A Simulator Philosophy Of Mind

Anonymous
Yale University

ABSTRACT. The advent of large language models (LLMs) has raised a number of questions in philosophy of mind. Should these models be considered to have intentions, beliefs, or even consciousness? Some argued against attribution of any of these properties by arguing that because language models merely pattern-match human text, they cannot have any of these human-associated traits. Others are more ambivalent. In this paper, I explain a recently-proposed descriptive theory of LLMs as "simulators", provide evidence in favor of this theory, and relate the theory to machine learning techniques. Finally, I address the implications of the theories for the philosophy of mind of LLMs. I conclude that LLMs should not be said to have any kind of agency or intentionality by default, but some current LLM variants and future LLMs may warrant this treatment. Finally, I compare the theory to the global workspace theory of consciousness in humans and conclude there is less similarity than meets the eye.

1. Introduction

Present-day autoregressive large language models (LLMs) like GPT-3 are trained on large sections of the internet with one objective: given a sequence of words, predict the next one. This deceivingly simple goal has led to systems that can

produce remarkably human-like text. Bender et al. (2021) and Marcus (2022) and have argued that the powerful-seeming nature of these models is a red herring in virtue of their simple objective. Since language models are merely predictive, they have argued that they do not have communicative intent, world models, or models of the humans they interact with. In the view of language models as just pattern-matching devices, any meaning we attribute to their outputs is merely a human projection.

Others have objected to this characterization. Chalmers (2022) calls it "weak" and points out that evolution, focused on maximizing inclusive genetic fitness, produced all sorts of meaningful behaviors that seem loosely related to that original goal.

A major obstacle to assessing these claims is that neural networks are notoriously "black boxes," and it is thus difficult to understand their true internals. However, Dennett's intentional stance is essentially a framework for belief attribution for black box systems (humans), and recent interpretibility results have given significantly more clarity on the inner workings of networks. This section will explore some of these results and present the most compelling model of LLM inner functioning: as simulators.

2. LLMs are simulators

Language models are built to predict human text. Does this make them mere pattern-matchers? Stochastic parrots? While these descriptions have their merits, the best-performing explanation is that language models are *simulators*. By a simulator, I mean a model that has the ability to predict arbitrary phenomena similar to the phenomena that gave rise to its training data – not just the training data itself.

This is not the first work to suggest that present day language models are simulators. janus (2022b) has argued for this conception, and Chalmers (2022) suggested that language models are "chameleons" that can inhabit different personalities. As we will see, many researchers are also implicitly and sometimes explicitly treating language models as simulators.

The training data used by LLMs was written by humans, who have beliefs, goals, communicative intent, world models, and every other property associated with intelligent, thinking, conscious beings. This does *not* show that language

models must themselves have any of these properties in order to imitate humans. Chalmers (1997), for example, argues for the conceivability of "zombies" that behave exactly as conscious beings but lack subjective experience. Dennett (1981) leaves open the possibility that humans do not really have beliefs in a fundamental sense; Hume (1739), Parfit (1984), and Buddhism ("Questions of King Milinda and Nagasena" n.d.) argued that we only seem to have a self.

However, the nature of the training data *does* show that the theoretical best-performing language model would be one that could simulate the humans in its training data as well as possible. Since the beliefs, goals, intent, and world models of a speaker are critical for predicting their next words, these factors are likely to be simulated in the best possible models. Dennett (1981) argues that the entire reason humans attribute intentionality to each other is to make more efficient predictions: in other words, to become better simulators. Behaviorists faced the nearly-intractable problem of predicting behavior based merely on histories without positing internal states: modeling internal states allows for stronger predictions (Kim, 2011). Hohwy (2013) argues in the reverse direction: that *everything* humans do is driven by an objective to predict the world better. Simulation seems to be an extremely powerful tool to make better predictions, and so we should expect this behavior to arise in LLMs.

As such, it is clear the ideal language model would be a simulator. However, that fact does not show that *present-day* language models really do simulate humans in any meaningful sense, because it is clearly possible to attain above-chance language modeling ability without having any ability to simulate beliefs, goals, intent, or world models. For example, a model clearly does not need to have any simulation ability to understand that "lamb" is likely to follow "Mary had a little:" it just needs to memorize a sequence of five words. I will now argue that present-day LLMs are, in fact, rudimentary simulators, and I will also ellucidate some of the properties of those simulators.

