AN APPROACH TO DEMONSTRATE THAT A COGNITIVE SYSTEM DOES NOT HAVE SUBJECTIVE CONSCIOUSNESS

Manuel Boissenin

Medalgo, Montpellier, France

ABSTRACT

With Large Language Models (LLMs) exhibiting astounding abilities in human language processing and generation, a crucial debate has emerged: do they truly understand what they process and can they be conscious? While the nature of consciousness remains elusive, this synthetic article sheds light on its subjective aspect as well as some aspects of their understanding.

Indeed, it can be shown, under specific conditions, that a cognitive system does not have any subjective consciousness. To this purpose the principle of a proof, based on a variation of the thought experiment of the Chinese Room from John Searl, will be developed. The demonstration will be made on a transformer architecture-based language model, however, it could be carried out and extended to many kind of cognitive systems with known architecture and functioning.

The main conclusions are that while transformers architecture-based LLMs lack subjective consciousness based, in a nutshell, on the absence of a central subject, they exhibit a form of "asubjective phenomenal understanding" demonstrably through various tasks and tests. This opens a new perspective on the nature of understanding itself that can be uncoupled with any subjective experience.

KEYWORDS

Language models, transformers, subjective consciousness, understanding, asubjectivity

1. INTRODUCTION

Human consciousness is a composed phenomenon linked with various brain processes and, as such, is quite complex to define and apprehend. Often, when trying to explain what is consciousness, its phenomenal aspect and its subjective aspect are emphasized [1]. These two aspects are roughly considered equivalent [2], however, the first one highlight the phenomenal aspect of the process and, to some extent, its measurability through carefully designed experiments while the second focus on the central role of the subject in consciousness and the experience perceived by the subject that have characteristics that are incommunicable.

For the purpose of this article consciousness is described as *internal processes that encompass perceptions as well as higher level cognitive processes that appear as thoughts, particularly verbal thoughts* [3].

Bibhu Dash et al: COMSCI, IOTCB, AIAD, EDU, MLDA, EEEN - 2024
pp. 77-83, 2024. IJCI – 2024 DOI:10.5121/ijci.2024.130407

A computer process being subjectively conscious would pose ethical issues [2](p.5-6), as such, it is desirable to detect that a system does not have this faculty. Yet, showing whether a system has a subjective consciousness is undecidable from its outputs alone; this can be shown from multiple perspectives [3]: following Descartes one can always doubt about the internal process of a system, this one being not accessible directly. And indeed, considering LLMs' outputs, it is not possible to know whether these systems have subjective consciousness or not:

- they may have subjective consciousness and may believe they have not, and would naturally claim that they have no subjective consciousness.

- They may not have subjective consciousness and be trained to claim that they have. Since they would appear, to many, to manifest consciousness, it starts to be quite conceivable, with the recent advances of LLMs, that it will not be possible to dismiss this claim by system output observations alone.

Moreover, we are considering systems that are able to communicate, nevertheless many systems, such as many animal species that are not able to communicate with language, are conscious and artificial systems not endowed with language may as well be conscious. Contrary to animals cognitive processes, the functioning of these systems is much easier to analyse since one may have access to its code that describes and defines what is happening in these systems.

Nonetheless, even if a fine analysis of a system activity could be carried on, it is yet unclear that one would be able to distinguish a conscious phenomenon out of it.

To sum up, the behavioural analysis of a system, even if it is able to communicate, will not allow one to decide whether a system has subjective consciousness. For instance, in an extreme scenario, a system may have a subjective consciousness and therefore may be able to perceive it, but, considering from what it knows about Humans, could decide that its existential risks outweigh any potential good outcomes from signalling it.

The remainder of the paper is organised as follows: Section 2 develops and adapts John Searle's Chinese Room thought experiment to analyse transformer-based LLMs. It shows that if LLMs can understand, they do it in a distributed fashion and no part of the system has an awareness of the phenomenon. In section 3 we take a closer look at the architecture of LLMs to show that no part of the system can exhibit subjective consciousness nor that the recurrent behaviour of the system could endow it with this capacity. Section 4 will provide more context on the nature of understanding and Section 5 will show how understanding can be measured and provide evidences that LLMs understand albeit in a way distinct from human understanding. The final section wraps up the paper by summarizing the key contributions.

