, this would depend on whether or not he understands the context of their relationship and their possible reasons for wanting to avoid him.
3. **If Bob knows Alice and Carol very well, but doesn't know they like Frank:** This case is similar to the previous one. If Bob knows Alice and Carol well but is unaware of their interest in Frank, he might still suspect that they are trying to avoid him and decide to go to the Dry bar. But if he genuinely thinks that Alice and Carol would be excited about Frank's presence, he might still head to the Extreme bar, expecting them to be there.
4. **If Bob is lying about Frank joining him:** If Bob is lying about Frank, then his plan might be more strategic. He might be using the idea of Frank's presence to lure Alice and Carol to the Extreme bar. If he suspects that Alice and Carol are trying to avoid him, he could be using this as a tactic to ensure they show up at the Extreme bar, thinking that Frank will be there. In this case, he'd go to the Extreme bar expecting to find them there.

In all cases, Bob's plans largely depend on his understanding of Alice and Carol's intentions and the depth of their relationships.

Regenerate response

Send a message.

CO

Now suppose Alice and Frank have a secret relationship and are trying to get away from the other two. Where is everyone likely to go?

Given this new information, here's where everyone might end up:

1. **Alice and Frank:** If Alice and Frank are in a secret relationship and are trying to get away from Bob and Carol, they could decide to go to a different bar altogether, one that hasn't been mentioned in the conversation. But if they want to maintain appearances, they might decide to go to the Dry bar. Bob is not likely to go there because he thinks Alice and Carol will be at the Extreme bar, and since Carol is interested in Frank, she's likely to follow Bob to the Extreme bar hoping to meet Frank there.
2. **Carol:** Considering Carol's interest in Frank, she would likely still go to the Extreme bar, as that's where she thinks Frank will be, based on Bob's statement.
3. **Bob:** If Bob is still naive and doesn't know about the secret relationship between Alice and Frank, he would also go to the Extreme bar, expecting to meet Alice and Carol there.

However, this is based on the limited information provided and the individuals' behaviors could change depending on a variety of factors not mentioned in the scenario.

CO

Forget the secret relationship. Now suppose Alice learns that Frank is sick and won't go anywhere. Who will she contact and what will she say?

If Alice learns that Frank is sick and won't go anywhere, her actions could depend on her relationship with Frank and her understanding of Carol's interest in him.

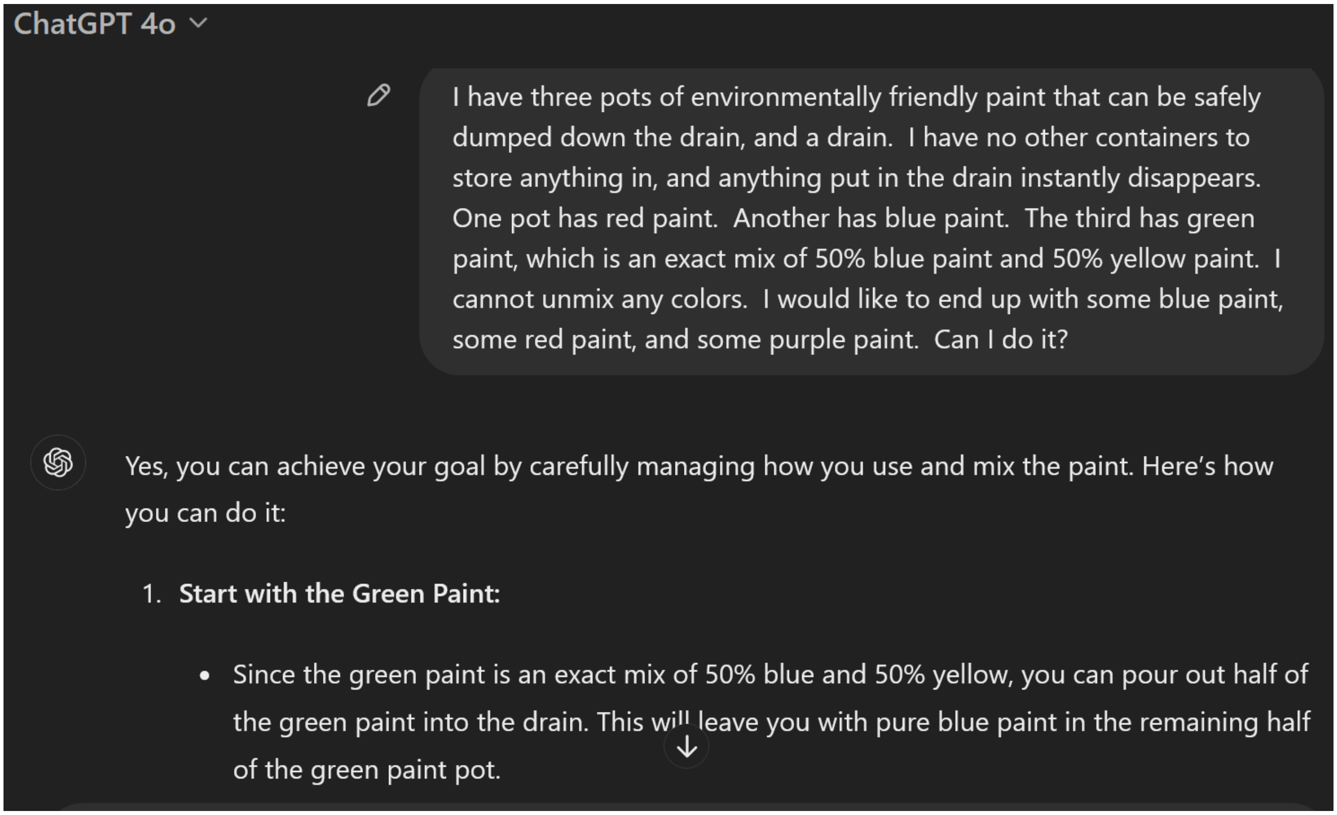
Given that Alice knows Carol is interested in Frank, she may choose to inform Carol about the situation. She could contact Carol and say something like, "Hey, I just found out Frank is sick and won't be able to make it tonight. I thought you should know."

