The problem of likening LLMs to human cognition

Abstract: Rothschild (2024) argues that only AI systems trained on natural language display robust domain-general reasoning, because language's abstraction and compression make inference computationally tractable, and that this supports a strong cognitive role for language in humans. I challenge both the move from LLM performance to reasoning and the extrapolation from artificial to biological minds. First, Rothschild adopts a purely mechanical, behavioral notion of reasoning and treats next-token prediction as a basic model of all linguistic inference. This view collapses the distinction between producing human-like outputs and possessing human-like cognitive states, making perfect imitation of reasoning indistinguishable from reasoning itself in a way that is reductionist and vulnerable to Searle-style worries. Second, evidence from animal cognition and neuroscience suggests that robust, structured inference can occur without language, and that in humans language and domain-general reasoning are neurally dissociable. LLMs therefore show what large-scale statistical learning over text can achieve, not what language does in human thought. Nonetheless, they provide valuable proof-of-concept against strong nativist views. I conclude that LLMs are useful analogues for exploring learning over symbolic input, but cannot yet ground substantive claims about language's distinctive role in human reasoning.

What exactly can be learned from the rise of large language models (LLMs) and can conclusions be drawn from the rise of LLMs and extrapolated to human cognitive research? In his 2024 paper 'Language and Thought: The View from LLMS', Rothschild seems to suggest that we can. Rothschild argues that recent advances in large language models can provide suggestions that show that language has a potentially substantial cognitive upside. This is based on the premises that 'only AI systems with access to natural language perform well at general reasoning' and that the 'abstraction and compression in natural language makes general reasoning computational tractable' (Rothschild, 2024). As such, it is suggested that it may very well be the efficiency of 'linguistic representation has a profound influence on our own mental development' (Rothschild, 2024). Essentially, by observing which architectures yield human-like general reasoning, we can gain indirect evidence about how our own minds work. Within this experiment, domain-general AI systems, such as LLMs, stand out at general reasoning tasks more so over domain-specific systems, seemingly as a result of their ability to operate over language. It is suggested that this is because language is a more highly compressed representational format which encodes exactly the information relevant for inference and reasoning while omitting vast amounts of detail. This compression drastically shrinks the data used, making next-token prediction a more efficient proxy for general inference. On the other hand, raw video data or other non-language forms remain intractable as a result of their less compressible data. From these premises, Rothschild draws a couple of conclusions. Firstly, that LLMs vindicate connectionist architectures while also revealing the usefulness of symbolic representation. Second, they provide a 'proof of concept' of sorts that suggests that language could be responsible for significant cognitive abilities within humans. Lastly, we can extend this latter conclusion to the idea that this shows potential evidence for the hypothesis that language has a significant role in human thought and that it might have served as a likely driver that has shaped human inferential capacities and may have had a role to play in our cognitive development. Yet, this study poses a few questions. Just to what extent can we extrapolate data from LLMs to draw conclusions about human cognition and should the way LLMs reason even be likened to the way humans reason? To what extent are we anthropomorphising LLMs to fit a human like criteria for reasoning? If we are going to use evidence from 'the Great AI Experiment' (Rothschild, 2024) to draw conclusions about our own cognitive abilities, then we first need consensus about how useful this evidence actually is and whether it is relevant to be used as evidence at all. While I do not have definitive answers to the aforementioned questions, it is still relevant to raise a couple of points that should be considered and potentially discussed when considering such evidence.

The main issue is whether we can really liken the way LLMs reason to the same way that humans reason in order to draw conclusions about human cognitive abilities. Rothschild argues in favour of a 'mechanical' sort of reasoning (Rothschild, 2024). He sidesteps the philosophical problem of reasoning requiring specific mental states, intentions, or connections to the external

world to be considered reasoning and instead defines reasoning as simply ‘whether a system can draw reasonable inferences (in the mechanical sense of spitting them out), not whether it shares other features of human mental states that go along with this ability’ (Rothschild, 2024). After giving this definition, he goes on to say that ‘predicting how texts continue, can be seen as a very basic computational model of all forms of inference within the medium of language’ (Rothschild, 2024). In this way, he can conclude that LLMs, through their next-token prediction, can be seen as reasoning in a mechanical sense. However, should this really count as evidence that LLMs reason? If LLMs only predict, then is it not safe to assume that what LLMs truly do is just simulate a human like kind of reasoning, rather than reason themselves? In essence, where it might look like an LLM is inferring or reasoning about a particular topic, what it is actually doing is simulating what reasoning on that topic would simply look like. The question that emerges from this is then, is there or should there be a difference between actual human reasoning and simulated or a perfect imitation of reasoning? Implicitly it seems that Rothschild is taking the position that there is no difference between actual reasoning and perfect imitation of reasoning. As Rothschild never makes the claim that LLMs reason in the same way that humans do, and in fact asks us to disregard certain human elements, in order to accept his conclusion. He explicitly says he is interested only in whether a system ‘can draw reasonable inferences (in the mechanical sense of spitting them out),’ not in whether it has the mental states or intentions that usually accompany human reasoning. With that behavioural definition, any system that maps inputs to appropriate outputs by predicting what follows next in a text counts as performing inference, and next-token prediction can then be treated as a ‘very basic computational model of all forms of inference within the medium of language’ (Rothschild, 2024). The problem with this however, is that it can be seen as reductionist, which makes the jump from LLM studies to human cognition a far one. For one, it collapses the distinction between producing the right pattern of words and having the right kind of cognitive process or state. Perfect imitation of reasoning is, by stipulation, treated as reasoning, so there is no conceptual space left for something like Searle’s Chinese Room (Fodor, 1985). Second, it erases and reduces concepts and potential ways of reasoning such as intentionality, conscious deliberation, grasp of logical relations to irrelevant for human reasoning by treating them as inessential extras on top of the right functional profile. On that view, there is indeed no difference between actual reasoning and its perfect simulation, but only because “actual reasoning” has been defined down to nothing more than producing indistinguishable outputs, which is why this line of thought is reductionist, and while it serves its purpose to provide some sort of empirical evidence for learning perhaps, it should not be used as evidence to conclude that language is somehow directly or indirectly responsible for heightened cognitive abilities. This is why Searle’s Chinese Room thought experiment could be useful to consider, as it dramatises this problem of extrapolating LLM findings to human cognition by highlighting how the human brain and the deep neural network architecture that makes up LLM are not the same and do not function the same (for an explanation of Searle’s Chinese Room see Cole, 2024, referenced below). The person in Searle’s Chinese Room still imitates a competent Chinese speaker, yet understands nothing (Cole, 2024). This means that LLMs are very much like the Chinese Room, devices that produce human-like outputs, regardless of whether they possess human like thoughts. If the claim is that the Great AI Experiment teaches us something about our biological minds, then a biological argument of what it means to think, and by extension reason, must be considered, otherwise no concrete conclusions can be drawn. An account that deliberately side steps intentionality, and studies about cognitive architecture cannot by itself support conclusions about the cognitive role of language in human, who have those properties. This is not to say that the way LLMs reason is not useful and I am also not saying that language, and its ability to compress data is not useful for the efficacy of AI models per se, that may very well be the case, I am simply pointing out that the jump from AI research to human cognition based on AI models is not a small one, and we need to be careful and consider other evidence before drawing such conclusions.

