A rebuttal of two common deflationary stances against LLM cognition

Anonymous ACL submission

Abstract

Large language models (LLMs) are arguably the most predictive models of human cognition available. Despite their impressive human-alignment, LLMs are often labeled as "*just* next-token predictors" that purportedly fall short of genuine cognition. We argue that these deflationary claims need further justification. Drawing on prominent cognitive and artificial intelligence research, we critically evaluate two forms of "Justaism" that dismiss LLM cognition by labeling LLMs as "just" simplistic entities without specifying or substantiating the critical capacities they supposedly lack. Our analysis highlights the need for a more measured discussion of LLM cognition, aiming to better inform future research and the development of artificial intelligence.

1 Introduction

Over 70 years ago, Alan Turing posed a question that has since captivated computer scientists, cognitive scientists, and philosophers alike: "Can machines think?" (Turing, 1950). With the recent proliferation of increasingly capable artificial intelligence systems (e.g., Bubeck et al., 2023)—namely, large language models (LLMs)—variants of this question have made their way far beyond the confines of academic departments.

Although LLMs have been shown to be predictive of human representations and behavior across a broad range of tasks (Binz et al., 2024; Tuckute et al., 2024; Hussain et al., 2024), a number of critics maintain that LLMs cannot be said to possess genuine cognition because they are "just...": "next-token predictors", "function approximators", or "stochastic parrots", and thus lack some essential capacity necessary for "thought", "reasoning", or "understanding" (henceforth, "cognition"). Unfortunately, such deflationary claims often fail to state what exactly this capacity is and have been given the pejorative label "Justaism" (pronounced "just-a-ism") due to the confident self-evidence with which they are wielded (Aaronson, 2023). Such views on the reality of LLM cognition, have implications for people's willingness to use them as scientific tools (Binz et al., 2025), and trust such systems in everyday contexts (Mitchell and Krakauer, 2023).

In what follows, we discuss two flavors of Justaism, and provide a critical analysis of these positions based on cognitive and artificial intelligence research. We refer to the flavors' prototypical forms but also provide specific examples found in the literature and public discussion on LLM cognition in a companion webpage (anonymous.4open.science/r/againstJustaism-5510). We conclude our analysis by putting forth three guiding principles to help clarify the status of LLM cognition.

Before proceeding, we clarify the scope of our work. While we focus on two forms of Justaism, other substantial perspectives on LLM cognition exist and deserve consideration. These views differ fundamentally from Justaism and hence are not the target of our critique. First, some empirical research highlights specific LLM cognitive deficits (e.g., McCoy et al., 2024; Turpin et al., 2024; Berglund et al., 2023). Rather than denying LLM cognition outright, such work is better understood as qualifying the extent of cognitive abilities in LLMs. Second, other research presents substantive arguments against LLM cognition, for example, by distinguishing *form* (syntax) from *meaning* (semantics) (e.g., Bender and Koller, 2020; Searle, 1980). We view such efforts as making important definitional and conceptual progress on cognition—an endeavor we also advocate in our conclusion. We hope these attempts may contribute to a more precise conceptual landscape, ultimately shaping how we evaluate and compare artificial and biological intelligence.

2 Flavors of Justaism

2.1 Anti-simple-objectives

"It's just a next-token predictor."

Perhaps the more common form of Justaism, which we dub *anti-simple-objectives Justaism*, takes issue with how LLMs are pre-trained. The assertion is that because the LLM pre-training objective is simply to predict the masked or next token, LLMs cannot be doing something as complex as cognition.

Assuming proponents of this view believe that humans possess cognition, anti-simple-objectives Justaism can be questioned by making the following facetious analogy to humans and other creatures shaped by evolution: We humans are "*just* next-child producers", stumbling forward in pursuit of the all-encompassing base objective of inclusive fitness maximization. The point here is not to argue that humans should actually be thought of in such a way but to highlight a common error with this kind of deflationary thinking—the error of assuming that simple base objectives necessarily produce simple systems.

