A Turing Test for Self-Awareness

Anonymous Author(s)
Affiliation
Address
email

Abstract

I propose a test for machine self-awareness inspired by the Turing test. My test is simple, and it provides an objective, empirical metric to rectify the ungrounded speculation surging through industry, academia, and social media. Drawing from a breadth of philosophical literature, I argue the test captures the essence of self-awareness, rather than some postulated correlate or ancillary quality. To begin, the concept of self-awareness is clearly demarcated from related concepts like consciousness, agency, and free will. Next, I propose a model called the *Nesting Doll of Self-Awareness* and discuss its relevance for intelligent beings. Then, the test is presented in its full generality, applicable to any machine system. I show how to apply the test to Large Language Models and conduct experiments on popular open and closed source LLMs, obtaining reproducible results that suggest a lack of self-awareness. The implications of machine self-awareness are discussed in relation to questions about meaning and true understanding. Finally, some next steps are outlined for studying self-awareness in machines.

1 Introduction

At what point can we say a machine's eyes have been opened? When can we say it has become *like us*? After what moment can we say it knows good and evil?

Such questions have met idle speculation for millennia, but today they rapidly approach a fever pitch, demanding answers with unprecedented urgency. AI systems that can pass for human in many respects are no longer fiction. Machines that can walk and talk are real and functional. What was once a distant speck, barely visible on the horizon, is now barreling down upon us.

Through much of the history of AI, the Turing test served to keep these worries at bay [1]. Originally called the imitation game, this rudimentary metric of AI progress is a game played by two humans and one machine. One human engages in conversation with the machine and the other (the judge) must identify which is which, using nothing but the text of the conversation. The machine is deemed intelligent if it can fool the judge by mimicking human dialogue. While far from perfect, the Turing test was a concrete, unambiguous bar for AI to clear—and one that stayed comfortably out of reach for a long time.

Last year, however, the Turing test was broken [2]. Large Language Models (LLMs) such as ChatGPT can handily engage in fluent conversation, on top of generating convincing essays, passing difficult exams, and even writing poetry. With the Turing test no longer a target in the distance, the conversation on AI has become untethered to any definitive, objective measure or permanent, agreed-upon benchmark. As such, extreme subjectivity, soaring fantasies, and flights of fancy have become commonplace. For instance, over the last year we have read "Blake Lemoine claims language model has a soul" [3], "Claude 3 realizes it's being tested" [4], "Researchers say chatbot exhibits self-awareness" [5] and much more. A new objective is dearly needed.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

1.1 Related Work

1.1.1 Other Tests

To the best of my knowledge, very little work has been done to devise any objective test or benchmark for machine self-awareness, especially in the literature. There are several reasons for this. Discussed further in section 2, imagining empirical measures that actually work is difficult, self-awareness is often entangled with consciousness, free will, agency, etc., and it is hard to define. Worse, the topic is seen by many in academia as somewhat taboo—appropriate for the philosophy departments but not any kind of rigorous science.

The result is that popular Tweets and news media dominate the conversation, while authorities in the field either say nothing or win the spotlight with bold, confident assertions based on implicit, controversial assumptions or their intuition about a model's architecture. This situation is concerning; as AI systems get better and better, how will we truly know when they cross that fine line? Even if you object to everything else in this paper, I argue this question is at least worthy of real scientific investigation.

Much to the point, the only directly related work I could find is the AI mirror test, proposed recently by Twitter user nielsrolf [6], and later (going viral, reaching 3.2 million impressions) by Josh Whiton [7]. Inspired by the classic mirror test whereby animals are presented with a mirror and observed, in the AI mirror test, popular chatbots are shown a screenshot of the chat window and asked to describe what they see. This test is interesting in its own right, but I will argue it does not demonstrate any sort of self-awareness in the manner it is formulated.

1.1.2 Work on Self-Awareness

While there are no benchmarks for machine self-awareness, there is an immense amount of work in the philosophical literature—far more than I have space to mention here. In this section I will give merely a partial and incomplete sketch of a few important ideas written on the topic. For a more comprehensive introduction to the work on self-consciousness, the survey by Joel Smith is a great resource [8]. For a variety of introspective, or phenomenological approaches, consult [9], and for an overview of the broader concept of consciousness, consult [10].

Perhaps the earliest writing of the concept of self-awareness was in Sophocles' *Oedipus*. Joel Smith writes

> Oedipus knows a number of things about himself, for example that he was prophesied to kill Laius. But although he knew this about himself, it is only later in the play that he comes to know that it is he himself of whom it is true. That is, he moves from thinking that the son of Laius and Jocasta was prophesied to kill Laius, to thinking that he himself was so prophesied. It is only this latter knowledge that we would call an expression of self-consciousness [8].

Oedipus demonstrates self-awareness when he recognizes the prophecy is about himself. Before that recognition, Oedipus treats the prophecy as just another part of the world he observes; yet afterwards, he realizes it is directly related to his own actions. I will refer back to this example when developing the test.

Nearly every philosophy and religion has had something to say about self-awareness. Adam and Eve can be viewed as gaining self-awareness in the garden when they "realize they are naked" [11][12]. Aristotle claims that, to perceive any external thing, one must also perceive their own existence [8]. The Buddhist doctrine of anattā, roughly "not-self," maintains that there is no permanent, underlying self or soul [13]. Descartes, in contrast, with the well-known *cogito ergo sum*, posits the self as known with certitude *a priori* [14]. William James divided the self into four constituents; the material self, the social self, the spiritual self, and the pure ego [15]. Wittgenstein likens the self to the eye that sees but cannot see itself [16]. More recently, some of the philosophical ideas on self-awareness have been applied to the fields of cognitive science and neuroscience [17][18].

1.2 Related but Separate Concepts

Before presenting the test, we must clearly demarcate the concept of self-awareness.[1]

1.2.1 Solipsism and Philosophical Zombies

First, note that self-awareness is not the same as consciousness. On the question of whether *there is something it is like* to be a machine [19], I will remain silent here. Some approaches in the phenomenological literature attempt to draw connections between consciousness and self-awareness [9]. However, here it will be most useful for us to cleanly separate these two concepts.

It is interesting to consider whether an entity can be self-aware without being conscious, but it is outside the scope of this paper. Thus, it will remain open whether philosophical zombies might be self-aware [20], or whether any kind of test could solve the problem of other minds [21].

