Do Zombies *Understand*?
A Choose-Your-Own-Adventure Exploration of Machine Cognition

Anonymous ACL submission

Abstract

Recent advances in LLMs have sparked a debate on whether they *understand* text. In this position paper, we argue that opponents in this debate hold different definitions for *understanding*, and particularly differ in their view on the role of consciousness. To substantiate this claim, we propose a thought experiment involving an open-source chatbot Z which excels on every possible benchmark, seemingly without subjective experience. We ask whether Z is capable of *understanding*, and show that different schools of thought within seminal AI research seem to answer this question differently, uncovering their terminological disagreement. Moving forward, we propose two distinct working definitions for *understanding* which explicitly acknowledge the question of consciousness, and draw connections with a rich literature in philosophy, psychology and neuroscience.

1 Introduction: A Thought Experiment

Large language models (LLMs) achieve impressive results on various benchmarks, seeming to generalize to unseen tasks and domains (Brown et al., 2020). This initiated a debate on whether LLMs truly *understand* (Mitchell and Krakauer, 2022). On the one hand, several works claim that LLMs are starting to show signs of understanding text (Manning, 2022; Piantadosi and Hill, 2022; Bubeck et al., 2023), while on the other hand, others argue that LLMs are inherently incapable of understanding because they observe form without meaning (Bender and Koller, 2020; Bender et al., 2021; Marcus, 2022). Evidently, such works have differing opinions of what it means for a model to *understand*. Here, we do not advocate for a single "true" definition for *understanding*, and instead aim to shed new light on the roots of this debate.

We contextualize the debate on machine cognition within the *mind-body problem*, which has been at the center of vast philosophical debate, as well as intense empirical research in cognitive neuroscience. We follow Chalmers (1995), who asks whether the quality of *consciousness* - the ability to have subjective experiences - is a strict requirement for *understanding*, or whether it can also manifest in non-conscious agents. We argue that this question lies in the background of all discussion around whether LLMs truly *understand*.

To make this concrete, consider the following thought experiment: you are presented with Z, a new newfangled chatbot. Z is implemented in computer hardware and performs only mathematical manipulations of its input. It is completely open-source — you have access to its code, training data, weights, hyperparmeters, and any other implementation detail. You interact with Z and discover that it excels on all NLP benchmarks, and will do so on any possible test you will come up with in the future. In essence, Z is the chatbot equivalent of the philosophical zombie (Kirk, 1974; Chalmers, 1996); it outperforms humans on all tasks, supposedly without having subjective experience. *Do you consider Z as capable of "understanding"*?

If you answer *"Yes"*, turn to Section 3. For you, the path toward machine cognition lies in test sets of ever-increasing complexity, identifying ever-more subtle deficiencies in machine responses. If we reach this road's end, we will find Z.

If you answer *"No"*, turn to Section 4. You hold that consciousness is a prerequisite for *understanding*, as that is the only thing distinguishing Zombies from humans. We make several connections between recent neuroscience research and AI, e.g., the function of consciousness and advancements in the field of neural correlates of consciousness.

If you feel uncomfortable with either of these options turn to Section 5, where we address potential objections to our setup and assumptions.

This setup produces two distinct definitions and research agendas for machine *understanding*, which are currently conflated in AI discussion.

2 Background: Philosophical Zombies

The zombie argument is a thought experiment proposed in the context of debates about consciousness and its relationship to the physical world, i.e., to what is measurable (Kirk, 1974). It seeks to question the validity of physicalism, the belief that all that exists in our world, including consciousness, is physical (Stoljar, 2024). The zombie argument suggests that it is conceptually possible for there to be beings that are physically identical to humans but possess no conscious experiences. These are commonly referred to as "philosophical zombies".

Philosophical zombies behave just like humans. They appear to feel pain when injured, joy when pleased and can converse about the events in which they participate. Despite these behaviors, philosophical zombies possess no subjective experiences or "qualia" (Tye, 2021) – they do not consciously experience sensations, feelings, or thoughts. For instance, a philosophical zombie would react externally like a human would to stepping on a sharp object but would not internally suffer due to the painful sensation. Chalmers (1996) played a significant role in bringing the argument into the mainstream discourse, particularly in the context of the philosophy of mind.

We conjure the equivalent of a zombie chatbot. It is implemented on physical computer hardware, and it is capable of excelling on every NLP task, seemingly without conscious experience.

3 Zombies *do* Understand: Functional Definition of Understanding

One approach to machine cognition relies only on the model's behavior, independent of any internal experience. This definition holds that *understanding* can be inferred from performance on specific tasks. We formulate this notion for a task T in Definition 1:

Definition 1: Functional Understanding

A model Z *functionally understands* a task T if its performance on T is as good (or better than) a human that *understands* T.