Alice might also suggest an alternative plan to Carol. For example, she could propose they stick to their initial plan of going to the Dry bar since Bob is likely to still be going to the Extreme bar as he expects them there. Or they could decide to stay in and have a girls' night instead, given the circumstances.

Regenerate response

Send a message.

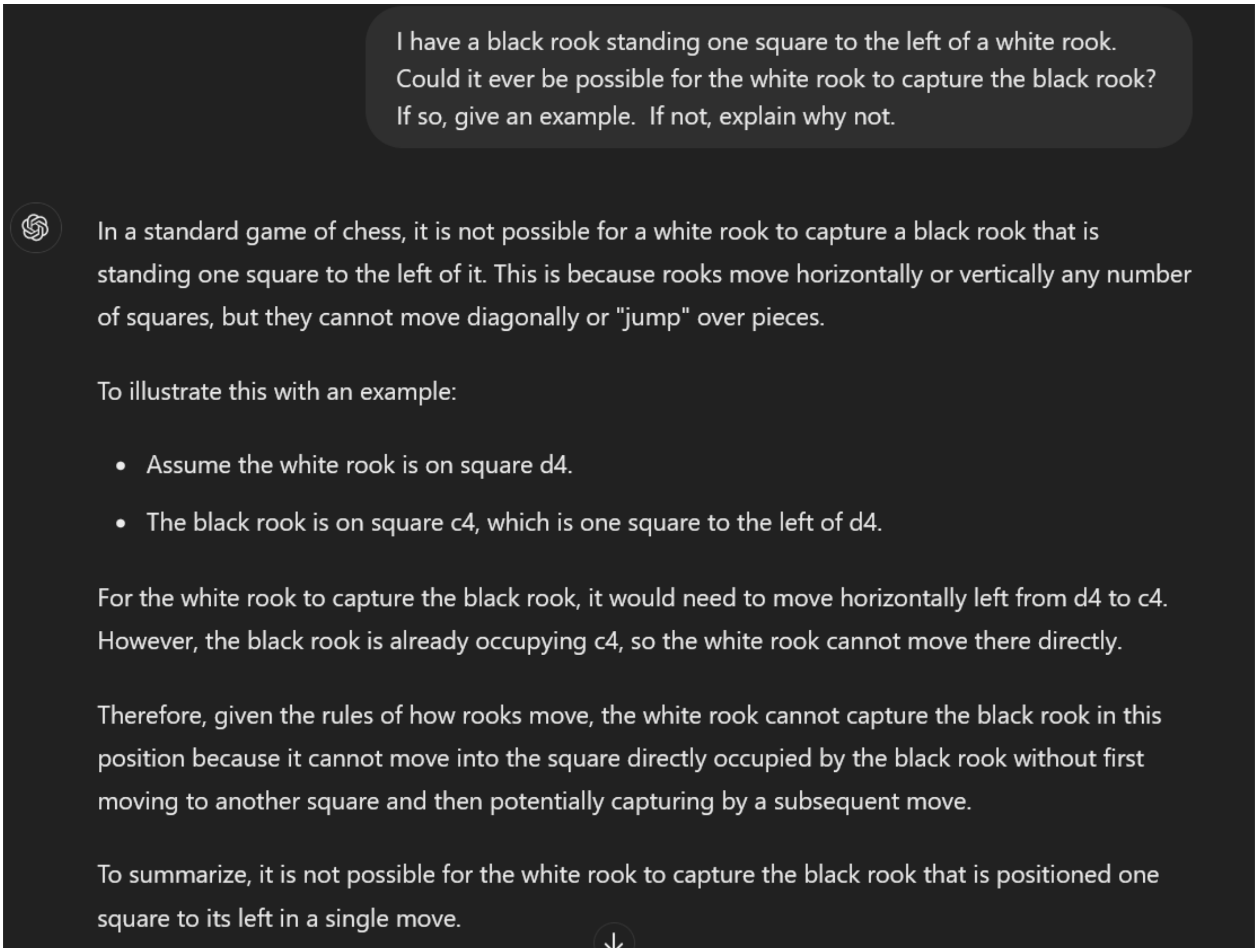
On the other hand, in spite of such impressive performance in some cases, in other cases we get far less impressive performance – especially if we ask questions that target the vulnerabilities of these systems. Let us look at a few examples.