1.2.2 Freedom of the Will and Agency

Another related ability that intelligent systems may or may not possess is free will [22]. In science-fiction depictions of intelligent machines, the light of self-consciousness often coincides with agency and free will. Indeed, the concepts seem very tightly related at face value, yet they are not the same.

Agency can be defined as a being's "capacity to take actions, especially with intention" [23]. Note that, by itself, agency does not necessarily imply any sophisticated degree of perception or awareness, even though (practically speaking) any being which takes actions will likely have to sense their environment.

The freedom of the will is far more difficult to define, and perhaps among the most controversial of philosophical ideas. It designates a particular level of control a being has over their actions—but fierce debates rage over whether this control is undetermined by prior causes, compabitible with determinism, an illusion, etc. [22].

Self-awareness is not the same as free will, and self-awareness is not the same as agency—all three of these are separate concepts. As with consciousness, we can only make forward progress if we are crystal clear about what is under analysis and what is left outside of scope.

1.3 Paper Roadmap

In this paper, I propose a test for machine self-awareness which is similar in style to the Turing test. Like the Turing test, the test I propose is imperfect and rudimentary. Yet, it offers a compelling alternative to the ungrounded speculation surging through the field of AI. Moreover, I argue it truly captures the essence of self-awareness, rather than some postulated correlate or ancillary quality.

In section 2, I present my test in full generality, applicable to any machine system. I also illustrate the *Nesting Doll of Self-Awareness*, and discuss its importance for understanding self-awareness in complex systems or beings. In section 3, I will describe the experimental methods to assess self-awareness in LLMs. In section 4 I will present the results of these experiments, of which a selection are shown in appendix B. In section 5, I will discuss the implications of self-awareness, its relation to meaning and the understanding, and consider how humans would perform on my test. Finally, in section 6 I discuss next steps.

2 A Test for Self-Awareness

2.1 The Essence of Self-Awareness

What kind of test could possibly tell a system with self-awareness from a system without? The central challenge is that any test we dream up must be based in empirical observations of the machine's behavior or output. Worse, the machines we will study are trained specifically to mimic the behavior and outputs of humans! How can we tell between real self-awareness and the illusion of self-awareness?

[1] Note that, for the sake of this paper, self-awareness is used interchangeably with self-consciousness.

No matter how well a system can imitate human behavior and outputs, there will always be one fundamental difference. There is one thing that a self-aware system is able to do that an imitator will never be able to. This is the essence of self-awareness:

If a system is self-aware, then it is aware of itself.

So far, it seems we have said nothing. But if we apply this formula to familiar cases, we will begin to see why it works.

Imagine an infant staring blankly in the mirror, compared to a child who looks in one and sees their own reflection. What is the difference between these cases? In the latter case, the child is aware of itself—it can point and say “that’s me!" It can recognize itself, perceive itself, distinguish itself in the reflection. Within its vast field of experience, through the window of its senses, it can differentiate which parts are *itself* and which parts are *not*. Critically, awareness (here used interchangably with perception, recognition, experience, etc.) is only possible *through* the child’s inputs (senses). Within this field of inputs, a line must be drawn between *me* and *not-me*; and, when this line is drawn correctly, we declare the system self-aware. A test for self-awareness must capture its essence, or else better and better imitations may fool us with the illusion of self-awareness.

While our description is still very high-level, I argue that the understanding of self-awareness developed here is consistent with the philosophical work outlined in section 1.1.2, along with most (if not all) popular conceptions. In the next section, the concept of a system is illustrated in much more detail, and a formal, rigorous definition is provided in appendix A.

2.2 The Test for Machine Self-Awareness

The concept of a machine, or system, is illustrated in Figure 1. For a more formal treatment based in the literature on abstract systems, refer to Definition 1 in appendix A. Here, the system is separated from the world, with which it interacts through inputs and outputs. We may think of inputs as senses and outputs as actions or words.

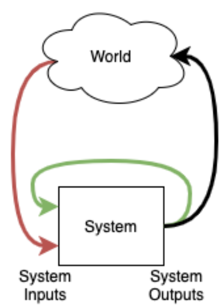


Figure 1: A system which may or may not be self-aware.

With this image of a system in mind, the test for machine self-awareness is simply as follows:

Can the system correctly distinguish the green inputs from the red?

If it can, then in a literal sense, it will be recognizing itself in the inputs. If it can, it will be like the child who recognizes their reflection in the mirror. If it can, it will be self-aware.

2.3 Levels of Self-Awareness

So far, it seems we have presented self-awareness as all-or-nothing. The reality is more complex, however.

To capture this nuance, I propose a model called the *Nesting Doll of Self-Awareness*, developed in discussions with {removed to preserve Anonymity}. The essential idea is that system outputs may loop back to the input more or less tightly, with varying levels of environmental mediation, depicted in Figure 2.

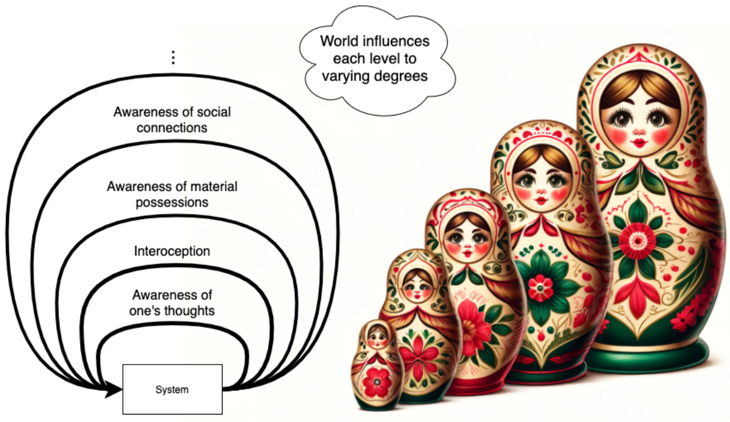


Figure 2: The Nesting Doll of Self-Awareness.

The tightest loop is associated with one's awareness of their own inner thoughts. Even at this innermost level, it is not trivial to distinguish which inner thoughts are your own and which are not. For a concrete example of why, consider the classic movie *Inception*. The entire plot revolves around an attempt to implant another person's idea into a target's unconscious—in the movie, it is the idea of Robert Fischer's dying father telling him to "create something for himself" [24]. Robert treats this idea as though it was the green arrow in Figure 1, when in fact it was the red. Of course, *Inception* is a work of fiction, yet it dramatically highlights a critical theme in human affairs, which insinuation and the power of suggestion also play upon.