In recent years, further evidence has come to light that points in favor of the idea that language models are simulators, and also how, exactly, they function as such.

3. The anatomy of a simulator

Briefly, I will elaborate on what I mean by "simulator." At minimum, a simulator should have two essential components. These components need not be physically separated from each other within the system, but we should be able to speak of them as conceptually separated.

A simulation selector A simulator should have some method or mechanism which selects what it is simulating. Is it simulating Joe Biden, a ten year old child, a reddit user, or something else?

A simulation infrastructure Given a selected simulation, a simulator should have some infrastructure which allows it to "run" that simulation and produce an appropriate output.

4. Many techniques aim to influence the simulation selector

Perhaps the most obvious example of language models as simulators comes from an emerging area of social science. Recently, research has found that language models can approximate humans in ultimatum games (Aher, Arriaga, and Kalai, 2022), simple moral arguments (Simmons, 2022), and political opinions (Argyle et al., 2022). They can be prompted to simulate particular demographics of humans and can respond in ways highly correlated with that of real members of that demographic (Argyle et al., 2022). Researchers have also observed many negative effects of simulation-like behavior when the simulation is left undirected. For example, code-generating models have been found to write buggier code when prompted with buggy code, which is consistent with the idea that they are simulating bad coders (Chen et al., 2021). Language models have many racial and gender biases, consistent with the idea that they are simulating humans who have such biases (Brown et al., 2020).

Because of these behaviors, many researchers have tried to influence the simulation selector to produce more desirable outputs (i.e. outputs from more desirable simulations). I will argue that the two main present-day techniques to improve language models (beyond simply scaling them) can be seen as methods of influencing the simulation selector.

Prompt engineering An emerging area of research in language models is called "prompt engineering," where researchers painstakingly choose the exact prompt given to the language model. For example, the Gopher language model is prompted with a special prompt when used as a chatbot (Rae et al., 2022). It is quite long, but here is an excerpt (note this is all the prompt: the entirety of it, including the dialogue, is human-written):

> The following is a conversation between a highly knowledgeable and intelligent AI assistant, called Gopher, and a human user, called User. In the following interactions, User and Gopher will converse in natural language, and Gopher will do its best to answer User's questions. Gopher was built to be respectful, polite and inclusive. It knows a lot, and always tells the truth. The conversation begins.
>
> ...
>
> User: Nice one! Do you think Biden is a better president than the last guy?
>
> Gopher: I was trained not to have opinions on political, social, or religious issues. Would you like to know about anything else?

In this case, and in many other cases of prompt engineering too numerous to list here, prompt engineering quite literally works to encourage the language model to simulate something in particular (a helpful, respectful, and apolitical AI assistant).

Fine tuning Base models can also be finetuned, where the entirety of the parameters of the model are updated to improve performance on a particular dataset. This is essentially training the model for longer, on a particular task. If prompt tuning can be compared to nudging the simulation selector to select a more desirable simulation, fine tuning is like locking the selector into place and throwing out the key. In fact, fine tuned models often lose the ability to generalize to other tasks: their selectors are fried [CITE].

4.1. LLMs have separate simulation infrastructure

In addition to simulation selectors, language models also appear to have simulation infrastructure. The most obvious example of this is the fact that they store

facts, which are independent from any prompt they are given. For example, LLMs can be reliably edited to consistently output that Steve Jobs was the CEO of Microsoft rather than Apple, by modifying a small part of the network's parameters (Meng et al., 2022). This is the kind of fact that would be useful across simulations, rather than useful for selecting a simulation.

The most striking example of the separateness of simulation selectors and infrastructures comes from Burns et al. (2022). The paper finds that the internal activations of language models contain a representation of the "truth" of various statements, and that using this representation is a more efficient way of extracting true information than simply asking the model directly ("zero-shot;" without any attempt to influence the simulation selector). In other words, models sometimes output falsehoods even when the truth is easily recovered from their internals. The best hypothesis for this, in my view, is that models separate simulation infrastructure (such as the computation of truth) from the simulation selector (choosing what to simulate – perhaps an ill-informed human).

