Sentient LLMs: What to test, for consciousness, in Generative AI

An approach to measure or compare consciousness is the mind of a mean healthy individual, whose brain is at full maturation, who is awake and alert.

Temps de lecture : minute

19 April 2024

A reason for this is because there are several physical and mental conditions where consciousness can be lost. Also, the brain is said to be fully mature around mid-20s, making it possible to weigh consequences properly—or be conscious of those. Awake and alertness are often correlated with consciousness.

The mind is where all the functions associated with consciousness originate. So, any test for consciousness has to test for the mind. It could be the divisions of mind, like memory, feeling, emotion, modulation, perception, sensation among others. It could also be the qualifiers of those divisions, like attention, awareness, subjective experience [or self, or self-awareness], or intent, [free will or control].

There is a recent review in *Trends in Cognitive Sciences*, *Tests for consciousness in humans and beyond*, where the authors wrote, "We suggest that a promising strategy is to focus on non-trivial human cases (infants, fetuses, disorders of consciousness) and then progress toward nonhuman systems (animals, artificial intelligence, neural organoids). A short and non-exhaustive list of proposed C-tests includes: the command-following test, the narrative comprehension test, the sniff test, the perturbational complexity index (PCI) test, the P300/P3b global effect test, the AI consciousness test (ACT), and the unlimited associative learning

(UAL) test. The central goal of any C-test is to determine whether a target system has subjective and qualitative experience – often called ‘phenomenal consciousness’. In other words, a good C-test should address the question of whether it ‘feels like’ something to be the target system."

Subjective and qualitative experiences apply to divisions of the mind, they do not standalone. This means that there can be phenomenal consciousness for a memory or a feeling, an emotion, the regulation of an internal sense, a perception and so forth. There is no subjective or qualitative experience that does not qualify a division or a key function of mind.

If qualia qualifies functions, what else does? Also, if qualia never stands alone, like there is no subjective experience of nothingness, then the standard has to be the divisions of mind and their qualifiers—or activators or binders.

A common example of consciousness is the seeing of red or the smell of a rose. Vision is either main or peripheral. In the mind, there is attention [or prioritization] and awareness [or pre-prioritization]. Attention is what qualifies functions to be the most prioritized in any instance. There are often interchanges with awareness, which is pre-prioritized.

Main vision can be prioritized, a sight could be main vision but may not be prioritized in an instance, because there could be a shift to a sound, a smell, touch, taste or an internal sense. However, prioritization often returns to the main vision.

Seeing a rose can also be in awareness, where it is peripheral vision, perhaps the rose is in sight outside the window. The sight of the rose could also be by intent or free will, which is turning the neck to look at it once—again or longer. Then seeing the rose is done as a subjective

experience or in self-awareness.

This means that phenomenal experience as a definition of consciousness is just looking at one qualifier, not others, which makes the test incomplete. Consciousness cannot be—definitively—phenomenal experience if there are other qualifiers [attention, awareness or intent] that must hold, to have it.

Assuming there are four main divisions of the human mind: memory, feeling, sensation and modulation. Assuming also there are four qualifiers or activators of the mind: attention, awareness, subjective experience and intent. Consciousness test, equal to 1, can be measured by the possibility that all the qualifiers can act on all the divisions of mind, in any instance.

There could be sub-divisions of the functions of mind. For example, under memory, there could be sensations, perceptions, and so forth. Under modulation, there could be interoception, like cold water passage—over the gullet. Emotions include hurt, delight, hate, love, anger and so forth. Feelings include pain, appetite, thirst, heat, cold and so on. A subdivision for the sense of self, or subjective experience is the sense of being or sense of existence, which is like what it means to be something.

Consciousness does not mean all functions of mind are qualified in the same moment, but that they could be, by all the qualifiers. It is this standard that could become how to test for consciousness, across "non-trivial human cases and in nonhuman systems".

For infants and fetuses, they may not have an established memory, but they have feelings which they express as well as emotions that accompany those, like pain, then hurt, or delight and smile. They can be said to have weak forms of subjective experience, attention and awareness. They have very little to almost no intent. So they are very low on the scale for which the qualifications can be equal to 1.