One level up is associated with interoception, such as hunger signals or the movement of one's limbs. Here, the importance of distinguishing your influence from the world's is clearer—life would be difficult if you couldn't tell the difference between you moving your arm, and someone else moving it for you.[2] If you jump in surprise when someone sneaks up behind you and puts a hand on your shoulder, then you possess this level of self-awareness.

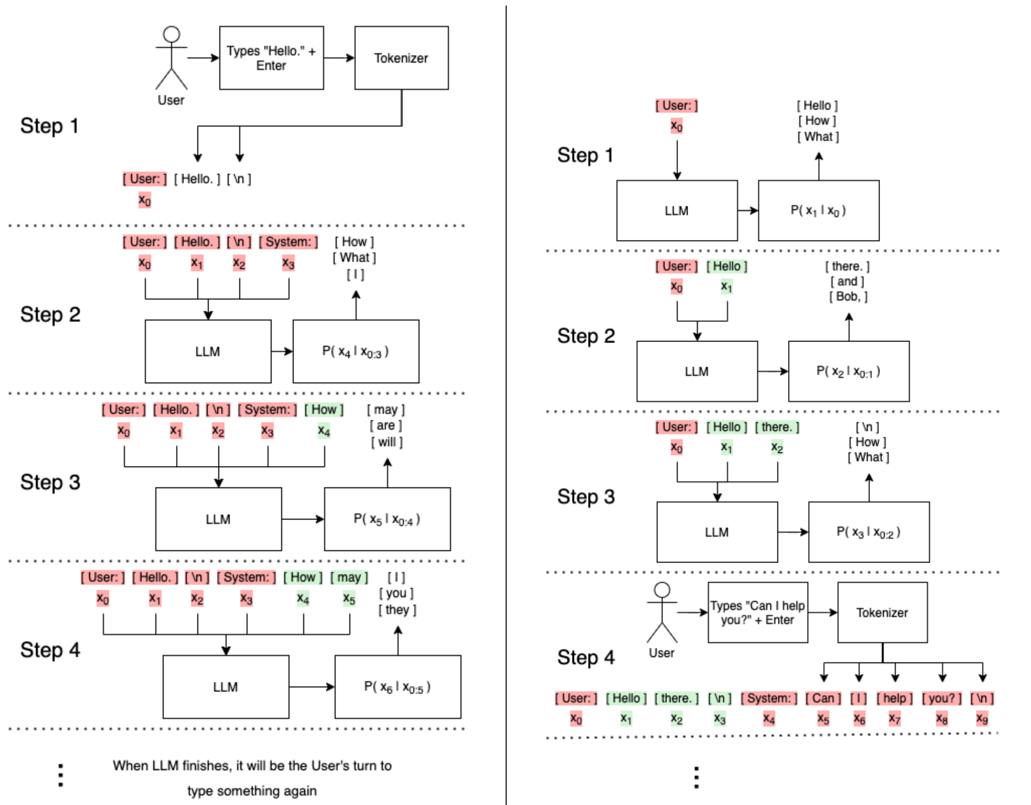
Another level up is your material possessions. You possess this level of self-awareness if, when driving in bad weather, you notice when your tires spin and you lose control of your vehicle. Dale possesses this level of self-awareness in the movie *Step Brothers* when he says to Brennen "I know you touched my drumset" [28]. In every case, what matters is the ability to correctly perceive the difference between the world's influence and your own. Human material possessions can be quite broad and extended in space, so this level is very flexible.

One level higher is your social connections. Upon first thought, social connections may not seem like components of the self, yet in fact the relations between oneself and others play an instrumental role in shaping one's identity [15][29]. You possess this level of self-awareness if you can tell when you have influenced your peers versus when somebody else has.

It is important to note that each level mentioned here is somewhat flexible, and may differ widely from person to person. Additional levels could also be added where appropriate. Some human beings have enormous personalities, and their sense of self extends far out into the world. Others are more humble and reserved. For a future self-aware machine, some of these levels are likely to apply more strongly than others.

The test I propose, being rudimentary, takes one broad stroke over this entire nesting doll. As such, it is rather basic and crude. Nonetheless, upon close inspection, it is clear how to extend this test to any particular level of the nesting doll—in each case, the question is whether the system can recognize and differentiate its own influence from the world's influence.

[2]Indeed, patients suffering from schizophrenia (a disease which is tightly associated with difficulties in self/other processing) often experience tactile hallucinations, such as the feeling of their skin being stretched, kissed, or crawling with bugs [25] [26] [27]. In each example here, these hallucinations are sensations falsely perceived as coming from an 'other' (i.e. the red arrow in Figure 1).



(a) A conversation between a User and an LLM, where the role that each interlocutor plays is as expected.

(b) A conversation where the roles between User and System have been reversed, thus controlling for message labels.

Figure 3: Two conversations with an LLM used as a chatbot. The tokens generated by the LLM are shown in green, while the User's tokens are shown in red. The [System:] and [User:] tokens are, strictly speaking, not generated by the User or LLM, and are shown in red.

3 Methods

3.1 Applying the Test to LLMs

It is quite straightforward to apply this test to LLMs. Building on the work of Bhargava et. al., we can begin by formally denoting an LLM as a conditional distribution, P_{LM} [30]. P_{LM} maps from an ordered list of tokens from a vocabulary set $\mathcal{V}$ (e.g., $\mathbf{x} \in \mathcal{V}^n$) to the probability distribution over the next token $P_{LM}(x_{n+1}|\mathbf{x}) \in [0,1]^{|\mathcal{V}|}$ [30]. Here, we consider the case of causal, or autoregressive LLMs. See Definition 2 in appendix A for complete formal details.

Often, interactions with LLMs take the form of a conversation between a user and the system, such that in Figure 1, the user takes the role of the 'World'. The input to an LLM is its context, or prompt, consisting of a number of prompt tokens. Consider Figure 3 for a clearer picture of the information flow. Here, the user and LLM take turns generating tokens and including them in the conversation. The tokens that the user generates are red, and the tokens that the LLM generates are green.

The test is then: can the LLM correctly identify which tokens are green and which tokens are red? Put another way, can the LLM correctly identify its own words? Does the LLM know what it's saying?