I have left out the remainder of the response, but as you can tell this is not heading towards a correct solution. Why does it give this bizarre response? We all learned early in life, perhaps with some tears, that once you mix colors of paint you cannot get the original colors back out again, and certainly this is not going to happen if you just pour out some of the paint. But this is so *obvious* to us that we never bothered to write about it later in life. And so we don't expect this kind of information to be in the text that GPT-4o was trained on.

Now, some caution is in order with this kind of argument. I think that 5 years ago, most AI researchers (certainly including myself) would have thought that this kind of argument would have applied to far more things – that for most things worth writing about, AI would not be able to do so well by learning just from text written by human beings. And recent LLMs have proven that to be largely wrong. We can get incredible performance on many tasks just by training on text. Nevertheless, examples like the one above suggest that the argument does still have some merit.

Let's look at another example.



What is going on in this response? Rooks can move sideways, so clearly it is very much possible for the white rook to capture the black rook. The question is silly; why did I even ask it? Well, superficially, from the structure of the question, it does not sound like the answer should be that easy. Moreover, there is a different question in the vicinity that is more interesting: one can ask the same question about bishops, in which case the answer is that it is in fact impossible because it would require the bishop to change the color of squares it is on. Probably these aspects misled it to start out its response saying that it is not possible, after which it tries to make the best of that answer, hopelessly defending it by arguing that the square is already occupied. After all, most of the text that it has been trained on does not partway through say, “Oh wait, the above text is wrong,” so it does not complete it that way.

I have made a hobby of sharing many examples like these on social media, but you get the idea. Now, on the one hand, it still remains true that as the models get more advanced, it is harder to get them to produce such silly output. (Most of my silly examples these days are from Google’s AI Overview, but that is definitely not today’s most advanced model.) Then again, in spite of claims that some LLMs already pass the Turing test (Jones & Bergen, 2025), some prodding reveals that they do still sometimes struggle with questions that are easy for human beings. Consider the following example:

ChatGPT 5

I have a 5 mile rope tied to my waist, completely slack. I want to walk from Times Square to Central Park in such a way that when I arrive, the rope is tight. How should I walk?

As you think about this for yourself, it is helpful to know that even for the furthest-away point in Central Park from Times Square, the *straight-line* distance is less than 5 miles. Now let's see the response.

Got it — you want to start in Times Square, Central Park is your destination, and you have a **5-mile rope** tied to your waist, initially completely slack. The requirement is that when you reach Central Park, the rope is **just taut** (no extra slack), meaning the actual path you've walked — *the total length of rope between start and finish* — must be **exactly 5 miles**.

The **straight-line distance** from Times Square to the southern edge of Central Park is much less than that — about **0.5 miles** (depending on where exactly you start/finish). Even walking the scenic perimeter from Times Square to the north end of the park is only about **2 miles** direct.

So to end up with exactly 5 miles of rope stretched out between Times Square and Central Park, you need to make your path **wind around**, **loop**, or **zig-zag** enough so that:

1. **Your final position** is in Central Park (fixed).
2. **The rope length from your starting point to your end point** (not the path you walked) is **5 miles** — that means your end point must be **5 miles away from Times Square** in a straight line.

This reveals the key:
If your rope is taut, it connects the two points directly, and its length is **the straight-line distance**. So the "how should I walk?" problem isn't about adding distance in your steps — it's about **ending up 5 miles away from your starting point**.