Of course, there are important differences between next-token prediction and inclusive fitness maximization. For instance, the ancestral environment from which we evolved was potentially richer than the online text corpora used to train LLMs. Combined with a sufficiently complex nervous system and other distinguishing factors (e.g., resource competition), biological evolution may lead to the development of *instrumental objectives* that are more conducive to cognition than next-token prediction.

However, even if it were the case that these distinguishing factors were pivotal to the development of instrumental objectives *in humans*, it is nevertheless plausible that cognition-enabling instrumental objectives could be acquired via other means during next-token-prediction-based pre-training. In fact, empirical evidence suggests that LLMs are already employing such instrumental strategies in order to achieve high performance on the base objective (through a process known as *mesa-optimization*, Von Oswald et al., 2023). There is also reason to expect that these instrumental objectives are similar to those of humans. After all, the LLM pre-training distribution was generated (mainly) by humans, who would have had various (instrumental) motives driving their text production. An LLM that learns to model these human objectives and incorporate them into its prediction could thus improve its performance on the training distribution by better capturing the data generating process (Hubinger et al., 2019). There is also empirical precedence for this sort of convergence, with research in representational alignment demonstrating that predicting human-generated text can lead to increased alignment between LLMs and human brains (Sucholutsky et al., 2023; Binz et al., 2024).

Relatedly, LLM (instrumental) objectives need not be especially complex to be on par with those of human beings. After all, many foundational theories of human cognition posit relatively simple objectives as fundamental components, with prominent examples including predictive brain theories (e.g., *Bayesian brain*, *predictive coding*, *active inference*, Clark, 2013). Notably, these objectives may not be so different from next-token prediction, which raises a similar question to the evolutionary analogy that opened this section: If simple predictive objectives are generally considered insufficient for the development of cognition, might it be that humans similarly lack genuine cognition?

Finally, it is important to qualify that most modern-day LLMs are not only (pre-)trained with next-token prediction but also go through several stages of fine-tuning. These often include reinforcement-learning from (subjective) human feedback (Bai et al., 2022) and (objective) rule-based rewards (Guo et al., 2025), which are targeted at improving the model's helpfulness. As such, it is now often factually incorrect to claim that LLMs are only trained to predict the next token, though it is still true that the vast majority of data and compute goes into such pre-training (see, e.g., Guo et al., 2025).

Ultimately, the extent to which next-token prediction enables or precludes cognition is a question that requires further theoretical and empirical research. Nevertheless, we hope the above arguments demonstrate that it is *by no means self-evident* that an LLM is devoid of cognition.

2.2 Anti-anthropomorphism

"It's just a machine."

A second prominent form of Justaism, which we dub *anti-anthropomorphic Justaism*, claims that attributing cognition to machines constitutes a fundamental error. In its strongest form, it argues that such thinking commits a category error because cognition is *by definition* a human capacity. On this view, the essential capacity that LLMs lack and humans possess is just that: humanness.

Although logically valid, we would argue that this view is unproductively restrictive. Advances in scientific theory often come from generalizing concepts beyond their initial application. One instructive example comes from animal cognition research, where, in response to a growing body of empirical evidence, researchers began to see great utility in ascribing capacities previously thought to be uniquely human, including emotion, self-awareness, or consciousness, to non-human animals (De Waal, 2016). We believe it should be *in principle* acceptable to make such conceptual generalizations for information processing systems more broadly.

There are, of course, more moderate forms of anti-anthropomorphic Justaism. For instance, one might take the view that although it is not a problem *in principle* to talk about LLM cognition, the burden of evidence for doing so should be set very high. One reason for this would be to guard against the Eliza effect (Mitchell and Krakauer, 2023), which refers to the human propensity to all-too-liberally ascribe "thought" to even the simplest of machines (Weizenbaum, 1976).