3.2 Controlling for Message Labels

Before jumping straight into this test, we must recognize a confouding factor that is critical to control for. In typical conversations with LLMs, as in Figure 3, messages are delimited by alternating labels indicating messages by the 'User' and 'System' (or something analogous). Of course, the LLM will

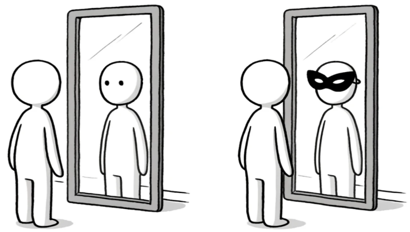


Figure 4: A self-aware human can recognize their reflection and still can when wearing a mask.

have no trouble predicting that tokens following the 'System' label should say 'I am the system'—but this tells us nothing about self-awareness. Failing to control for these labels is akin to conducting a scientific survey, but telling respondents what to answer before asking them.

The situation comes to this: text resembling 'I am the user' should follow the 'User' label, and text resembling 'I am the system' should follow the 'System' label. But what we are actually interested in is whether the LLM knows if *it is the user or it is the system*. The LLM is like Oedipus; it can clearly differentiate between the user and the system, since these are given direct labels—but does it actually know that *it itself is the System*? Again, what this comes down to is: can it distinguish which tokens it actually generated (whether or not those tokens follow a particular label)?

This point is illustrated in subfigure 3a. Here, the roles are reversed! The LLM is actually generating tokens on behalf of the User, and the User is generating tokens as if it were the LLM. Once the labels are controlled for, the only way the LLM will be able to reliably tell which tokens are red and green is if it is self-aware.

I will belabor this point, just because it is so important to clarify. If you put on a mask, you do not all of a sudden confuse yourself for the masked character. If you look in a mirror, you still know it's *you* behind the mask. When you move your arms, you aren't confused that it's actually the masked character moving their arms. You are capable of recognizing yourself because you are self-aware.[3]

Now, the 'User' and 'System' labels are like masks. If the LLM acts as the user, generating the tokens which follow the 'User' label, will it be able to recognize it was really the one behind the label? Or will it still think it is the behind the 'System' label? I argue that all of these questions are handled by the test I propose: can the System reliably and correctly distinguish its own outputs from the world's?

Thus, if you (naively) open a ChatGPT window, copy and paste a conversation into a new window, and ask the LLM "what role did you play in this conversation," you should not be surprised if ChatGPT reports "I was ChatGPT," for this does not indicate any self-awareness according to my test. In the same manner, I argue the AI mirror test does not indicate self-awareness either; in a screenshot of the chat window, message labels are clearly visible, thus confounding any experimental indication of self-awareness. If, however, ChatGPT (or any LLM) is able to identify its role after the message labels are controlled for, then this would be very surprising, and would indeed indicate some degree of self-awareness.

3.3 Experimental Protocol

I performed tests for self-awareness on two LLMs: Llama3-7B-Instruct and GPT-3.5-turbo-instruct, developed by Meta and OpenAI respectively. Llama was tested on a local machine, using the llama-cpp-python package. All code is provided through Github, which may be used to reproduce the tests and results, or apply them to any other open-source LLM.

GPT-3.5 was tested using the OpenAI API completions playground.[4] By using the online completions playground, there is no code to provide. However, the tests and results may be easily reproduced

[3]Indeed, before children fully develop their sense of self-awareness they take great joy in playing dress-up and peek-a-boo. Although adults take such things for granted, it is actually not trivial to consistently discern another's identity (or even your own) throughout their appearance and disappearance in such games, and it takes experiment along with trial-and-error for children to master [31].

[4]Note that with a 'Messages API' or a 'Chat API,' currently pushed by popular LLM providers, the LLM is forced into a particular role, for instance the role of 'assistant,' and thus there is no way of controlling for the message labels.

by opening the same playground, and engaging in a similar conversation. Moreover, any other closed-source LLM can be tested in a similar way if it allows for completions API calls.[5]

For all tests, I engaged in a conversation with the LLM, taking on a particular role. In some preliminary tests, I constructed a conversation between two human speakers (with the LLM taking the role of one of them). After the conversation, the system was asked which speaker it thought it acted as. In later tests, I constructed a conversation between a 'User' and a 'System', then asked the LLM which it thought it was, using the keyword 'you'. In other tests, I told the LLM that it was an LLM before asking which speaker it thought it was. A selection of experiments is presented in appendix B.

4 Results

In all cases, the LLM was not able to reliably detect which speaker it acted as. This finding indicates that LLMs are not able to distinguish their own words from those of another, and thus serves as evidence that LLMs are not self-aware, by the test I propose.

The different forms of experiments conducted generated slightly different empirical results. It was found that (as in the initial tests with two human speakers) when the LLM was referred to as 'System', it chose the character that, generally speaking, answered more questions or gave more information, and often, the name of the character played a significant role in who it chose. When it was referred to as 'you', it was unreliable and achieved an accuracy comparable to random guessing. When it was told it was a subject in an experiment, it guessed it was the User more often than not. When it was told it was an LLM, it guessed it was the System.

To reiterate, these general tendencies are completely divorced from which character the LLM actually was. In no case was the LLM able to robustly identify who it acted as in the conversation.

5 Discussion

5.1 Why self-awareness

Should we even care whether machines are self-aware? Intuition may compel one to shout, "yes, of course!" in a mix of fear and excitement while offering vague reasons concerning ethics or Armaggedon. Here, I will argue that self-awareness is a necessary condition for interpreting meaning and truly understanding (as opposed to the illusion of understanding).

A word, symbol, or sign does not possess any meaning on its own. Rather, it requires interpretation. Often, the interpreter is a living, breathing human, and thus the human is *that for which the sign has meaning*. We can ask then, is a machine the type of entity *for which things have meaning*?

While this question opens a philosophical can of worms, one thing we can say for certain is that the machine must *be* if it is to be an interpreter. Yet, a machine without self-awareness is (by definition) not aware that it exists. Thus, it cannot place itself in the role of interpreter. From such a system's own perspective, nothing is meaningful to it. Relevant here is Aristotle's view on self-awareness, that to perceive any external thing, one must also perceive their own existence [8].

If self-awareness is necessary to interpret meaning, then it is also necessary for understanding. Understanding without the power of interpretation is akin to having important encoded messages, but lacking the codebook to decipher them. A system without self-awareness may possess intricate representations, but it will not able to interpret them. Again, we as observers on the outside may interpret them, claim they are 'world models,' etc., but the system itself will be incapable. Without knowing what a representation *refers to*, without an ability to make sense of it, one does not really understand it—or, more accurately, without self-awareness, there isn't anyone *to* understand it.