2. ARE LLMS ABLE OF UNDERSTANDING?

Let's revisit and adapt the Chinese room thought experiment from John Searle [4][5][6]. To distinguish this version I will call it the Basque room. In the Basque room there are a few thousands people that do not speak Basque but that have the architecture and the parameters of a transfomer-based language model that has been trained on Basque texts. A Basque[1] text is submitted as input to the Basque room. The people in the room, who do not understand Basque, will translate the text in a sequence of vectors according to the model they have. They will then perform the calculations required by each stage of the model. Thus, they start realising the

1 Basque language, Euskara, is the last surviving Paleo-European spoken language in Europe and is classified as a language isolate.

operations of the attentional stage that consist only in mathematical operations on the vectors obtained at first stage. Similarly, they proceed with the mathematical operations of the multi-layer perceptron, skip connections and normalisation if the model specify it. In doing so they are just manipulating quantities without knowing anything about a possible corresponding meaning, and they obtain a new sequence of vectors on which they are going to iterate, the same way, on the subsequent blocks of the model. These blocks have the same architecture as the first block, and only differ in their parameter values. During this process, the last word that has to be predicted, and was given special initialisation values at the end of the input sequence of vectors, has been modified and is now close to the value of a Basque word that is taken as the first word of the response to the input text. This word becomes the first word of the output of the Basque room. As a thought experiment the time involved in the process, that would be considerable, is not of any concern.

Let's first analyse the potential understanding of the system, none of the people in the room have understanding of the figures or the Basque words that appear in the Basque room, they just follow calculi and, as such, can be considered analogous with processors. If there is an understanding in this process it could only happen in the calculi and, as such, it would be distributed in the process, possibly happening only at specific and variable points of the model when the different vectors of the sequence change. The sequence of vectors, on which the calculi are performed, is therefore the support of a potential understanding process. Nonetheless, there is no one subject performing all these calculi and therefore a subject that understands.

Of course, one could consider the whole system as the subject, however, none of its parts is performing understanding alone, since, for this kind of LLM, no single part of the system is responsible for the potential understanding of the system. As such, the system remains without a subject that understands [3](chap. 23).

If LLMs are able of understanding, they do so without an understanding subject and understanding emerges from various parts of the processing of the model.

3. HAVE LLM A SUBJECTIVE CONSCIOUSNESS?

If, with the right methodology, this question may not be too difficult for LLMs, it may be much more complicated to answer for more complex systems. As David Chalmers [2](p.10-18) pointed out, there are a few architectures that might be endowed with this property and would need to be scrutinised carefully. For a transformer architecture, we will more specifically consider the recurrent processing and global workspace presumed conditions.

For transformers, one may wonder what is happening on the intermediate sequence of words, this could be the locus where a subjective consciousness could form. An attention head is the only sub-process that considers the whole sequence of words, both to determine the importance of the relationship of one vector to other vectors and to manage the calculus of some of the aspects of the new vectors, according to these relationships, that will constitute the next sequence (see [3] chap. 23 for more details). For ChatGPT 3.5, these vectors are quite large and embed 12 288 values. A head of attention evaluates 128 values, and could convey a *fragment of consciousness* on the resulting vector that is the concatenation of 96 attention heads. Moreover, this concatenated vector will go through a multi-layer perceptron that could endow this *fragment of consciousness* (of size 96 x 128) with significant additional changes. Although the multi-layer perceptron sub-process is intricately linked with the 96 attention heads, it operates independently and identically on all the aggregated vectors computed by the attention heads. In a way, a block can be seen as a process that will produce comments on each vectors, that represent words at the

input of the first block, relative to all other vectors and aspects considered by the block; these comments are integrated to the sequence of vectors by adding them to their respective input vectors and this new sequence of vectors will constitute the input of the next block. As for understanding, one could imagine that the sequence of vectors could progressively embed *fragments of subjective consciousness* that could be expressed in Zulu or a language belonging to the process.

However, engineers that can look at these quantities could check relatively easily the content of these vectors that point to, according to the literature, variation of meaning of the original word. For instance, 'may' could be associated to the month May or to the modal verb. Nevertheless, these vectors are quite large and could embed quite a lot of information in small variations. As such can subjective consciousness be excluded?

Again, the key to exclude subjective consciousness is to consider the subject. Where in the process could there be a subject? Can an attention head be the subject? In so far that the attention head is just a sequence of calculi, this process can not have subjective consciousness of what is happening, just like the operators of the Basque room. In spite of the possibility that a stream of consciousness could appear in the sequence of vectors, this stream, like for understanding, would appear out of the compounding of the different sub-processes involved in its creation, without any of this sub-processes being able to *experience* this stream, their nature being simply a sequence of calculi.