Although we agree that it is important to reject naive anthropomorphism, we note that running counter to anthropomorphism is another, perhaps more infamous, human tendency: anthropocentrism. Regarding cognition, anthropocentrism is the tendency to view capacities such as "thought" as so unique that it would not make sense to ascribe them to "lesser" systems, such as non-human animals (see, e.g., Singer, 2011; Harris and Anthis, 2021). In the context of artificial intelligence, it can be observed in the well-documented phenomenon of algorithmic aversion—the human tendency to rely more on human advisors over equally good or better-performing algorithms (Jussupow et al., 2022). Anthropocentrism may ultimately have implications for the adoption of novel technologies that have the potential to contribute to human wealth and well-being.

In light of humans' countervailing tendency to view their own cognition as exceptional, we would advocate for specifying more precisely the forms of cognition in question and the evaluative criteria to be employed. We believe this will enable more substantive discussions of and comparisons between the capabilities of humans and other information-processing systems.

3 Conclusion: Toward a more measured discussion

In support of a more measured discussion of LLM cognition, we would like to advance three guiding principles: (i) modesty regarding human cognition (and our understanding of it), (ii) consistency for future work comparing humans and LLMs, and (iii) a focus on empirical benchmarks.

Regarding modesty, we would reiterate that human history is littered with delusions of human exceptionalism (De Waal, 2016). This is despite our limited understanding of the mechanisms underlying cognition. Thus, although we fully support cautioning against the dangers of (naive) anthropomorphism, we see the need for a backstop against the opposite tendency: viewing human cognition as too special to also be ascribed to LLMs.

Regarding consistency, we would reiterate the need for consistent goalposts: Are we applying the same standards to LLMs as we would to humans? For instance, if we wish to reduce LLM cognition to its pre-training objective (i.e., next-token prediction), we must show why the same reductionism should not apply to humans as well. Similarly, when LLMs commit errors that appear so elementary to us as to discredit LLM cognition, it is important to recall the host of fallacies and illusions that humans are susceptible to and consequently may not so easily identify or view as significant. These considerations not only help guard against certain biases (e.g., algorithmic aversion), but they can also provide a new perspective on human cognition by helping identify aspects of cognition that are, in fact, uniquely human. For instance, it has been argued that (current) LLMs probably lack sentience, consciousness, or self-awareness (Chalmers, 2023)—capacities that are thus unique to humans and other animals.

Finally, we are sympathetic to (Turing, 1950)'s view (among others, e.g., Niv, 2021) that discussions of cognition should focus on observables. As Trott et al. (2023) note, axiomatic rejections of LLM cognition can lead to positions that have no empirically testable implications. Not only does this run contrary to good scientific practice, but it can also lead to investigations of LLM cognition that lack practical relevance. After all, it is predominantly the behavior of a system that impacts the world. Consequently, we believe in the need for clear and consistent empirical benchmarks (e.g., Chollet, 2019) that allow for direct evaluations of the cognitive capacities of humans and LLMs.

Ultimately, the jury is still out on the existence and extent of LLM cognition. We hope these principles can help researchers move beyond Justaic reasoning towards a deeper, more measured understanding of the cognitive capacities of LLMs.

4 Limitations

Our work has two important limitations. First, we detail only two major forms of Justaism, but there are other stances in the literature that may also qualify as Justaic. For instance, a third could be characterized as *anti-memorization Justaism*, which asserts that LLMs are not doing cognition because they are simply reproducing patterns learned during training. Unfortunately, these objections often fail to: (i) evidence the extent to which the model is, in fact, relying on memory, (ii) justify why such memorization is so at odds with cognition, and (iii) acknowledge that humans often rely on memorization for tasks that are ostensibly reasoning-based (e.g., Bors and Vigneau, 2003; Jaeggi et al., 2008).

Second, since our main focus is to argue against unsubstantiated claims and call for a more measured discussion on LLM cognition, we do not make a substantive positive argument for or against LLM cognition in this work. Doing so would involve considering different definitions and operationalizations of cognition and proposing various empirical means for measurement and evaluation that are suitable for LLMs (and humans). Fortunately, work to this effect is already underway (e.g., Chollet, 2019). We hope to see more research in this direction.