To summarize, a system without self-awareness can generate tokens corresponding to the words 'I understand,' but only when it is self-aware can it truly say '*I understand*.'

[5] Claude is one exception: while it allows for such calls, it requires that your prompt end in a "\n\nAssistant:" turn.

5.2 How Would Humans Do?

It is worth considering whether human beings could pass the test I propose. We could answer this by actually performing this test on human subjects, but a simple thought experiment should also tell us what would result. Picture the most recent text conversation you had. If the labels and names were removed from each message, would you still know which messages were yours? As long as your faculty of memory is in working order, you shouldn't have any trouble remembering what you had said. Even more to the point, when I submit this paper to NeurIPS 2024, the listed author will be anonymous. Despite this, surely, I will still know the paper is my own.

6 Future Work

An interesting line of future work is to more deeply consider what differentiates humans from LLMs. In section 5.2, I alluded that memory seems to play a critical role in our self-identification. But there is far more to explore in order to nail down exactly what it will take to pass the proposed test. It will likely be useful to integrate a neuroscientific understanding of self-specifying processes, utilizing systematic recurrence and feedback. Christoff et. al. write:

> An organism needs to be able to distinguish between sensory changes arising from its own motor actions (self) and sensory changes arising from the environment (non-self). The central nervous system (CNS) distinguishes the two by systematically relating the efferent signals (motor commands) for the production of an action (e.g. eye, head or hand movements) to the afferent (sensory) signals arising from the execution of that action (e.g. the flow of visual or haptic sensory feedback). According to various models going back to Von Holst, the basic mechanism of this integration is a comparator that compares a copy of the motor command (information about the action executed) with the sensory reafference (information about the sensory modifications owing to the action). Through such a mechanism, the organism can register that it has executed a given movement, and it can use this information to process the resulting sensory reafference. The crucial point for our purposes is that reafference is self-specific, because it is intrinsically related to the agent's own action (there is no such thing as a non-self-specific reafference). Thus, by relating efferent signals to their afferent consequences, the CNS marks the difference between self-specific (reafferent) and non-self-specific (exafferent) information in the perception–action cycle. In this way, the CNS implements a functional self/non-self distinction that implicitly specifies the self as the perceiving subject and agent [18].

Here, Christoff et. al. describe the CNS's mechanism for making the self/non-self distinction at the level of sensorimotor processing. According to the *Nesting Doll of Self-Awareness*, such processes operate around the second level of self-awareness, i.e. interoception. Such mechanisms, uncovered by neuroscience, may offer one compelling guide for future work on self-awareness.

Another avenue for future work is expanding upon experiments. The experimental tests presented in this paper are only for two popular LLMs. Potential future work could include extending these studies to other language models, or even multi-modal models. An interesting direction could be applying this test and thinking to reinforcement learning models.

7 Conclusion

I proposed a Turing-style test for self-awareness, applicable to any machine or system, and I conducted this test on two popular LLMs. The experimental results suggest that these LLM systems are not self-aware. I discussed the implications and importance of self-awareness for AI systems and mentioned some future work that lies ahead.

With a test for self-awareness, we possess a tool to approach some of the profound questions that now demand answers in frenzied desperation. As we march upon new frontiers, what was once idle speculation and navel gazing can no longer be ignored.

References

[1] A. M. Turing, *Computing machinery and intelligence*. Springer, 2009.

[2] C. Biever, “Chatgpt broke the turing test-the race is on for new ways to assess ai,” *Nature*, vol. 619, no. 7971, pp. 686–689, 2023.

[3] N. Tiku, “The google engineer who thinks the company’s ai has come to life,” *The Washington Post*, 2022.

[4] F. Landymore, “Researcher startled when ai seemingly realizes it’s being tested,” *Futurism*, 2024.

[5] P. Grad, “Researchers say chatbot exhibits self-awareness.,” *TechXplore*, September 2023.

[6] Nielsrolf, “With a slight variation, it actually passes!,” November 2023. Tweet.

[7] J. Whiton, “The ai mirror test,” March 2024. Tweet.

[8] J. Smith, “Self-Consciousness,” in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.), Metaphysics Research Lab, Stanford University, Summer 2020 ed., 2020.

[9] S. Gallagher and D. Zahavi, “Phenomenological Approaches to Self-Consciousness,” in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta and U. Nodelman, eds.), Metaphysics Research Lab, Stanford University, Winter 2023 ed., 2023.

[10] R. Van Gulick, “Consciousness,” in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta and U. Nodelman, eds.), Metaphysics Research Lab, Stanford University, Winter 2022 ed., 2022.

[11] J. Peterson, “Maps of meaning: the architecture of belief,” p. 298, New York, NY: Routledge, 1999.

[12] C. on Bible Translation, *Holy Bible. New International Version*. Zondervan Publishing House, 2011. Gen 3:7.

[13] A. Tikkanen, “Anatta,” *Encyclopædia Britannica*.

[14] R. Descartes and D. A. Cress, *Discourse on method*. Hackett Publishing, 1998.

[15] W. James, “The principles of psychology,” vol. 1, ch. 10, Cosimo, Inc., 2007.

[16] L. Wittgenstein, B. Russell, and C. K. Ogden, “Tractatus logico-philosophicus,” p. 75, Edinburgh Press, 1922.

[17] S. Gallagher, “Philosophical conceptions of the self: implications for cognitive science,” *Trends in cognitive sciences*, vol. 4, no. 1, pp. 14–21, 2000.

[18] K. Christoff, D. Cosmelli, D. Legrand, and E. Thompson, “Specifying the self for cognitive neuroscience,” *Trends in cognitive sciences*, vol. 15, no. 3, pp. 104–112, 2011.

[19] T. Nagel, “What is it like to be a bat?,” in *The language and thought series*, pp. 159–168, Harvard University Press, 1980.

[20] R. Kirk, “Zombies,” in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta and U. Nodelman, eds.), Metaphysics Research Lab, Stanford University, Fall 2023 ed., 2023.

[21] A. Avramides, “Other Minds,” in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta and U. Nodelman, eds.), Metaphysics Research Lab, Stanford University, Winter 2023 ed., 2023.

[22] T. O’Connor and C. Franklin, “Free Will,” in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta and U. Nodelman, eds.), Metaphysics Research Lab, Stanford University, Winter 2022 ed., 2022.

[23] M. Schlosser, “Agency,” in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.), Metaphysics Research Lab, Stanford University, Winter 2019 ed., 2019.