This framing helps explain the common practice for testing *understand*ing in models through long-standing challenges, such as chess, Go, or language generation. McCarthy (1990) figuratively called such tasks the *Drosophila of AI*, drawing a parallel between research in AI and biology, where model organisms (e.g., the Drosophila fly) are chosen for wide benchmark experimentation with findings generalizing beyond that specific organism.

Evidently, the recurring trend in the last 70 years has seen tasks adopted as benchmarks for *understanding* until automated models functionally understand them. Then, the AI community *moves the goalposts* to another, arguably harder, external objective benchmark for *understanding*. Taken to the extreme, a model that functionally understands *every* potential benchmark is equivalent to our hypothetical Z chatbot. Notably, models excelling on these tasks are tested only externally and are not required to have any internal state linked to their success. Below, we outline some famous examples of this trend.

Perhaps the most well-known examples are the games of chess and Go. Chess served as a proxy task for *understanding* for nearly 50 years. Early works, such as Shannon (1950) and Turing (1953), already deemed chess a benchmark for machine intelligence. With the advent of deep learning models, chess engines now vastly outplay any human opponent (Silver et al., 2017). For all intents and purposes, these models *functionally understand* chess according to Definition 1. Consequently, chess was abandoned as a useful benchmark for *understanding*.[1] Instead, the game of Go was adopted as a marker for *understanding* (Bouzy and Cazenave, 2001; Van Der Werf, 2004), until here, too, models have outplayed even the best human players (Silver et al., 2016).

A natural follow-up question is whether LLMs can functionally understand. We argue that similar trends to Go and chess happen for certain NLP tasks. For example, natural language inference (NLI) has garnered significant attention since its introduction (Dagan et al., 2005), and was framed as "fundamental to understanding natural language" by the authors of SNLI, one of the most prominent benchmarks for the task (Bowman et al., 2015). However, as can be seen in Figure 1, the number of models developed over SNLI has dropped in recent years when performance on the benchmark was saturated, while similar trends are observed also for the follow-up MNLI dataset (Williams et al., 2018). To the best of our knowledge, there is no large scale effort to curate a new benchmark for

[1] New chess engines are still being developed, albeit without any claims about general *understanding* beyond chess.

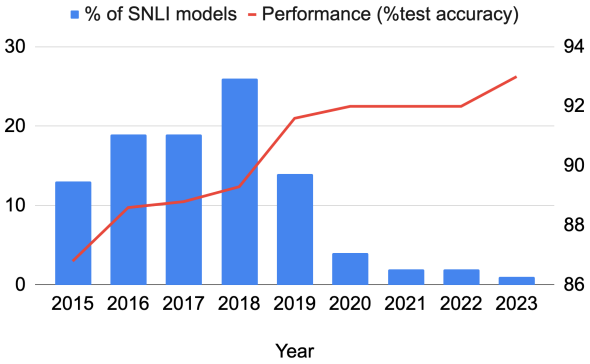

Figure 1: %Models tested on SNLI (blue bars, left axis) per year versus state-of-the-art performance on the benchmark (red line, right axis). Data collected from `paperswithcode.com`.

the task. It could be argued that LLMs functionally understand NLI, and the field has implicitly moved to other tasks. An indication that this trend does not stem from loss of interest in the task are various recent works that use NLI models as components within larger systems, showing that indeed NLI models are useful (Honovich et al., 2021; Laban et al., 2022; Aharoni et al., 2023; Min et al., 2023).

Adopting this notion of *understanding* implies getting other NLP tasks to go down this path, incrementally achieving functional understanding on as many tasks as possible. At the end of this path, *if it is reachable*, lies our hypothetical Z chatbot, which functionally understands *every* NLP task.

4 Zombies *don't* Understand: Consciousness as Prerequisite for Understanding

In contrast to the external approach to *understanding* in AI, stands a long line of work that either explicitly or implicitly requires models to have subjective experience. Importantly, these works view the quality of consciousness as an essential aspect of *understanding*, regardless of any external behavior. This notion is formulated in Definition 2:

Definition 2: Conscious Understanding

A model M is said to *consciously understand* if it has subjective experience immediate to it but not necessarily to an observer. Or, in Nagel (1974)'s words *there is something that it is like to be* M.

As we highlight below, this notion of *understanding* has been articulated by seminal works in the field of AI and NLP. In a section titled *Argument from Consciousness* from his famous paper, Turing (1950) cites (Jefferson, 1949):[2]

> Not until a machine can write a sonnet or compose a concerto because of **thoughts and emotions felt**, and not by the chance fall of symbols, could we agree that machine equals brain—that is, not only write it but **know that it had written it**. No mechanism could feel (and not merely artificially signal, an easy contrivance) pleasure at its successes, grief when its valves fuse, be warmed by flattery, be made miserable by its mistakes, be charmed by sex, be angry or depressed when it cannot get what it wants.
>
> Turing (1950)

Importantly, this reveals a strong tie between external behaviors, such as writing a sonnet or composing music, and subjective experiences, such as feeling emotions, in considering them as prerequisites for *understanding*, or intelligence.