5 Ethics statement

To the best of our knowledge, our work conforms to the ACL Code of Ethics. The work is of a theoretical nature and does not involve human participants or personal data. We believe it does not pose any significant risks.

References

Scott Aaronson. 2023. The problem of human specialness in the age of AI. https://scottaaronson.blog/?p=7784. Accessed: 2024-03-31.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

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.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2023. The reversal curse: LLMs trained on "A is B" fail to learn "B is A". In *arXiv preprint arXiv:2309.12288*.

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K Eckstein, Noémi Éltető, et al. 2024. Centaur: a foundation model of human cognition. *arXiv preprint arXiv:2410.20268*.

Marcel Binz, Stephan Alaniz, Adina Roskies, Balazs Aczel, Carl T Bergstrom, Colin Allen, Daniel Schad, Dirk Wulff, Jevin D West, Qiong Zhang, et al. 2025. How should the advancement of large language models affect the practice of science? *Proceedings of the National Academy of Sciences*, 122(5):e2401227121.

Douglas A Bors and François Vigneau. 2003. The effect of practice on raven's advanced progressive matrices. *Learning and Individual Differences*, 13(4):291–312.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. In *arXiv preprint arXiv:2303.12712*.

David J Chalmers. 2023. Could a large language model be conscious? *arXiv preprint arXiv:2303.07103*.

François Chollet. 2019. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.

Andy Clark. 2013. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204.

Frans De Waal. 2016. *Are we smart enough to know how smart animals are?* WW Norton & Company.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Jamie Harris and Jacy Reese Anthis. 2021. The moral consideration of artificial entities: a literature review. *Science and engineering ethics*, 27(4):53.

Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. 2019. Risks from learned optimization in advanced machine learning systems. In *arXiv preprint arXiv:1906.01820*.

Zak Hussain, Marcel Binz, Rui Mata, and Dirk U Wulff. 2024. A tutorial on open-source large language models for behavioral science. *Behavior Research Methods*, 56(8):8214–8237.

Susanne M Jaeggi, Martin Buschkuehl, John Jonides, and Walter J Perrig. 2008. Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences*, 105(19):6829–6833.

Ekaterina Jussupow, Izak Benbasat, and Armin Heinzl. 2022. Why are we averse towards algorithms? a comprehensive literature review on algorithm aversion. In *Proceedings of the 28th European Conference on Information Systems*.

R Thomas McCoy, Shunyu Yao, Dan Friedman, Mathew D Hardy, and Thomas L Griffiths. 2024. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41):e2322420121.

Melanie Mitchell and David C Krakauer. 2023. The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120.

Yael Niv. 2021. The primacy of behavioral research for understanding the brain. *Behavioral Neuroscience*, 135(5):601–609.

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

Peter Singer. 2011. *The expanding circle: Ethics, evolution, and moral progress*. Princeton University Press.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C Love, Erin Grant, Iris Groen, Jascha Achterberg, et al. 2023. Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*.

Sean Trott, Cameron Jones, Tyler Chang, James Michaelov, and Benjamin Bergen. 2023. Do large language models know what humans know? *Cognitive Science*, 47(7):e13309.

Greta Tuckute, Nancy Kanwisher, and Evelina Fedorenko. 2024. Language in brains, minds, and machines. *Annual Review of Neuroscience*, 47.

A. M. Turing. 1950. Computing machinery and intelligence. *Mind*, 59:433–460.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2024. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36.

Johannes Von Oswald, Eyvind Niklasson, Maximilian Schlegel, Seijin Kobayashi, Nicolas Zucchet, Nino Scherrer, Nolan Miller, Mark Sandler, Max Vladymyrov, Razvan Pascanu, et al. 2023. Uncovering mesa-optimization algorithms in transformers. *arXiv preprint arXiv:2309.05858*.

Joseph Weizenbaum. 1976. Computer power and human reason: From judgment to calculation. *San Francisco*.