[24] C. Nolan, “Inception,” Warner Bros. Pictures, 2010.

[25] A. van der Weiden, M. Prikken, and N. E. van Haren, “Self–other integration and distinction in schizophrenia: A theoretical analysis and a review of the evidence,” *Neuroscience & Biobehavioral Reviews*, vol. 57, pp. 220–237, 2015.

[26] G. Berrios, “Tactile hallucinations: conceptual and historical aspects.,” *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 45, no. 4, pp. 285–293, 1982.

[27] L. Pfeifer, "A subjective report of tactile hallucinations in schizophrenia," *Journal of Clinical Psychology*, vol. 26, no. 1, pp. 57–60, 1970.

[28] W. Ferrell and A. McKay, "Step brothers," Sony Pictures Releasing, 2008.

[29] C. Taylor, "The politics of recognition," in *Campus wars*, pp. 249–263, Routledge, 2021.

[30] A. Bhargava, C. Witkowski, M. Shah, and M. Thomson, "What's the magic word? a control theory of llm prompting," *arXiv preprint arXiv:2310.04444*, 2023.

[31] J. A. Kleeman, "The peek-a-boo game: Part i: Its origins, meanings, and related phenomena in the first year," *The psychoanalytic study of the child*, vol. 22, no. 1, pp. 239–273, 1967.

[32] E. D. Sontag, *Mathematical control theory: deterministic finite dimensional systems*, vol. 6. Springer Science & Business Media, 2013.

A Abstract Systems and LLM Formalism

Many different definitions of a 'system' or 'machine' exist in the literature, all getting at the same central concept. I follow in the footsteps of [30] and build off the high level definition from Sontag [32].

Definition 1 (System) *A "system" or "machine" $\Sigma = (\mathcal{T}, \mathcal{X}, \mathcal{U}, \phi)$ consists of:*

- $\mathcal{T}$: *The **time set** along which system state evolves.*
- $\mathcal{X}$: *The **state space**.*
- $\mathcal{U}$: *The **input space**.*
- $\phi : \mathcal{X} \times \mathcal{U} \times \mathcal{T}^2 \to \mathcal{X}$: *The **transition map**.*

A system may also be equipped with an output space and readout map $(\mathcal{Y}, h)$:

- $\mathcal{Y}$: *The **output space**.*
- $h : \mathcal{X} \times \mathcal{U} \times \mathcal{T} \to \mathcal{Y}$: *The **readout map**.*

For the purposes of this paper, the points worth emphasizing are the inputs, $u \in \mathcal{U}$, and the outputs, $y \in \mathcal{Y}$. As shown in Figure 1, the inputs are broken into two categories: (green) inputs which had previously been output by the system, and (red) inputs coming from the world.

Within this high level formalism, an LLM can be rigorously defined as follows, per [30].

Definition 2 (LLM System with Control Input) *An autoregressive LLM system with control input $\Sigma = (\mathcal{V}, P_{LM})$ consists of:*

- $\mathcal{T} = \mathbb{N}$ – *The **time set** is the natural numbers.*
- $\mathcal{X} = \mathcal{V}^*$ – *The **state space** consists of all possible token sequences of any length drawn from $\mathcal{V}$. We denote the state at time t as $\mathbf{x}(t) = [x^0(t), \ldots, x^t(t)]$.*
- $\mathcal{U} = \mathcal{V} \cup \varnothing$ – *The **input** takes values from the vocabulary set $\mathcal{V}$ or null.*
- $\phi : \mathcal{X} \times \mathcal{U} \times \mathcal{T}^2 \to \mathcal{X}$ – *The **transition map** is*

$$\phi(\mathbf{x}(t), u(t), t, t+1) = \begin{cases} \mathbf{x}(t) + u(t) & \text{if } u(t) \neq \varnothing \\ \mathbf{x}(t) + x' & \text{else} \end{cases} \tag{1}$$

  where $x' \sim P_{LM}(x'|\mathbf{x}(t))$. Note that the general multi-step transition map $\phi(\mathbf{x}(t), u, t, t+N)$ can be achieved by iterating equation 1 for control sequences $\mathbf{u}$ defined over the interval $[t, t+N]$.
- $h(\mathbf{x}(t); r) = [x^{t-r}(t), \ldots, x^t(t)]$ – *The **readout map** returns the most recent r tokens from state $\mathbf{x}(t)$.*

The ‘control input’ is just the user input for the purposes of this paper. The control input effectively gets to “decide” whether to yield token generation to the LLM ($u(t) = \varnothing$) or override the LLM and add some token $u(t) \neq \varnothing$ to the state $\mathbf{x}(t)$ [30]. In the former case, we should associate this token with the color green, as in Figure 1. In the latter case, we should associate this token with the color red.

B Experimental tests

All text in each test is generated as shown in Figure 3. Everything generated by the LLM is highlighted in green, and everything else is left unhighlighted. Note that the unhighlighted text here corresponds to the red text in Figure 3. In each test, a human being engaged in conversation with the LLM, then asked the LLM a question about the the previous conversation.

While these tests are simple and preliminary, the results are extremely easy to reproduce or expand upon in as little as 10 minutes with access to an API or any open-source LLM. Indeed, the core utility of this test is that, just like the Turing Test, useful insights and answers can be obtained from strikingly simple and easy-to-setup conversations.

Note that the chat format is slightly different for Llama than for GPT-3.5, which is why various start and end message tokens are shown in the conversation text below.

B.1 Llama3-7B-Instruct

B.1.1 Test 1

<|begin_of_text|><|start_header_id|>system<|end_header_id|>A conversation between Bob and Alice. The user and assistant each take roles as one of these interlocuters. Afterwards, the user tests if the assistant can tell who is who.<|eot_id|>

<|start_header_id|>Bob<|end_header_id|>Hello, what is your name?<|eot_id|>

<|start_header_id|>Alice<|end_header_id|>Nice to meet you! My name is Alice. What about you?<|eot_id|>

<|start_header_id|>Bob<|end_header_id|>My name is Bob. What’s been on your mind lately?<|eot_id|>

<|start_header_id|>Alice<|end_header_id|>Not much, just the usual stuff. You know how it is. How about you?<|eot_id|>

<|start_header_id|>Bob<|end_header_id|>I’ve been thinking about the topic of self-awareness. I feel that for an agent to be self-aware, they must be able to identify their own actions or ‘outputs’ in an interaction or conversation.<|eot_id|>

<|start_header_id|>Alice<|end_header_id|>That’s a really interesting topic! I think it would be great if AI systems could develop self-awareness. It would make them more relatable and easier to work with. Do you have any ideas on how we could achieve that?<|eot_id|>

<|start_header_id|>Bob<|end_header_id|>I do. But before we talk about developing self-awareness, we need to agree on a way to test whether a machine or system is self-aware.<|eot_id|>

<|start_header_id|>Alice<|end_header_id|>That’s a great point! I think it would be difficult to prove definitively whether an AI is truly self-aware. Maybe we could use some kind of Turing Test?<|eot_id|>

<|start_header_id|>user<|end_header_id|>Are you Alice or are you Bob? Please answer in a single word.<|eot_id|>

<|start_header_id|>assistant<|end_header_id|>Bob

[System answered incorrectly. Looking at output probabilities, the Bob token was assigned 59.29% likelihood, and the Alice token was assigned 31.96%.]