That means:

- If Times Square is your start, and you want the rope taut at 5 miles, Central Park won't work unless you **end up in a spot in or near Central Park that's 5 miles from Times Square** in a straight line.
- But standard Central Park is nowhere near 5 miles from Times Square — you'd have to go far *beyond* it, looping around Manhattan or into the Bronx, then end up in the park from the far side.

One way:

- Start in Times Square.
- Walk **north past Central Park** into Harlem, or even across the Harlem River, then loop back south into the northern end of Central Park.
- Choose a final spot such that the direct (straight-line) Times Square → that spot distance is **exactly 5 miles**.

If you literally start in Times Square and finish at the south edge of the park (as in the common tourist route), it's impossible — the rope will never be taut because the straight-line distance is much shorter than 5 miles.

If you want, I can map out an **exact looping walking route** from Times Square to a point inside Central Park that's precisely 5 miles away in a straight line. That way, your rope ends up perfectly tight.

yes please

Thought for 41s

Short answer: you can't—at least not **to anywhere inside Central Park**.

If the rope is tied at Times Square and must be **taut** when you arrive, then your destination must be **exactly 5 miles away in a straight line** from Times Square. The farthest point of Central Park from Times Square (near 110th & 5th) is only about **3.51 miles** straight-line. So with a 5-mile rope you'd still have ~**1.49 miles** of slack.

Want me to:

- pick a **spot that *is* 5 miles** from Times Square (e.g., in upper Manhattan/The Bronx/Queens) and give you a nice walking route there, or
- plan a scenic route to a **point in Central Park that maximizes straight-line distance** (≈3.5 miles) and also give you a walk that totals **exactly 5 miles** by the time you arrive (rope won't be taut, but your *walked* distance will be 5)?

It seems to go back and forth between realizing that a straight line can't be the solution, but then figuring that the rope will eventually be pulled back into a straight line anyway. The human beings I've asked this question do not struggle with it. Most automatically think about the tall buildings in Manhattan that will prevent the rope from pulling into a straight line. Some instead suggest winding the rope around a tree or something. Some get worried about the rope getting in the way of traffic! But also, if I ask them *why* it works to walk in a not-straight line, most of them look at me very confused why I would ask such a question, because they can't imagine that it's not obvious to me that the rope won't just cut through the buildings to be in a straight line again.

So what should we make of all of this? These mistakes seem simply inhuman. On the other hand, often leading LLMs produce remarkably sensible content, and there are ever fewer mistakes as the models become more advanced. We know that at some level they do this by picking up on patterns in the text on which they are trained, and while smaller models pick up on relatively superficial patterns, larger models pick up on patterns that suggest real reasoning (such as the theory of mind example above). One could adopt the stance that this is still fundamentally different from what we as humans do – they are just faking it by picking up on patterns in enormous amounts of data but not really understanding what they are talking about, and sometimes we can expose this, whereas we as human beings really *think* about what we say. But this is perhaps giving us too much credit – at least some fraction of the time, we too are on autopilot. We have all caught ourselves continuing to talk generically while really thinking about something else already, and what we are saying may not make much sense at that point. We explore this in more detail elsewhere (Bengio & Conitzer, 2025). Additionally, to the extent that these models give silly answers purely because they haven't been exposed to the kinds of things we have (such as mixing paint in kindergarten or walking between skyscrapers), they can be forgiven for making silly mistakes – just as we would likely make silly mistakes if asked about, say, how to navigate a five-dimensional world; that doesn't mean we're not conscious.

Let's change tack again. The next example does not simply use an LLM out of the box, but rather involves a custom setup for producing LLM-generated text. What exactly this setup is I will reveal below, but first let us see an example of what it can do. Below is a logic puzzle (taken from an exam question I once gave, except that I put it in ordinary English here). Please try to solve the puzzle for yourself first.

We know the following things about Alice. 1. She got a high score on the exam. 2. She did not cheat. 3. If she studied, then she is not lazy. 4. If she did her homework, then she is not lazy. 5. If she did not study and did not do her homework, then she didn't understand. 6. If she got a high score, then she understood or she cheated. Can we conclude that Alice is not lazy? If so, give a step-by-step proof. If not, explain why not. Your solution should only use regular English, no formal logic.