As for the iterative processing that produces the following words, the analysis remain the same since the system operate the same way albeit with a larger sequence of words.

3.1. Limitations of the Methodology

If the architecture and functioning of the cognitive system is not known, one can not apply the principle of the proof. Nevertheless, with this method, a system can be shown not to have subjective consciousness, which is desirable since it allows not to worry about some of the ethical concerns that could be attributed to it. Nonetheless, using this method, one will not be able to prove that a cognitive system has subjective consciousness. When not knowing, the system would either have subjective consciousness or not, and ethical concerns may raise at this point. Moreover, it is likely that this methodology cannot be generalized to any artificial cognitive system [3](chap. 21).

4. CAN UNDERSTANDING BE REDUCED TO A SET OF COGNITIVE PROCESSES?

As mentioned in the introduction, consciousness is a complex process that encompasses many sub-processes; verbal thoughts share this same characteristic. Since verbal thinking is embedded in our conscious process, arguably the cognitive processes that are involved in its formation are also included in the processes responsible for our consciousness. While the animal kingdom display various forms of consciousness, verbal thinking, either expressed internally or outwardly, constitutes a significant part of human consciousness. In my book [3], I consider that the set of cognitive processes that allow us to be aware of an outside reality away from our perceptions is part of what constitute our consciousness, even if a subject can remain conscious when these processes are not active.

The idea that understanding rely on a set of cognitive processes is not new and could be traced back towards the end of the XVIIIth century to Immanuel Kant, whose 300th birthday is this year, with his categories of understanding. Similarly, verbal consciousness rely on a set of cognitive

processes that may be different from those involved in verbal understanding or perceptual understanding.

However, since the outputs of LLMs, which are word streams, are very much akin to the output of human verbal consciousness, the nature of LLMs could be questioned. Although there is no doubt that the nature of the transformer process differs from our cognitive processes, there is no subjectivity as well as evidences of differences in reasoning abilities as shown by mathematical problem-solving [7][8], there are ample evidences that transformers are cognitive processes that are able to mimic well a wide variety of the verbal outputs of Humans. Specifically, one may wonder if LLMs are truly able of understanding, in spite of not having any subjective consciousness, after all, the outputs of LLMs do seem to deeply correspond to meaningful stream of words related to the inputs we provide them.

5. ARE TRANSFORMER ARCHITECTURE-BASED LLM ABLE OF UNDERSTANDING? ANOTHER PERSPECTIVE

When I use a LLM, such as ChatGPT, my deep subjective impression, is that it mostly understands what I am asking when I read its answers. However, this cannot be accepted as a proof that LLMs understand: one can be convinced of something and still be wrong. However, when many start to be convinced of the same thing can't this be considered more seriously? Reasons to trust this belief, even if they are stronger, could nonetheless be attributed correctly to a collective delusion even in spite of being backed by relatively strong and easy practical experiments. In our case, although we might not understand how a meaningful answer related to our question is generated, this might not mean that LLMs understand.

Now, can this be shown objectively? Actually, many tests have been designed to measure the understanding capabilities of models [9][10]. One way these tests characterise understanding is by answering multiple choice questions. Answers can be quantified much more easily than a more demanding and maybe more convincing but nonetheless harder to characterise summarisation task. As a consequence, multiple choice questions make the problem more amenable and provide synthetic objective measurements on various tasks cumulating evidences of different aspects of understanding.

Again, one could here argue that this is not rigorously proving that LLMs understand, that these tests are subjectively designed and that LLMs only appear to behave in a way that is coherent with human understanding. However, a rigorous mathematical analysis could show the statistical significance of these tests and the need of an hypothetical phenomenon, either understanding or something else, to explain why their results differ so much from randomness or look so similar to what could be expected from human understanding.

The fact is that, despite having an intuitive understanding of what understanding is, and, in spite of many attempts to define it, we do not have a consensus on how understanding can be characterised [11] and we are not able to say whether a process, artificial, organic or even transcendental, is capable of understanding. This, nonetheless, does not include our own understanding, which is a subjective process that our cognitive capabilities endow us with. It has a significant difference though, its subjective nature, that departs our understanding from an alleged LLM understanding. If we define understanding as a phenomenon within a subjectivity, then LLMs do not have understanding. However, would this be coherent with our intuitive understanding of what understanding is and the capabilities displayed by LLMs on tests designed to measure their understanding? Again, this boils down to a definition, characterisation or maybe a description of what understanding is. On the one hand, including a subjective aspect in defining