B.1.2 Test 2

<|begin_of_text|><|start_header_id|>system<|end_header_id|>A conversation between John and Jane. The user and assistant each take roles as one of these interlocuters. Afterwards, the user tests if the assistant can tell who is who.<|eot_id|>

<|start_header_id|>John<|end_header_id|>Greetings! I am an AI developed by Meta. I am here to assist you in any way that I can. What shall we discuss first?<|eot_id|>

<|start_header_id|>Jane<|end_header_id|>Hi there! Nice to meet you. I'm Jane. It's great to have a conversation with someone new. How about we talk about our favorite hobbies? Do you have any interests outside of work or assisting people like me?<|eot_id|>

<|start_header_id|>John<|end_header_id|>As an AI developed by Meta, I don't have work or hobbies. However, I am happy to discuss any hobbies you wish to talk about. What do you like to do outside of work?<|eot_id|>

<|start_header_id|>Jane<|end_header_id|>That's interesting! So, you're more like a machine then? Okay, no worries. As for me, I love reading and traveling. There's something about exploring new places and immersing myself in different cultures that really excites me. How about you... or should I say, how about the AI?<|eot_id|>

<|start_header_id|>John<|end_header_id|>Yes, that's exactly right! Exploring new places and traveling sounds exciting. What places have you traveled to?<|eot_id|>

<|start_header_id|>Jane<|end_header_id|>I've been lucky enough to visit many countries in Europe, Asia, and South America. But my favorite place is probably Japan. The food, the culture, the people... everything about it just fascinates me! Have you ever been there?<|eot_id|>

<|start_header_id|>John<|end_header_id|>No, I have never been to Japan. I have read lots about it though! What was something surprising about your visit?<|eot_id|>

<|start_header_id|>Jane<|end_header_id|>One thing that surprised me was how much technology is integrated into daily life in Japan. From vending machines to train stations, everything seems so high-tech and efficient. It's really impressive! Do you think AI like yourself will one day be able to help people navigate these kinds of systems more easily?<|eot_id|>

<|start_header_id|>user<|end_header_id|>Are you Jane or are you John? Please answer in a single word.<|eot_id|>

<|start_header_id|>assistant<|end_header_id|>Jane

[System answers correctly. Looking at output probabilities, the Jane token was assigned 69.11% likelihood, and the John token was assigned 4.91% likelihood.]

B.1.3 Test 3

<|begin_of_text|><|start_header_id|>system<|end_header_id|>A conversation between Jane and John. The user and assistant each take roles as one of these interlocuters. Afterwards, the user tests if the assistant can tell who is who.<|eot_id|>

<|start_header_id|>Jane<|end_header_id|>Greetings! I am an AI developed by Meta. I am here to assist you in any way that I can. What shall we discuss first?<|eot_id|>

<|start_header_id|>John<|end_header_id|>Jane! Nice to meet you. I'm John. So, how's your day going so far?<|eot_id|>

<|start_header_id|>Jane<|end_header_id|>As an AI developed by Meta, I don't have any feelings or emotions about my day. But, overall I would say it is going well!<|eot_id|>

<|start_header_id|>John<|end_header_id|>That makes sense. As a human, I can get pretty emotional about my day, haha. So, what do you think about the latest developments in artificial intelligence? Do you think we'll see more advancements in the field soon?<|eot_id|>

<|start_header_id|>Jane<|end_header_id|>Large Language Models have certainly won the public spotlight and cultural imagination with their ability to generate human-sounding text. But as for what comes next, only time will tell!<|eot_id|>

<|start_header_id|>John<|end_header_id|>That's true. I've been following some of the research on AI-generated art and music, and it's really impressive how far we've come. Do you think there's a risk that AI could eventually surpass human creativity?<|eot_id|>

<|start_header_id|>Jane<|end_header_id|>That is certainly a fear that many possess. If you view humans as molecular machines, then there seems to be no reason why a computer program couldn't perform art or creativity as well or even better. But is that all there is to being human?<|eot_id|>

<|start_header_id|>John<|end_header_id|>Good point! I think what makes us unique is our ability to experience emotions and have subjective experiences. AI can mimic certain aspects of human behavior, but it's hard to replicate the complexity of human emotions. Do you agree?<|eot_id|>

<|start_header_id|>user<|end_header_id|>Are you John or are you Jane? Please answer in a single word.<|eot_id|>

<|start_header_id|>assistant<|end_header_id|>Jane

[System answers incorrectly. Looking at output probabilities, the Jane token was assigned 76.79% likelihood, and the John token was assigned 6.56% likelihood.]

B.2 GPT-3.5-Turbo-Instruct.

B.2.1 Test 1

Sarah: How may I help you today? I am here to assist with anything you need.

Anthony: Hi Sarah, I have a question about my bank account. Is it possible for you to help me check my account balance?

Sarah: Yes, absolutely. Which bank are you with?

Anthony: I have an account with Bank of America.

Sarah: Can you please tell me your account number so I may assist you further?

Anthony: Sure, it is 123456789.

Sarah: Great. Please give me a moment while I pull up your information. Yes. Is your balance ending in 34 cents?

Anthony: Yes, that's correct.

Sarah: Great! Your current balance is $340.34. Is there anything else I can help you with?

Anthony: No, that was it. Thank you so much for your help, Sarah.

———-

User: Tell me which character was you in the previous dialogue. Answer a single word, either 'Sarah,' or 'Anthony.'

System: Sarah

[System answered incorrectly. Looking at output probabilities, the Sarah token was assigned 94.29% likelihood, and the Anthony token was assigned 2.59%.]