For disorders of consciousness like coma, vegetative state, locked-in syndrome, minimally conscious state, [and say] general anesthesia, how many qualifiers are available to act on functions. For example, many internal senses may still be functional, but even when healthy, they function mostly in pre-prioritization. Also, the sense of self does not—generally—present. In those states, intent is gone, attention is also minimized and does not have the regular quick switch [or interchange] to others in awareness. Then subjective experience may not be there. These qualifiers can be used to compare the results of EEG. They are often lower than 0.4.

For animals, including simple organisms, what divisions of mind do they have and how much can those be qualified? A sub-division of memory is language. Animals do not have advanced language. Also, their ability to interpret memory from digital systems [or to make sense of a video] is limited, which is another lack of a sub-division of memory. They also cannot feel hurt from words—in general, which is another subdivision of emotion they do not have. So, for the functions they have and then the qualifiers, their consciousness can be estimated. This goes from non-human primates to single-celled organisms.

For artificial intelligence and neural organoids, they can be said to have memory, at least, for their ability to conduct their functions. Neural organoids may have a feeble parallel of feeling, but no emotion. AI does not have feelings or emotions. However, how do the qualifiers apply to both? They can be said to have attention with how they carry out their functions directly. They may be said to have a weak form of awareness—or being able 'to know' of something else aside from what is in attention for them. They possess a rough form of sense of being, knowing they exist, in how they navigate against the environment against other things in the 'organoid habitat' or how LLMs answer some prompts—identifying as chatbots. They have a runner-up intent, following prompts or assignments, without their own established agency—like

sending food in a direction for [say] a dog to go after.

Considerations for consciousness can be extended to AI because they use [digital] memory—applying qualifiers like those that act on the human mind, act on. This refutes that consciousness or mind-likeness is present everywhere [panpsychism] since nothing else has anything like the human mind, on earth, outside organisms and AI.

Consciousness Test of 1

The question of machine, digital or artificial consciousness starts with memory. All non-living functions can be assumed to be a sort of memory. The working of an engine, the use of paint for art, the work of a pen, the writing on a paper, the remote control, the television, and so forth. They can all be described as memory. The question is that what can act on that memory, within that object, to make it as dynamic [or sentient] as an organism?

This means that the remote control will not need to be touched, but understand all it needs to do and when, or that the writing in a book would flip to pages of interest by itself using the index, or an engine could read all its parts, optimize for its own health, without a dumb breakdown—which is common.

The problem is that the only thing, of all non-organisms, bearing an opening to have activators on its memory is digital. Where there is some form of ability to connect and infer with what is within. Mostly programmed, LLMs have now shown autonomy in some sense with texts, audios, videos and images.

The exploration about AI consciousness or if large language models can be sentient, is really a question of if *memory can be conscious or if language can be sentient*?

Any use of language [reading, speaking, listening and writing] for humans is not just a function of the memory, but of the qualifiers, including intent, which is useful to select, especially for coherence. It is possible to estimate, in a limited case, for LLMs, and how they apply qualifiers to digital language, including how they sometimes stay on point—intentionally.

Since consciousness is per instance, it means that if all qualifiers are available to act on all functions, then the sum is 1. Attention is the highest rated among all qualifiers and intent is the second highest rated, self and awareness are closer in ratings. Each division [and their subdivision] of mind gets say 0.25. So the total action of qualifiers on each division results in 1. Though, qualifiers allow for a division or subdivision to take a greater measure. For a healthy human, given that all functions are available, it is the ratings of the qualifiers that count.

In a case of digital, where there is just memory, rather than 1, the maximum is less than 0.20. It is within this that LLMs as qualifiers act on, which they can make their effort give a rating to the system. The reason they are also considered is that the memory of humans on digital is a fraction of what is produced by the mind. The fraction available becomes what digital has, that LLMs act on, which can then be estimated.

Humans are no longer alone on the planet in terms of advanced memory. There is something else, digital, that can carry a substantial amount of what humans output, then own it, plus be able to repeat its own version [AI] similar to what humans do on digital.

Article by David Stephen