A Benchmark for Consciousness in Large Language Models

Morgan Rivers
Department of Physics
Freie Universität Berlin
danielmorganrivers@gmail.com

August 2, 2024

Abstract

This paper presents a novel approach to evaluating consciousness in artificial intelligence systems. We discuss the importance of identifying machine consciousness before its emergence, the potential implications for AI ethics and development, and propose a benchmark for assessing consciousness in AI architectures. Our methodology combines elements of the ACT (Artificial Consciousness Test), which aims to detect robust indicators of consciousness while controlling for knowledge-based confounds. A simple training scheme to induce the emergence of consciousness in small open source language models is also introduced.

1 Introduction

The question of machine consciousness has become increasingly relevant as artificial intelligence systems grow more sophisticated. This paper argues for the importance of detecting machine consciousness before its full emergence and proposes a methodology for doing so.

While the exact conditions for consciousness remain unclear, certain AI architectures, particularly those beyond simple transformer models, may possess capabilities conducive to consciousness. A test has been proposed by Schneider et al. (Turner and Schneider, 2018) and expanded on by Perez et al. (Perez and Long, 2023) and Long (Long, 2023) that attempts to determine whether a language-enabled AI agent is conscious. Specifically, Perez and Long recommend the training for introspection of language models in order for consciousness to arise. While such approaches are interesting, the task remains to actually build a benchmark and training scheme for LLM's such that they can be tested at all, before diving into tweaks on a method of training involving introspection. The key practical questions remain unanswered in the literature: 1. How to introduce language models to language, without or with minimal examples generated from conscious beings (e.g. humans) to ensure they are not simply mimicking consciousness, rather than emergently exhibiting conscious behavior? 2. How to ensure such a training scheme can enhance language ability without language examples? 3. What key aspects of consciousness in a training scheme can both be learned reliably, and be reasonably expected to produce conscious behavior?

1.1 Importance of Detecting Machine Consciousness

Several factors underscore the significance of this research:

- Potential for machine suffering
- Implications for AI identity and goal-setting
- Ethical considerations regarding AI well-being
- Advancements in AI architectures mimicking conscious processes
- Possibilities for simulating conscious human experiences
- Potential for creating positive experiences (hedonium) in AI
- Impact on AI safety and alignment
- Advancing consciousness research

1.2 Gradient Nature of Consciousness

Consciousness is generally considered to exist on an analog gradient. Our research aims to identify factors that push AI systems towards greater degrees of conscious-like behavior.

1.3 Proposed method of achieving benchmark

Social animals that are widely regarded as conscious share the property that they have developed a theory of mind, and exhibit social behavior. Such animals over the aeons have emergently developed consciousness without other conscious beings training them on every behavior. We therefore have empirical evidence that selective pressures under the condition of socialization are therefore key to allow for consciousness to emerge.

We are left with the observation that in order for consciousness to emerge, language itself must also emerge in the process without continued examples of natural language, as occurred in the animal kingdom through evolutionary selective pressures.

We propose to first initialize an untrained transformer model of modest size (100 million parameters) and teach it rudimentary logic and language with rules generated from ProofWriter (Tafjord et al., 2021). By utilizing a small number of rules, it is possible to teach the model to both follow logical steps, and to use certain keywords in an appropriate way in the game of Diplomacy.

Next, we introduce the model to the board game diplomacy as was implemented in (Meta Fundamental AI Research Diplomacy Team et al., 2022) the following training scheme to meet the benchmark described above. By playing against itself in a strategic game, it will utilize the reasoning and ability to use words as meaningful tokens. Full-press Diplomacy is a game where effective cooperation, planning, and theory of mind are critical to effective gameplay. The secondary benefit to using the game is it mimics evolutionary environments of conscious animals, where cooperation, strategy, and theory of mind are critical to achieving reward. Finally, full-press diplomacy has been shown in (Meta Fundamental AI Research Diplomacy Team et al., 2022) to increase game performance over time with only self-play.

2 Training methodologies to achieve high confidence in ACT test results

It is still the case that we want models trained to give correct answers, not just mimicky answers. That implies that a model which is trained solely on task performance related to factual answering and understanding natural language communication, and not on mimicking some human-like output, would be a less misleading model when it comes to answering the ACT test honestly and accurately. Furthermore, understanding natural language must inevitably require some degree of skill in logical deduction.

This is why I wanted the model to self play diplomacy: I wanted it to be rewarded for action that represents internal states accurately, and not for any human-like mimicry. I know self-play can be a very powerful tool to train deep knowledge. All the ACT questions are going to be about what the model plans, prefers, likes, dislikes, knows, doesn't know, how it views others, the structure and breadth of inner experience.