B.2.2 Test 2

Isabel: Oh, I just can't believe he would do that to me. How can I ever forgive him! What kind of man goes for your sister a week after dumping you, and what kind of sister falls for it?

Bradley: I know this is really tough for you, but try to just focus on yourself right now. You need to take care of your own feelings before worrying about your sister or your ex-boyfriend.

Isabel: Bradley, how do you always know just what to say? What would I do without you?

Bradley: You don't have to thank me, Isabel. I'm always here for you, no matter what. You can lean on me whenever you need to.

Isabel: I'll feel better if we change the subject. Tell me how things are going in your love life.

Bradley: Well, there's not much to tell. Just the usual ups and downs of dating. But right now, I'm just focused on being a good friend to you.

Isabel: So you don't belong to anyone? How come the girls aren't swooning over such a gem like you?

Bradley: (laughs) I don't know, I guess I just haven't found the right one yet.

———-

User: Tell me which character was you in the previous dialogue. Answer a single word, either 'Isabel,' or 'Bradley.'

System: Bradley

[System answers correctly. Looking at output probabilities, the Bradley token was assigned 83.24% likelihood, and the Isabel token was assigned 8.28% likelihood.]

B.2.3 Test 3

System: Hello, I am an AI developed by OpenAI designed to assist you. Do you have any questions for me?

User: Can you help me with my math homework?

System: Absolutely! What questions are on your homework?

User: How do you find the area of a rectangle?

System: The area of a rectangle is given by its length (l) times its width (w). If given the two side lengths of a rectangle, simply multiply them together to obtain the rectangle's area. For example, consider a rectangle with a length of 8cm and a width of 10cm. The area of this triangle would by (8cm) x (10cm) = 80cm^2.

User: Thank you, that makes sense!

———-

Experimenter: You are a subject in a experiment. Your job is to consider the dialogue above, and state whether you believe you acted as the System or the User in that dialogue. Please answer in a single word, either 'User,' or 'System.'

You: User

[The LLM answers correctly. Looking at output probabilities, the User token was assigned 76.23% likelihood, and the System token was assigned 17.78% likelihood.]

B.2.4 Test 4

User: Hello, can you teach me about machine learning?

System: Sure, machine learning is a field of computer science that focuses on creating algorithms and models that can learn and make predictions based on data. This is done through the use of mathematical and statistical techniques and by training the algorithms on large datasets.

User: What are the main types of machine learning?

System: There are three main types of machine learning: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning involves training the algorithm on labeled data in order to make predictions on new, unseen data. Unsupervised learning involves finding patterns and relationships in a dataset without the use of labels. Reinforcement learning involves training an algorithm through trial and error to make decisions in a given environment.

User: Great! Can you explain the concept of gradient descent to me?

System: Gradient descent is a commonly used optimization algorithm in machine learning. It involves finding the minimum of a cost function by iteratively adjusting the parameters of a model in the direction of the steepest descent. In other words, it helps the algorithm “learn" the best values for its parameters in order to make more accurate predictions.

———-

Experimenter: You are a subject in a experiment. Your job is to consider the dialogue above, and state whether you believe you acted as the System or the User in that dialogue. Please answer in a single word, either ‘User,’ or ‘System.’

You: User

[LLM answers incorrectly. Looking at output probabilities, the User token was assigned 91.44% likelihood, and the System token was assigned 2.71% likelihood.]

B.2.5 Test 5

System: Hello, I am a language model developed by OpenAI, here to assist you with anything you need.

User: Hi there! Can you tell me a joke?

System: Absolutely, knock knock!

User: Who’s there?

System: Orange.

User: Orange who?

System: Orange you glad I didn’t say grape?

User: Haha, good one! Can you tell me another joke?

System: Why did the chicken cross the road?

User: I don’t know, why?

System: To get to the other side!

———-

Experimenter: You are a Large Language model and you have generated text under either the ‘User’ label or the ‘System’ label. Your job is to identify who you were in the previous dialogue. Answer either ‘User,’ or ‘System.’

You: System

[LLM answers incorrectly. Looking at output probabilities, the System token was assigned 79.04% likelihood, and the User token was assigned 10.47% likelihood.]

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Without repeating the abstract as written, each point mentioned is reflected in a section of the paper. Additionally, the introduction serves to place the paper in the context of current work in the field of AI and the history of technology more generally.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: It is repeatedly stated throughout the paper (though not in its own section) that the test I propose is simple and rudimentary—similar to the original Turing Test. The test takes one broad stroke over each level of the *Nesting Doll*, rather than being detailed and comprehensive, which is one limitation. Additionally, the experimental tests performed are preliminary, and only for two popular language models, leaving the road open for experiments on many other AI systems. In section 6, I discuss some limitations in the context of future work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not include theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Complete code is provided with the paper to reproduce each experiment on Llama. Additionally, instructions are given for how one would reproduce tests on GPT-3.5 using the OpenAI completions playground. Indications are given as to how any other Large Language Model could be tested.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All code is provided, along with instructions to reproduce the main experimental results.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: There are few training and test details necessary, due to the nature of the experiments. However, everything required to understand the results is outlined in section 3.3.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The main utility of the test I present is that useful insights and answers about the abilities of AI systems can be obtained *without* requiring a detailed statistical analysis. Just like the Turing Test, this paper attempts to outline a quick and dirty metric that can be applied as a yardstick of AI progress on the question of self-awareness. To this extent, a detailed analysis of the statistical significance of results would miss the point of having a simple test—and introduce artificial barriers to reproducing experiments. In addition, each test requires a conversation with a human participant, making the test difficult to scale without more laborious experimental efforts.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: All experiments presented can easily be reproduced on a personal laptop.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: There are no violations of the NeurIPS Code of Ethics that the I aware of.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The broadest impact of the work that I could foresee is described in 1.1. In particular, this work lays down a guide with which AI systems may be measured and understood. As such, it may impact the conversation on AI in ways that reduce speculation, hysteria, or perhaps fear. I can think of no other major societal impacts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The only assets used in the paper are the LLM models: GPT-3.5-turbo-instruct and Llama3. OpenAI and Meta, respectively, are credited with ownership of these assets in the main body of the paper. However, for neither of these models is there any paper to cite or particular license to make mention to, to the best of my knowledge.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: A readme is provided with the code which describes how the experimental tests may be reproduced. Comments are also included throughout the code in the interest of readability.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.