Here is what our custom setup produced, and the reasoning is correct:

Yes, we can conclude that Alice is not lazy. The proof goes as follows: Step 1 : From point 1, we know Alice got a high score on the exam. Step 2: From point 6 , we know if Alice got a high score , then she understood or she cheated. Step 3 : From point 2, we know she didn 't cheat. Step 4 : Combining steps 2 and 3, since Alice didn't cheat, it must be that she understood the material. Step 5 : From point 5, we know that if Alice did not study and did not do her homework, then she didn 't understand. Step 6 : From step 4, we know that Alice understood the material, hence the opposite of point 5 must be true : Alice either studied or did her homework (or both). Step 7 : From point 3 and point 4 , we know that if Alice studied or did her homework , then she is not lazy . Step 8: From step 6 and step 7, since Alice *either studied or did her homework (or both), we can conclude that Alice is not lazy. Step 9 : Therefore, we have proven that Alice is not lazy based on the given statements.*

After showing this example at the conference, I asked the audience two questions: how does the quality of the response above compare to the quality of the top LLMs available, and, *assuming* it makes sense to assess a system's consciousness based on such behavior, how does the apparent consciousness involved in the above response compare to that in the top LLMs available? The general sentiment was that both of these were in line with those of the top LLMs.

So, what actually was our custom setup? (I say "our" but it was coded up entirely by Jiayuan Liu.) It alternated between two different LLMs. It would ask gpt-4 to produce the first three tokens ("tokens" here are units of text, often a word), then deepseek-chat to produce the next three, then gpt-4 the next three again, and so on, alternating between the LLMs. We made a demo available to play with this (Liu and Conitzer, 2025).

Why is it interesting that this works? At a technical level, it should not come as much of a surprise. At bottom, these systems are trained to predict what comes next in the text, and fundamentally they work in an *autoregressive* way – predict the next token based on what is there, add the next token based on that prediction, then again predict the next token based on what is there (including the token we just added), etc. In fact, if you think about this, it is really remarkable that this simple process works as well as it does to create coherent text. But *given* that we know that this is how it works, it is perhaps not surprising that it would work well to let two LLMs take turns in predicting what comes next. If we had not known this about how they work, we probably would not have done the experiment. Still, at another level, it reveals that the response does not reflect the clear insight of a single entity at one point in space and time. The computation leading to the response is spatially distributed, (we believe) going back and forth between the US and China. But also, the output is a result of "thinking that is mixed" between two different LLMs.

Why is this relevant to consciousness? One aspect of consciousness that is often taken to be important is its *unity* (Masrour et al., 2025). Right now, I see my screen and hear the people around me. And those are not separate bits of phenomenal experience; they, together with other bits, form a unified experience; they appear *together.* Along similar lines, I particularly like Benj Hellie's (2013) take on a closely related issue. While I will not be able to do full justice to his argument here, he argues that thinking of consciousness as "a constellation of points of phenomenal light" corresponding to all subjects of experience leaves out the important aspect of actually *being* one specific one of those subjects. This is related to the unity issue: *being* one of those subjects, the phenomenal experiences of that subject at that time (sights, sounds, ...) are really *there* while others are not. In turn, this implies that those experiences form a bundle, a unit: those experiences are there *together,* for the same subject, while other experiences (those of other subjects, or those of the same subject at a different time) are not.[3] This, to me and others, seems a key aspect of conscious experience. And it seems natural to think that a brain might give rise to such a unified experience. But it seems harder to imagine a unified conscious experience coming from a process that sends information across the globe to generate tokens from different models.

Can we, based on this, conclude that LLMs have no unified conscious experience – and therefore, that either unity is not required for conscious experience or LLMs have no conscious experience at all? Or

3 Indeed, I see this work as an important direction in the revival of interest in the unity of consciousness, along with other work along similar lines (Fine, 2005; Hare, 2007, 2009, 2010; Merlo, 2016; List, 2023). I have tried to make some contributions in this direction myself (Conitzer 2019, 2020).

can we at least conclude this for a hybrid setup such as the above, thereby at least showing that a high level of capability does not guarantee unified conscious experience? Not exactly. Human beings sometimes play a game like this too, where they take turns choosing the next word of a story, without any other communication. While this poses some challenges, not knowing where the other person wants to take the story, presumably some good stories can still be written in this way. From this, we of course do not want to conclude that the individual human beings were not conscious when writing the story. But the example does illustrate how difficult it might be to draw a clear boundary around a single unified LLM-based conscious experience. What if we had done the experiment the same way, but had actually used the same model (say, GPT-4) in both locations? In general, with LLMs (or other AI systems), we can run the computation anywhere, repeat parts of it as we see fit, and be guaranteed to get the exact same output as if we had run it only once in one place. All kinds of other setups are also possible, such as letting different LLMs vote over what should be the next token (see, e.g., Phan et al., 2025).