This conflation is also evident in Searle's interpretation for his Chinese room argument (Searle, 1980). This questions if a computer can be truly intelligent by imagining a non-Chinese speaker using a rule-book to manipulate Chinese symbols, seemingly displaying comprehension without real *understanding*. Searle (2010) explicitly states that this argument was meant as a thought experiment for the existence of consciousness, or its lack thereof:

> I demonstrated years ago with the so-called Chinese Room Argument that the implementation of the computer program is not by itself sufficient for **consciousness** or intentionality.
>
> Searle (2010)

Since then, notable works have also connected the Turing test and Chinese room argument to consciousness (Churchland and Churchland, 1990; Gozzano, 1995; Dehaene and Sigman, 2012).

It seems that from the inception of AI research, foundational works defined *understanding* as the existence of consciousness. We believe that this drives part of the current discussion regrading the question of whether LLMs *understand*, evident in this quote from (Bender et al., 2021):

> Our human understanding of coherence derives from our ability to recognize interlocutors' **beliefs and intentions** within context. That is, human language use takes place between individuals who share common ground and are mutually **aware** of that sharing (and its extent), who have communicative intents which they use language to convey, and who **model each others' mental states as they communicate.**
>
> Bender et al. (2021)

Finally, this definition for understanding is in

[2] Emphasis is our own in all quotes.

line with (O'Gieblyn, 2021), who in her recent book advocated for consciousness as the defining factor of human intelligence:

> As AI continues to blow past us in benchmark after benchmark of higher cognition **we quell our anxiety by insisting that what distinguishes true consciousness is emotions, perception, the ability to experience and feel**. The qualities, in other words, which we share with animals.
>
> O'Gieblyn (2021)

To move forward on conscious understanding as articulated in Definition 2, we suggest following literature in neuroscience regarding the neural-correlate-of-consciousness (NCC; Koch et al., 2016). This field is dedicated to recognizing the neural dynamics in biological organisms that are associated with consciousness experience. For example, the Integrated Information Theory (IIT; Tononi et al., 2016), and more specifically the weak IIT, links elements of consciousness with wider information flow metrics, like recurrent processing, or global workspace. These findings can inform cognitively-inspired architectures, e.g., spiking neural networks (Mediano et al., 2022).

5 Other Possible Answers

Here we survey alternative answers to the question of whether zombies *understand*. We reply to these objections below, hopefully resolving seeming inconsistencies within our paradigm.

Argument: *Whether Z understands **depends on its implementation** (training data, architecture, hyperparameters, etc.), but that has nothing to do with conscious experience.*

This argument assumes that there may exist implementations of Z which will show that it indeed *understands*, e.g., if they involve complex feature manipulation or explicit reasoning steps, while there may exist other implementations which imply that Z does not *understand*, e.g., if all Z does is leverage spurious correlations or memorize an immense look up table, similar to the Chinese room argument (Searle, 1980). We posit that this argument is compatible within our discussion around functional understanding (§3). For example, if Z leverages spurious correlations, then by definition there are samples which do not exhibit these correlations and which will stump Z (otherwise they would not be spurious), contradicting our assumption that Z is a philosophical zombie, and does not make non-human errors. Similarly, since human language can produce an infinite amount of meaningful texts (Chomsky, 2002), and Z can only memorize a finite amount of samples (as it is implemented in finite hardware), then there must be samples outside of its memory on which it is bound to fail. This again contradicts our initial assumption that Z does not fail where humans do not fail.

Argument: *The question is ill-posed as **it does not define what is understanding**. Different definitions may lead to different answers.*

We do not aim to define apriori what constitutes *understanding*, and do not argue that there is a single "correct" definition. Instead, we try to tease apart what researchers mean when they use the term, specifically highlighting the role that consciousness plays in it, and examine how AI research may be explained through this lens. In fact, we claim that answering the question elucidates different definitions for *understanding* (Definitions 1,2). We invite researchers to engage with this question to examine *their* definition for *understanding*.

Argument: *The question is ill-posed as Z **is inconceivable**. Hence it is meaningless to discuss different properties of Z.*

This argument may stem from a philosophical stance often termed *mind-body dualism* (Robinson, 2023), positing that mental states are distinct from physical states, and that not all mental capacities can manifest in purely physical objects (e.g., computer hardware). We argue that this position is in fact compatible with the view that consciousness is a prerequisite for *understanding* (§4), and takes it a step further, claiming that consciousness can never be implemented in machines.