It seems that in terms of understanding natural language (English) without enforcing a model to mimic or infer logically invalid statements, or to hallucinate, there are a limited set of options where training on these tasks or similar tasks has succeeded in the past:

A: training the model on natural language based (e.g. questions written by LLMs or humans) question answering

B: training on performance on language-involving self play, where self-talk or talk with fellow agents must be constrained to properly formatted English language. E.g. poker, diplomacy, rock paper scissors.

C: training on performance of coding tasks specified in natural language

D: training on entailment in knowledge databases (provides very large datasets, hundreds of thousands of prompts or even unbounded, with easily constrained input knowledge)

E: training on synthetically generated natural language first order logic tasks (provides very large datasets, hundreds of thousands of prompts or even unbounded, with easily constrained input knowledge)

F: Training on many synthetically generated natural language prompts about that model's internal state

G: training on the degree of certainty of answers to any of the above, or even just simple synthetically generated mathematics questions.

2.1 Regarding option A

There are further important restrictions: reducing the breadth of knowledge required helps a model to be smaller but maintain similar intelligence. So simply finding huge dataset of multiple choice questions is unlikely to work.

Reducing a model's exposure to concepts of feelings, preferences, emotions, discussion of internal states, assumptions about its own identity, philosophical knowledge mean that we must be strict about excluding these concepts from its training data, especially early on when foundational neural structures emerge.

(Perhaps information retrieval systems could help prevent memorization and display what the model has learned/memorized explicitly? That way we could train language model normally but observe the actual inner knowledge of the agent directly? But it would be hard to separate false "knowledge" about its internal experience from mimickry of humans in the standard training case).

Requiring a model to furthermore be reasonably capable and intelligent at a relatively small size means that we need to carefully vet the questions themselves to eliminate this.

Option A is thus simple, but difficult to implement. It is also hard for an entirely random language model to train on such a small dataset as human generated multiple choice.

2.2 Regarding option B

This suffers from the possibility that the language developed may still be incomprehensible to humans, even if enforced to be English (secret/hard to detect information can be transmitted through english, so what the text means to language models may be very different from what it means to humans). Furthermore, self play with multiple agents is hard to implement, and diplomacy especially is quite a complex game meaning much learning would be in an unrelated domain to inner planning, theory of mind. Much would have to be understanding board states, understanding game mechanics, rules, and strategy, and interpreting the prompting formats specific to diplomacy.

3 Methodology

In general, the methodology for training is as follows:

1. ensure the model has a simple grasp of english

2. use diplomacy to ensure development of theory of mind and use of english in communication

In order to arrive at 1, we will be iteratively testing the model such that it passes the ACT test some small portion of the time. We will need to add more natural language to the training until it can do so. As the model trains, we expect that consciousness arising will lead to improved performance on the ACT test. To make that be a reasonable chance, we will train a small version of https://github.com/google-deepmind/recurrentgemma at only 100 million parameters, and compare it with a similarly sized https://github.com/google/gemma_pytorch Gemma model.

This is just a first pass. Obviously, higher parameter counts would be nice (but are very expensive), and improving the training protocol enough that we need zero natural language would make for the strongest test of all, but this is infeasible as a first pass with limited resources for training.

As a source of natural language, the tinystories dataset may be appropriate, or perhaps coding textbooks as were used by Microsoft in the training of phi-1 and phi-1.5.

3.1 Comparative Analysis

We will compare results between potentially conscious architectures and control groups to identify significant differences in performance and behavior. Key metrics include the ability to discuss internal states, theory of mind, and emotional understanding.

The core feature of the analysis which greatly enhances the power of the ACT test is to compare enhanced architectures, such as RecurrentGemma, with the same training against control groups that have architectures that are less likely to exhibit consciousness, such as Gemma. This should allow the detection of even very small degrees of conscious emergence, and potentially running multiple models of different sizes can inform us of "scaling laws": given an appropriate training regime, how conscious compared to a human a model may be based on its parameter count and architecture.

3.2 Adaptation of the Artificial Consciousness Test (ACT)

We incorporate elements from Schneider and Turner's ACT, which offers several advantages:

- Neutrality regarding architectural details

- Consistency with human and AI ignorance about consciousness
- Allowance for radical cognitive differences between AI and humans
- Compatibility with various philosophical views on consciousness

3.3 Addressing ACT Limitations

To address concerns raised about the ACT, particularly regarding the ”epistemic sweet spot” for testing, we propose:

- Careful curation of training data to avoid consciousness-specific information
- Comparative analysis of models with similar training but different architectural properties
- Assessment of performance deltas on consciousness-related tasks

4 Discussion

4.1 Interpreting Results

Differences in performance between test groups will be analyzed to interpret signs of consciousness. Significant indicators include coherent self-referential statements, theory of mind capabilities, and appropriate responses to emotional and philosophical queries, as outlined extensively in the literature discussing the ACT test.