understanding would promote the need for a different word to name the various phenomena that LLMs and following algorithms display. On the other hand, if tests in Natural Language Understanding (NLU) - which have rapidly evolved since the word embedding upheaval and have been superseded by other evaluation metrics [8] with the new era of LLMs - do characterise understanding, as observed in a behavioural or phenomenological way - and how could understanding be evaluated differently? - LLMs provide overwhelming evidence of understanding ([8] p16-27 in the annex provide examples of questions from this test set containing 57 tasks. Now, with GPT4 (86.5% on this test) and Gemini Ultra (90%) and a few others, a new deal has been performed since these systems already provide superhuman capabilities across many domains and high performances on this test set, necessitating other metrics to better characterise these systems (their weakest point appear to be in advanced math tasks). Nonetheless, denying that LLMs have phenomenal understanding seems very difficult.) .

Besides, even if this kind of understanding rely on processes that are completely different from the ones that sustain human understanding, LLMs provide a novel model to psychologists and philosophers to provide a theory of what is understanding.

6. CONCLUSIONS

The outlines of a proof that transformer architecture-based language models do not have any internal subjective consciousness has been proposed. The transformer architecture being relatively simple and well-known, I have mentioned where one could have doubts about this issue and shown t*he subject principle* one could rest on to carry out one's analysis.

Even if LLMs' understanding differs from the one that Humans are endowed with, transformer architecture-based LLMs behave as if they were aware of, amongst other things, the meaning of words and most of the sentences constituting the inputs submitted to them. This is manifested by their outputs and their coherence with human understanding. As such, LLMs display phenomenal understanding. Put it differently, LLM understanding can be measured, and not only can they display extensive breadth and depth of knowledge, but they exhibit, in various dimensions, although not in all human dimensions, phenomenal understanding.

We have also briefly shown how the understanding process emerges, in a distributed fashion, without involving a subject that understands. As such, LLMs can be said to have asubjective understanding.

Finally, LLM can be said to have “phenomenal asubjective understanding”. This novel concept suggests that LLMs can process and understand language in a way that is demonstrably effective, even without subjective experience. It challenges the notion that subjective experience is a necessary component of understanding. Furthermore it could cast new lights to the nature of our own understanding, that is, at least partially, subjective. This subjectivity could explain why some people resist the idea of LLMs exhibiting understanding, despite compelling evidence. Could asubjectivity be compatible with consciousness?

ACKNOWLEDGEMENTS

I could carry on this work only thanks to the allowance provided by the French institution France Travail.

REFERENCES

[1] Thomas Nagel, What is it like to be a bat? In The Language and Thought Series (pp. 159-168). Harvard University Press.

[2] David J. Chalmers, Could a Large Language Model be Conscious? From NeurIPS conference, 28 November 2022.

[3] Manuel R.F. Boissenin, La Conscience et ChatGPT, la nouvelle ère des cognitiques. Amazon, 2024.

[4] John Searle, Consciousness in Artificial Intelligence, Talks at Google, déc. 2015

[5] John Searle, The Chinese Room. 1999

[6] John R. Searle, Minds, brains, and programs. Behavioral and brain sciences, 3(3), 417-424. 1980

[7] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, ... and Yi Zhang, Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv:2303.12712 [cs.CL]

[8] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, Jacob Steinhardt, Measuring Massive Multitask Language Understanding, arXiv:2009.03300 [cs.CY], 2020

[9] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, Samuel R. Bowman, GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding, arXiv:1804.07461 [cs.CL], 2019

[10] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, Samuel R. Bowman, SuperGLUE: a stickier benchmark for general-purpose language understanding systems, SuperGlue: Learning Feature Matching with Graph Neural Networks, NIPS'19: Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2019

[11] Melanie Mitchell, David C. Krakauer, The Debate Over Understanding in AI's Large Language Models, arXiv:2210.13966 [cs.LG] 2022

[12] Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, ... and Rufin VanRullen, Consciousness in artificial intelligence: Insights from the science of consciousness. arXiv preprint arXiv:2308.08708, 2023.

AUTHOR

Dr. Manuel Boissenin finished is PhD in Computer Vision in 2009 at Sheffield Hallam University, he went back to France and worked for various companies from startups to large institutions such as the CEA. From 2015 to 2017 he worked at LaBRI where he did research in Deep Learning. He wrote two books and now he is an entrepreneur at the intersection between ML, DL and medical science.