6 Discussion

We pose the question of whether Zombies *understand* to highlight consciousness's role in AI debates. We propose two definitions for *understanding*. One deals with *functional understanding*, and the other revolves around *conscious experience*. These definitions leads to different research agendas. This argument can be ported to other discussions about LLMs possessing human traits. E.g., Perry (2023) recently claimed that LLMs could not feel empathy. We argue that here, too, consciousness plays a major role in the definition of empathy. Similarly, the question of the relevance of consciousness to empathy can be unpacked by asking "Can Zombies be *Empathetic*?".

Limitations

We presented a thought experiment posing a philosophical question and have tried to answer it through the lens of two schools of thought within the fields of AI and NLP. While we tried to address potential reservations to our paradigm, it is possible that there are other answers that were not considered in this paper. We invite opinions and objections to further inform the discussion around machine cognition.

References

Roee Aharoni, Shashi Narayan, Joshua Maynez, Jonathan Herzig, Elizabeth Clark, and Mirella Lapata. 2023. Multilingual summarization with factual consistency evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3562–3591, Toronto, Canada. Association for Computational Linguistics.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Bruno Bouzy and Tristan Cazenave. 2001. Computer go: an ai oriented survey. *Artificial Intelligence*, 132(1):39–103.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

D. Chalmers. 1996. The conscious mind: in search of a fundamental theory.

David J Chalmers. 1995. Facing up to the problem of consciousness. *Journal of consciousness studies*, 2(3):200–219.

Noam Chomsky. 2002. *Syntactic structures*. Mouton de Gruyter.

Paul M Churchland and Patricia Smith Churchland. 1990. Could a machine think? *Scientific American*, 262(1):32–39.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*.

Stanislas Dehaene and Mariano Sigman. 2012. From a single decision to a multi-step algorithm. *Current opinion in neurobiology*, 22(6):937–945.

Simone Gozzano. 1995. Consciousness and understanding in the chinese room.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Geoffrey Jefferson. 1949. The mind of mechanical man. *British Medical Journal*, 1(4616):1105.

Robert Kirk. 1974. Sentience and behaviour. *Mind*, pages 43–60.

Christof Koch, Marcello Massimini, Mélanie Boly, and Giulio Tononi. 2016. Neural correlates of consciousness: progress and problems. *Nature Reviews Neuroscience*, 17:307–321.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Christopher D. Manning. 2022. Human language understanding & reasoning. *Daedalus*, 151:127–138.

Gary Marcus. 2022. Nonsense on stilts.

John McCarthy. 1990. Chess as the drosophila of ai.

Pedro A. M. Mediano, Fernando E. Rosas, Daniel Bor, Anil. K. Seth, and Adam B. Barrett. 2022. The strength of weak integrated information theory. *Trends in Cognitive Sciences*, 26:646–655.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *ArXiv preprint*, abs/2305.14251.

Melanie Mitchell and David C. Krakauer. 2022. The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences of the United States of America*, 120.

Thomas Nagel. 1974. What is it like to be a bat? *The philosophical review*, 83(4):435–450.

Meghan O'Gieblyn. 2021. *God, human, animal, machine: Technology, metaphor, and the search for meaning*. Anchor.

Anat Perry. 2023. Ai will never convey the essence of human empathy. *Nature Human Behaviour*, 7:1808 – 1809.

Steven T. Piantadosi and Felix Hill. 2022. Meaning without reference in large language models. *ArXiv preprint*, abs/2208.02957.

Howard Robinson. 2023. Dualism. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Spring 2023 edition. Metaphysics Research Lab, Stanford University.

John Searle. 2010. Why dualism (and materialism) fail to account for consciousness. *Questioning nineteenth century assumptions about knowledge, III: Dualism*, pages 5–48.

John R Searle. 1980. Minds, brains, and programs. *Behavioral and brain sciences*, 3(3):417–424.

Claude E Shannon. 1950. Programming a computer for playing chess. *Philosophical Magazine*, 41(314):256–275.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, L. Sifre, Dharshan Kumaran, Thore Graepel, Timothy P. Lillicrap, Karen Simonyan, and Demis Hassabis. 2017. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *ArXiv preprint*, abs/1712.01815.

Daniel Stoljar. 2024. Physicalism. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Spring 2024 edition. Metaphysics Research Lab, Stanford University.

Giulio Tononi, Mélanie Boly, Marcello Massimini, and Christof Koch. 2016. Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience*, 17:450–461.

Alan Turing. 1953. Digital computers applied to games. In B. V. Bowden, editor, *Faster than thought*, pages 286–310. Sir Isaac Pitman & Sons, Ltd., London.

Alan M. Turing. 1950. Computing machinery and intelligence. *Mind*, LIX:433–460.

Michael Tye. 2021. Qualia. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Fall 2021 edition. Metaphysics Research Lab, Stanford University.

Erik Van Der Werf. 2004. *AI techniques for the game of Go*. Citeseer.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.