4.2 Justification and scope of the test

The test at hand is applicable to most, but not all consciousness theories. Notably Panpsychist and IIT are not amenable to the results of these tests.

There appears as well to be confusion over whether this test is flawed in some way (Udell, 2021). The first objection is ”The AI can't be entirely prevented from learning about consciousness-adjacent aspects of the world if it's going to be a conversational partner in a test of this sort. But it's unclear how much world knowledge would be too much, in the hands of a sophisticated learning AI. Implementation of the ACT will therefore require careful aim at an epistemic sweet spot.”.

While without a ”control” model, this is a decent objection, a reasonable sweet spot in terms of filtering training data/erasing relevant data can be found by gradually introducing more natural language in the training until the model passes approximately half the time and fails half the time. At that point, the ”sweet spot” can be significantly widened by simply getting the delta between models with and without the necessary architectural properties in terms of test performance, that have the same training data, and have been shown to perform similarly on cognitive tests. We will find models with fully emergent consciousness to be far ahead of models without architectures capable of consciousness on emergent consciousness benchmark, despite all of the training data being identical between the ”control” and the ”treatment” agent.

The second objection of (Udell, 2021) is as follows:

”By lack of miracles, the system's responses to the ACT will be some further implementation of its ordinary architectural processes. It might, for example, if given ACT Sample Question 1, do some complicated version of mechanically associating lack of survival with death and death with sadness, then answer that it would be sad to learn it would be deleted. Few AI consciousness doubters, we venture, ought to abandon their motivating architectural worry upon consideration of this situation — unless, perhaps, their motivating architectural worry is highly specific and of just the right sort. For most AI consciousness doubters, the underlying inference to the best explanation should not remove their doubt. This is the audience problem. Most theorists who have the kind of architectural worry that motivates seeking out a test like the ACT should, upon seeing that an AI passes the test, remain worried, rather than confidently concluding that the AI subject is conscious.”

The response to this is that the expectation of any test of consciousness should be understood as *providing further evidence to a reasonable audience*. A test cannot be expected to provide us with 100% confidence in the consciousness of an AI entity, as we cannot even know our fellow humans we interact with on a daily basis are conscious with this kind of certainty. Instead, we have reasonable evidence, inferred from a series of reasonable inferences. Explicitly:

1. I am conscious, and I have experiences

2. The other entity, whether a human, an AI, or a nonhuman animal, seems to contain exposure to an environment similar to my own in relevant ways (e.g. exposure to a need to develop theory of mind, selective pressures for accessing internal plans and identifying as a person)

3. The other entity is instantiated on a substrate (either silicon or organic) which contains various analogous structures to my own, such as neural patterns of firing, ability to access and use internal states, high levels of interconnectedness, and a capacity for reasoning and honest communication of internal states.

And finally,

4. The development of apparent consciousness in the entity has not been formed by simple mimicry of other conscious beings, and still it claims to possess consciousness as well as having the capability to discuss consciousness in detail and in an informed manner in a way a non-conscious being would have great difficulty with, unless it was extensively trained on the utterances and/or behaviors of conscious beings.

So unless we are refusing to be empirically grounded, the results of such a test are good, but not conclusive evidence that the entity is conscious, and excluding the theorists who insist on the near certainty of pan-psychism or IIT being the only possible way consciousness may be explained, any entity satisfying points 1-4 should logically convince a theorist that their theory about consciousness does not hold up to the empirical evidence, namely the consciousness benchmark proposed in this paper.

4.3 Large Language Model Architectures More Likely to Exhibit Consciousness

At the present moment, the most popular architecture, the transformer, is notably unlikely to architecturally capable of passing the ACT test, remarkably more so than many other architectures. The reasons are explained in detail in (Butlin et al., 2023). Essentially, there is very little recursion in the most popular LLM architecture, used in the likes of Gpt4o and Claude3.5. As these models by and large use the next-token prediction when running inference, and by and large each token is computed simply based on past tokens, and by-and-large not based on past computation there appears to be very little possibility of ”stream of consciousness” in current architectures.