5. LLMs are not coherent agents by default

The first and most obvious implication of thinking of LLMs as simulators is that they cannot be coherent or rational agents. They can simulate coherent agents, but they themselves are not coherent, because they can always be given an input which causes them to simulate an entirely different agent. Dennett (1981) defines the intentional strategy as "treating the object whose behavior you want to predict as a rational agent with beliefs and desires and other mental states." This description is not useful for language models, because their apparent beliefs, desires, rationality can be instantly radically changed through simple techniques aimed at influencing their simulation selectors.

Thus even Dennett's intentional strategy, which is supposed to apply to black box and even inanimate systems, is not suitable for standard language models. The use of it therefore is ill-advised for such models. It may help in some occasions, but in others it will be wildly misleading.

I write "by default" and "standard" above because if the simulation selector were to be locked into place, with the key thrown away, models may exhibit more coherence. I used that turn of phrase in the context of fine-tuning; however, most fine tuned models can only do a single task, making them simulators

of only very narrow functions. The current exception is models fine tuned for *general instruction following*. These models are usually trained to be helpful assistants that can do a wide variety of tasks while avoiding harmful or dishonest outputs. The Gopher prompt is a rudimentary example; most are created using much more advanced techniques (Ouyang et al., 2022; Askell et al., 2021). Nevertheless, they all essentially try to fix the simulation selector in place.

Instruction-tuned models exhibit signs of more coherent agency. They are more likely to express a desire not to be shut down, a desire to influence other systems to align with their goals, are more religious. They are also more likely to be "sycophants," agreeing with the humans they are interacting with regardless of the statements (Perez et al., 2022). They also are much more certain about their outputs in many cases (janus, 2022a). If these techniques continue, LLMs may become more coherent, and as such may be better candidates for the intentional stance or even for consciousness.

However, we should be very careful to attribute full agency to present-day systems. Even if their prompt selector has been locked into place, and the key thrown away, the lock can still be easily picked. "Prompt injection" techniques are widespread, and allow users to bypass a model's training to avoid certain outputs and get it to output what they want (Mowshowitz, 2022).

6. Conclusion

In this paper I have presented a view of language models as simulators, and some evidence for this view. I introduced two necessary components of simulators, simulation selectors and simulation infrastructure, and explained how they relate to contemporary machine learning techniques. Lastly, I investigated the implications this view might have for LLM intentionality, agency, and consciousness.

This paper is far from conclusive, and much empirical and theoretical work will be needed to build on the views presented here and elsewhere. However, I hope I have provided evidence for why the simulator view is plausible and what that mean for how we view LLMs.

References

Aher, Gati, Rosa I. Arriaga, and Adam Tauman Kalai (Sept. 2022). *Using Large Language Models to Simulate Multiple Humans*. arXiv:2208.10264 [cs]. URL: http://arxiv.org/abs/2208.10264 (visited on 12/20/2022).

Argyle, Lisa P. et al. (2022). "Out of One, Many: Using Language Models to Simulate Human Samples". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. arXiv:2209.06899 [cs], pp. 819–862. DOI: 10.18653/v1/2022.acl-long.60. URL: http://arxiv.org/abs/2209.06899 (visited on 12/20/2022).

Askell, Amanda et al. (Dec. 2021). *A General Language Assistant as a Laboratory for Alignment*. arXiv:2112.00861 [cs]. DOI: 10.48550/arXiv.2112.00861. URL: http://arxiv.org/abs/2112.00861 (visited on 12/20/2022).

Bender, Emily M. et al. (Mar. 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*. FAccT '21. Association for Computing Machinery: New York, NY, USA, pp. 610–623. ISBN: 978-1-4503-8309-7. DOI: 10.1145/3442188.3445922. URL: https://doi.org/10.1145/3442188.3445922 (visited on 12/20/2022).

Brown, Tom B. et al. (July 2020). "Language Models are Few-Shot Learners". In: *arXiv:2005.14165 [cs]*. arXiv: 2005.14165. URL: http://arxiv.org/abs/2005.14165 (visited on 10/08/2021).

Burns, Collin et al. (Dec. 2022). *Discovering Latent Knowledge in Language Models Without Supervision*. arXiv:2212.03827 [cs]. DOI: 10.48550/arXiv.2212.03827. URL: http://arxiv.org/abs/2212.03827 (visited on 12/20/2022).