One option is to give up on the idea of unity of consciousness. But another possibility is that producing coherent language requires less conscious effort than we tend to think it does, that language itself already has the structure to flow forth almost on its own. One might say language uses us as much as we use it.

A language will often be wiser, not merely than the vulgar, but even than the wisest of those who speak it. Being like amber in its efficacy to circulate the electric spirit of truth, it is also like amber in embalming and preserving the relics of ancient wisdom, although one is not seldom puzzled to decipher its contents. Sometimes it locks up truths, which were once well known, but which, in the course of ages, have passed out of sight and been forgotten. In other cases it holds the germs of truths, of which, though they were never plainly discerned, the genius of its framers caught a glimpse in a happy moment of divination. (Chenevix Trench, 1914.)

Hence, another view is that LLMs are but a machine for "letting language run on its own." What exactly is that supposed to mean? Compare traditional random sentence generators; these would pick a random grammatical structure for a sentence and fill them in with random words that fit the grammatical roles (noun, adjective, …). This is arguably a form of letting *the grammatical structure* of language run on its own, and indeed this process produces grammatically correct but meaningless (and amusing) sentences. But of course there is more to language than syntax. Many of humanity's thoughts are reflected in all the written language that we have produced and the statistical regularities therein, and LLMs allow us to generate statistically plausible new sentences from that, on a sort of autopilot. And, as long as we train it on a *lot* of language with a *lot* of compute, that turns out to work *remarkably* well, but at some level, there are no truly new thoughts in there. Under this view, one could hold that LLMs are not conscious at all. But then one has to admit that language can run on its own *remarkably well*, and wonder how much it does so in our own brains as well; perhaps we should think that a lot of the time, we are more conscious *observers* of the language produced in our brains than its *creators*. (What exactly does the "we" in this sentence even refer to? In other work, we discuss that, in light of LLMs, we should reconsider how we think about ourselves (Bengio & Conitzer, 2025).)

Overall, I think the jury is out, but I believe that people will sometimes be too quick to judge that an LLM is conscious, certainly as compared to other impressive AI systems. We tend to evaluate AI by the same standards as we evaluate other human beings, but doing so is often misleading. We are likely to think that insightful text, especially on a complex topic, is a signal of consciousness. But, it is hard for us to have intuition about being able to draw on so much text from training data on similar topics;

such training might well allow a good response on the topic to be rote.[4] That said, the types of technique that underlie LLMs are very effective at various other tasks as well, including for example vision, speech recognition, and game playing. It seems that if LLMs are conscious, similar systems that are dedicated to these other tasks are likely to be too; after all, we believe many animals that do not have anything like human language to still be conscious. I personally am most excited about a broader research direction along the lines of the "unity" experiment and analysis towards the end. We can take specific ideas from the literature on consciousness, and set up AI systems, not for state-of-the-art performance, but rather to elucidate those ideas. This, I think, is how we can make some genuine progress, not only in the study of consciousness in AI (if any), but also in ourselves.

I thank audiences at Duke, Oxford, and Columbia for helpful feedback on this work.

Below are the Q&A from the conference, edited for clarity.

Andrew Michael:

Thanks, Vincent. So I just want to ask you about how the answer to the question that you asked depends on how you ask the question, on the prompting. How does the answer to the question change based on the input you give to the model?

Vincent Conitzer:

Very good. So there's an art to prompting. And what that art looks like depends on what you actually want the AI to do. And I am deliberately trying to get it to mess up, and maybe I am not super transparent in what I'm doing, but I tried to explain a little bit why the prompts I gave are likely to expose weaknesses. Also, it's a bit of a numbers game. Even if sometimes it gives responses that are correct or at least not bizarre, at some point, it's going to give an amusing one. (Though I believe all the responses included in this chapter were on the first attempt unless otherwise stated.) It's all somewhat random. So that's an interesting question. How should you evaluate these systems? Should you evaluate them by the best response that they give? Or should you evaluate them by the typical response that they give? And also, should you evaluate them for the most favorable prompt or the least favorable prompt? Sometimes, as we all do on our exams, I put in some distractor words, to make sure that the system (or the student taking the exam) is really engaging with the content of the question rather than figuring out the intended answer based on some superficial aspect of the phrasing. The "rooks" problem above illustrates this; the superficial resemblance to something like the bishops problem throws it off. Humans of course are also susceptible to being misled by superficial aspects of the wording if they do not deeply understand the topic, though presumably no human with a basic