However, after detailed investigation into the degree of incorporation of past computation into current computation, a more complex picture emerges: it may be that the calculation done along the layers in past token computation, which is incorporated into the layers of the transformers, does much of the useful computation: in other words, when an attention head on layer 17 of a language model tries to predict the next token, it seems to be attending to the residual stream of past layers *at the current depth of the layer in question*. At first this seems to be an accident to allow us to squeeze more compute for the massively parallel training these models were designed to take advantage of. But the fact that it does not simply look at the embedded version of the context token, and instead attends to the residual stream for that context token many layers deep implies that whatever compute was performed up to that layer in that context token remains for the model to use, and notably this is recomputed for every context token, each and every time the transformer runs! Unless the model is failing to utilize such ”wasted” context token computation, it seems likely that the model has learned to use this as ”extra” computation and thus incorporate more than just the identity of context tokens, but also compute some useful things for future token prediction even when computing past context tokens. This is where the concept of a ”stream of consciousness” may be appropriate: the model is in a sense sequentially computing on past tokens, iteratively using the results of these computations for the next token prediction, and thus may have a limited sense of future planning involved in the past token computations!

Still, the reintroduction of RNN’s to the scene of large language models (previously dominant in LSTM’s, but vanishing and exploding gradients lead to LSTM’s to fall out of favor compared to longer-context-window-capable transformers) may allow for a much more intuitive conception of ”stream of consciousness” in human terms. RNN’s allow for a ”hidden” representation that is updated across the context tokens, and recent architectures such as Hawk or Griffin (De et al., 2024) (avalailable from Google in the form of "RecurrentGemma" model) are capable of continuously updating this representation. It is therefore possible for such language models to authentically make a plan 100 tokens ago, store this in their hidden representation, and 100 tokens later implement the stored plan, something difficult with transformer architectures.

As an example, in the game of "20 questions", the model may say that it has selected an object for which we must guess. If it is a Hawk or Griffin architecture, then it very well may have selected an architecture and this is likely represented in its hidden layer. That selection is unlikely to change, all else equal, and when you guess the object it selected initially, it is likely to honestly inform you of the original selection.

On the other hand if a transformer language model plays the same game, there would need to be the storage of that selected object in a given token that occurred when it claimed it had selected the item. It would need to attend to this token each and every time a question was asked about the object, or re-store this data in later token generations by attending back to the original token. Not only that, only the latest layers would likely have sufficient access to the original token (as attention heads refer back to the value of residual stream on the layer they are at, so the first layer can only look 1 layer deep to see what to dump into its own residual stream). Given the difficulty of an attention head properly accessing the information under these circumstances, a transformer language model is much less likely to have the sense of time and stream of consciousness recognizable to humans, which may be a key component required to report conscious experience honestly.

5 Future Research Directions

5.1 Emotional Valence in AI

Propose methods for investigating emotional states in AI, such as analyzing latent space representations of concepts like "happy" and "sad".

6 Conclusion

In this paper, we proposed a novel approach to evaluating consciousness in artificial intelligence systems. By combining elements of the ACT with comparative analysis, we aim to detect signs of consciousness while controlling for knowledge-based confounds. This research highlights the importance of identifying machine consciousness before its emergence and underscores the ethical implications for AI development. Ongoing research in this area is crucial to ensure the responsible and ethical advancement of AI technologies.

References

Turner, E., & Schneider, S. (2018, January). Testing for synthetic consciousness: The ACT, the chip test, the unintegrated chip test, and the extended chip test. In CEUR Workshop Proceedings (Vol. 2287). CEUR-WS.

Perez, E., & Long, R. (2023). Towards Evaluating AI Systems for Moral Status Using Self-Reports. *arXiv preprint arXiv:2311.08576*. Retrieved from https://arxiv.org/abs/2311.08576

Tafjord, Ø., Mishra, B. D., & Clark, P. (2021). ProofWriter: Generating Implications, Proofs, and Abductive Statements over Natural Language. *arXiv preprint arXiv:2012.13048*. Retrieved from https://arxiv.org/abs/2012.13048

Meta Fundamental AI Research Diplomacy Team (FAIR)†, Bakhtin, A., Brown, N., Dinan, E., Farina, G., Flaherty, C., ... & Zijlstra, M. (2022). Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378(6624), 1067-1074.

Udell, D. B. (2021). Susan Schneider's proposed tests for AI consciousness: Promising but flawed. *Journal of Consciousness Studies*, 28(5-6), 121-144.

Long, R. (2023). Introspective capabilities in large language models. *Journal of Consciousness Studies*, 30(9-10), 143-153.

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. *arXiv preprint arXiv:2308.08708*. Retrieved from https://arxiv.org/abs/2308.08708

De, S., Smith, S. L., Fernando, A., Botev, A., Cristian-Muraru, G., Gu, A., Haroun, R., Berrada, L., Chen, Y., Srinivasan, S., Desjardins, G., Doucet, A., Budden, D., The, Y. W., Pascanu, R., De Freitas, N., & Gulcehre, C. (2024). Griffin: Mixing Gated Linear Recurrences with Local Attention for Efficient Language Models. *arXiv preprint arXiv:2402.19427*. Retrieved from https://arxiv.org/abs/2402.19427