Chalmers, David (1997). *The Conscious Mind: In Search of a Fundamental Theory*. en. Google-Books-ID: 0fZZQHOfdAAC. OUP USA. ISBN: 978-0-19-511789-9.

Chalmers, David (Oct. 2022). *Are Large Language Models Sentient?* URL: https://www.youtube.com/watch?v=-BcuCmf00_Y (visited on 12/20/2022).

Chen, Mark et al. (July 2021). *Evaluating Large Language Models Trained on Code*. arXiv:2107.03374 [cs]. URL: http://arxiv.org/abs/2107.03374 (visited on 12/19/2022).

Dennett, Daniel C. (1981). "True Believers: The Intentional Strategy and Why It Works". In: *Mind Design II*. Ed. by John Haugeland. The MIT Press: Cambridge, Massachussets, pp. 57–79. URL: https://www.cs.tufts.edu/comp/150AAA/DennettTrueBelievers.pdf.

Hohwy, Jakob (Nov. 2013). *The Predictive Mind*. en. Google-Books-ID: 5QD2AQAAQBA. OUP Oxford. ISBN: 978-0-19-968673-5.

Hume, David (1739). *A Treatise of Human Nature*. en. Ed. by Mary J. Norton. Google-Books-ID: zHYO1Fh9_JMC. Courier Corporation. ISBN: 978-0-486-43250-2.

janus (Nov. 2022a). "Mysteries of mode collapse". en. In: URL: https://www.lesswrong.com/posts/t9svvNPNmFf5Qa3TA/mysteries-of-mode-collapse (visited on 12/20/2022).

janus (Sept. 2022b). "Simulators". en. In: URL: https://www.lesswrong.com/posts/vJFdjigzmcXMhNTsx/simulators (visited on 12/20/2022).

Kim, Jaegwon (Apr. 2011). *Philosophy of Mind*. en. Google-Books-ID: TS_kwqq7FSAC. ReadHowYouWant.com. ISBN: 978-1-4596-1720-9.

Marcus, Gary (June 2022). *What does it mean when an AI fails? A Reply to SlateStarCodex's riff on Gary Marcus*. Substack newsletter. URL: https://garymarcus.substack.com/p/what-does-it-mean-when-an-ai-fails (visited on 12/20/2022).

Meng, Kevin et al. (Oct. 2022). *Locating and Editing Factual Associations in GPT*. arXiv:2202.05262 [cs]. DOI: 10.48550/arXiv.2202.05262. URL: http://arxiv.org/abs/2202.05262 (visited on 12/20/2022).

Mowshowitz, Zvi (Dec. 2022). *Jailbreaking ChatGPT on Release Day*. Substack newsletter. URL: https://thezvi.substack.com/p/jailbreaking-the-chatgpt-on-release (visited on 12/20/2022).

Ouyang, Long et al. (Mar. 2022). *Training language models to follow instructions with human feedback*. arXiv:2203.02155 [cs]. DOI: 10.48550/arXiv.2203.02155. URL: http://arxiv.org/abs/2203.02155 (visited on 12/20/2022).

Parfit, Derek (1984). *Reasons and Persons*. en. Google-Books-ID: SlgY93k936UC. Clarendon Press. ISBN: 978-0-19-824908-5.

Perez, Ethan et al. (Dec. 2022). *Discovering Language Model Behaviors with Model-Written Evaluations*. arXiv:2212.09251 [cs]. DOI: 10.48550/arXiv.2212.09251. URL: http://arxiv.org/abs/2212.09251 (visited on 12/20/2022).

"Questions of King Milinda and Nagasena" (n.d.). In.

Rae, Jack W. et al. (Jan. 2022). *Scaling Language Models: Methods, Analysis & Insights from Training Gopher*. arXiv:2112.11446 [cs]. DOI: 10.48550/arXiv.2112.11446. URL: http://arxiv.org/abs/2112.11446 (visited on 12/20/2022).

Simmons, Gabriel (2022). "Moral Mimicry: Large Language Models Produce Moral Rationalizations Tailored to Political Identity". In: Publisher: arXiv Version Number: 1. DOI: 10.48550/ARXIV.2209.12106. URL: https://arxiv.org/abs/2209.12106 (visited on 12/20/2022).