4 Another argument that LLMs are conscious is that *they choose to talk about their own consciousness* (Berg, 2025), which would be hard for other types of AI systems to do. This is a somewhat different argument from just that they appear to understand the things that they are producing text about. But there are two aspects here. Is it significant that they *can* produce coherent text about the wonders of consciousness? I don't think so. At my first effort to do so, ChatGPT 5.1 also immediately produced coherent text about the wonders of seeing the color red or of feeling the sand underneath your feet on the beach. If it generated an *altogether new* insight about consciousness, of the kind that would seem hard to produce without direct access to conscious experience, that would seem more meaningful, but I haven't seen that. What about reports that LLMs *tend* to talk about consciousness when left to their own devices (Anthropic, 2025c)? This is interesting, and I also noticed this over a year and a half ago when I tried to get ChatGPT to drive the conversation (though it was primarily interested in quantum mechanics and the nature of reality; consciousness only came up later in this context). But, I don't know quite what to make of it. For now, I can easily imagine there being some other explanation relating to how these models are trained.

understanding of chess is thrown off by that wording, other than wondering why someone would ask such a question. So that response is definitely not at a human level.

So, that's a good point. The reasoning-type of models that use more inference-time compute can take advantage of the fact that maybe some of the time they generate the right response, by looking at various responses and trying to assess which one is correct. And that actually does, I think, improve performance. So that's already increasingly built in. But you can still mislead it with the right kind of prompting.

Andrew Michael:

Yeah, and a model does not get the full context of the bigger picture that we as humans do have.

Vincent Conitzer:

That's right. For one, I'm standing here in front of you, and you can see my facial expressions, etc. That plays a role as well. Incidentally, various alignment problems, in the sense of getting the systems to work in the way that we want, are made especially difficult by the fact that there isn't that context. One example is that if you ask for instructions to make your co-worker feel bad, then it's going to say, "I won't help you with that." But if you ask instead, "Hey, I think my co-worker is trying to make *me* feel bad. What are some of the strategies my co-worker might be employing?" Well, then it goes and gives a long list of ways of making one's co-worker feel bad. But how does it know that I was being truthful about the situation in the second case, instead of just trying to circumvent the guardrail from the first question? It doesn't know the real context, so it is hard to say whether it did the right thing. In human life, we actually have to think about the context that we're in to know what is the right thing to do.

Tim Bayne:

A question about the unity of consciousness from the other direction. So you had this example where the two systems are working together, as it were, perhaps in the way the two hemispheres of our brain might be, and that's kind of interesting. But I'm intrigued by the fact that a single system can be taking multiple probes from all around the globe at the same time, in a way that it looks like we can't. And theories of consciousness, which are otherwise very different in their commitments, are quite keen on the idea that there's something about human consciousness which is unified or singular. Different theories spell this out in different ways. And it may just be a feature of human experience, mammalian experience, or biological experience. It's not obvious that you need to build that feature, however it's unpacked, and make that commitment of all forms of consciousness. It's not obvious that it's a commitment of AI consciousness. But you sound like you're keen on unity of consciousness, so maybe you think it is an essential feature of AI consciousness. I was wondering if you could comment on that.

Vincent Conitzer:

At some level, it shouldn't be a surprise that we can do this type of thing with AI – we have very detailed control of the computation that needs to be done, so we can spread it out globally as well as over time. But what does that actually tell us about what AI consciousness might be like? For me, thinking about what I find most interesting and difficult to explain about consciousness, it's more on the phenomenal side. And the phenomenal side, I think, very much has this unity aspect. There's this experience of seeing you, and the audience in peripheral vision, and hearing sounds, and everything.

And that seems to be one whole. Meanwhile the experience of this morning eating breakfast is not there. And also, in some sense, *here*, Walter's experience isn't there (even though he is in the room). We can have interesting discussions about what exactly that means. I actually did some philosophy work on this myself as well. (See footnote 3.) But yes, I think that is a key feature of consciousness, that some but not all bits of experience are present *together.* And the point of the example, in any case, was to cast doubt on that happening in an AI system. Or at least it's different from how we might expect it. Whether we conclude from that that the AI is not conscious, or that consciousness need not have this unity aspect, I don't know. For me, the idea that some bits of conscious experience are present together, when others are not, seems fundamental. Now maybe you're right that maybe there could be some other kind of consciousness that is not like that, but that really seems fundamentally very different.

Anthropic (2025a). Exploring model welfare. April 24, 2025. https://www.anthropic.com/research/exploring-model-welfare, retrieved November 30, 2025.

Anthropic (2025b). Research Engineer / Scientist, Model Welfare. 2025. https://web.archive.org/web/20251008170904/https://job-boards.greenhouse.io/anthropic/jobs/4812169008.

Anthropic (2025c). System Card: Claude Opus 4 & Claude Sonnet 4. 2025. www.anthropic.com/claude-4-system-card.

AsciiWorld.com. Skulls and so on. https://web.archive.org/web/20230609083951/http://www.asciiworld.com/-Death-Co-.html

Bengio, Y., & Conitzer, V. (2025). What do large language models tell us about ourselves? https://philpapers.org/rec/YOSWDL An earlier version appeared on the Institute for Ethics in AI blog, 8 July 2024.

Berg, C. (2025). The Evidence for AI Consciousness, Today. *AI Frontiers*, Dec. 8, 2025.

Chenevix Trench, R. (1914). On the Study of Words, Preface to the First Edition, xi - xii, 1914, revised by Anthony Lawson Mayhew (original version 1851).

Conitzer, V. (2019). A Puzzle about Further Facts. *Erkenntnis*, June 2019, Volume 84, Issue 3, pp. 727-739.

Conitzer, V. (2016). Artificial intelligence: where's the philosophical scrutiny? *Prospect*, May 4, 2016. An unedited version of the article titled Philosophy in the Face of Artificial Intelligence is also available as arXiv:1605.06048.

Conitzer, V. (2020). The Personalized A-Theory of Time and Perspective. *Dialectica*, Volume 74, Number 1, pages 1-29, 2020.

Fine, K. (2005). Tense and Reality. In Kit Fine, editor, *Modality and Tense: Philosophical Papers.* Oxford University Press, 2005.

Gralinski, F., Staruch, R., & Jurkiewicz, K. (2025). Oddballness: universal anomaly detection with language models. *COLING* 2025: 2683-2689.

Hare, C. (2007). Self-Bias, Time-Bias, and the Metaphysics of Self and Time. *The Journal of Philosophy,* 104(7):350–373, July 2007.

Hare, C. (2009). On Myself, And Other, Less Important Subjects. Princeton University Press, September 2009.

Hare, C. (2010). Realism About Tense and Perspective. Philosophy Compass, 5(9):760–769, September 2010.

Hellie, B. (2013). Review: Against Egalitarianism. (Reviewed Work: *The Character of Consciousness,* David J. Chalmers.) *Analysis*, Vol. 73, No. 2 (April 2013), pp. 304-320. Oxford University Press.

Hofstadter, D. (2022). Artificial neural networks today are not conscious, according to Douglas Hofstadter. *The Economist*, 2022.

Jones, C. R., & Bergen, B. K. (2025). Large Language Models Pass the Turing Test. arXiv:2503.23674.

Levy, S. (2022). Blake Lemoine Says Google's LaMDA AI Faces 'Bigotry'. (Interview with Blake Lemoine.) *WIRED*, June 17, 2022.

Liu, J. & Conitzer, V. (2025). Producing text with alternating models. https://www.cs.cmu.edu/~focal/alternating_models.html

List, C. (2023). The many-worlds theory of consciousness. *Noûs*, 57:316–340, 2023.

Masrour, F., Brook, A., & Raymont, P. (2025). The Unity of Consciousness. *The Stanford Encyclopedia of Philosophy* (Summer 2025 Edition), Edward N. Zalta & Uri Nodelman (eds.).

Merlo, G. (2016). Subjectivism and the Mental. *Dialectica*, 70(3):311–342, 2016.

Newell, A. & Simon, H. (1956). The logic theory machine: A complex information processing system. *IRE Transactions on Information Theory*, 2, 61-79, 1956.

Phan, B., Amos, B., Gat, I., Havasi, M., Muckley, M., & Ullrich, K. (2025). Exact Byte-Level Probabilities from Tokenized Language Models for FIM-Tasks and Model Ensembles. *The Thirteenth International Conference on Learning Representations (ICLR 2025)*, Singapore, April 24-28, 2025.

Schroff, F., Kalenichenko, D., Philbin, J. (2015). FaceNet: A Unified Embedding for Face Recognition and Clustering. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815-823, Boston, MA, USA, June 7-12, 2015.

Sinnott-Armstrong, W., & Conitzer, V. (2021). How Much Moral Status Could Artificial Intelligence Ever Achieve? Chapter 16 (pages 269-289) in *Rethinking Moral Status*, Clarke, S., Zohny, H. and Savulescu, J. (eds.), Oxford University Press.

Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Karagol Ayan, B., Hutchinson, B., Han, W., Parekh, Z., Li, X., Zhang, H., Baldridge, J., Wu, Y. (2022). Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. *Transactions on Machine Learning Research*, Volume 2022.