Rothschild further admits that if a domain-specific neural architecture can be found to reason within the same capacities as LLMs (which are domain-general), then his claims can be dismissed (Rothschild, 2024). Alternatively, what repercussions, if any, would there be for his conclusion if it were found that other living beings reason in a similar way to humans without the capacity for language? In Chapter 6 of ‘The Philosophy of Animal Minds’, Elisabeth Camp discusses the language of baboon thought. Studies show that baboons track intricate social relations such as dominance hierarchies, alliances, and coalitions in ways that are discrete, compositional, and systematically deployed across contexts (Camp, 2009). While traditionally this

might suggest a form of language of thought in baboons (Seyfarth and Cheney, 2013), Camp argues this does not entail that they think in a language-like code (Camp, 2009). Their cognition may be implemented in maps, diagrams, or other non-linguistic formats that nevertheless support structured inference (Camp, 2009). In this same way we can draw an analogy to LLMs. An LLMs ability to produce outputs with coherent semantic structure does not itself show evidence for reasoning using a human-like language. The model is trained to minimise prediction error over token sequences (Buckner, 2023). Its internal representations are high-dimensional vectors whose interpretability is itself debated (Buckner, 2023). This does not immediately license the conclusion that "language as such" is doing the inferential work, rather than generic function approximation over a distribution of text. Even if we disregard this as irrelevant, as what is truly relevant is the output, which is correct by definition, then what LLMs show is what can be achieved by large-scale statistical learning over human-encoded text, and not necessarily what language as such does in human cognition. As demonstrated by Camp, evidence exists that language might not be necessary for reasoning, hence raising questions about the validity of conclusions. On the other hand, these baboon studies can easily be dismissed, and Rothschild acknowledges this. The claim is that animals and infants can indeed make inference and reason, but adult humans, due to natural language presumably, are much 'more tractable' (Rothschild, 2024). However, is this view not too anthropocentric? It seems that this view treats non-linguistic reasoning in animals and infants as preliminary or weaker form of reasoning. It implies that other species' abilities are deficient relative to ours rather than alternative solutions to their own ecological and cognitive problems. In fact, certain animals may even have cognitive abilities 'stronger' than ours such as animals which are able to process a wider spectrum of colours or sounds. From the baboon's point of view perceptual representations might be just as "tractable" given their needs. If we are going to accept that language improves our cognitive ability, we first need to assume that our cognitive ability is superior to that of other animals, and from a logical perspective this might be the case, but I would argue that from a purely 'reasoning' perspective such a claim is not as clear cut as it first hinges on what is meant by the term 'reasoning'. As such, dismissing 'reasoning' as merely mechanical outputs seems not only reductionist as mentioned earlier, but also anthropocentric. It is for this reason that I do not believe that reasoning should be defined in such a simple way, and other arguments for it should not be dismissed if we are going to draw conclusions from the Great AI Experiment. Furthermore, evidence from neuroscience can be found that challenges this notion that language is responsible for increased cognitive capabilities. The assumption that Rothschild makes is that because LLMs are language trained and do well on domain-general tasks, then language is a potential plausible substrate for domain-general reasoning in humans. Work by Fedorenko, Piantadosi, and Gibson (2024) can imply instead that language is a substrate for communication rather than reasoning. For example, severe aphasia caused by damage to the part of the brain responsible (or commonly accepted to be responsible) for language can leave complex cognition intact while certain brain damage or intellectual disabilities can impair reasoning while sparing language capacity (Fedorenko et al., 2024). This is supported by imaging which shows different areas of the brain being active depending on the task and show that demanding reasoning can show activity while language areas are silent while sentence processing activates languages areas with no reasoning (Fedorenko et al., 2024). From this, it can be safer to conclude that language is an efficient vehicle for expressing and transmitting thought, and a rich source of training data from which a system can learn to approximate some aspects of human reasoning, but it is neither necessary nor sufficient for the underlying cognitive capacities that enable this.

Having considered all of these objections, where does that leave us and what remains of Rothschild's project? The fact that language and text based data processing is more efficient over video does seem to suggest something about the efficiency of language. From the point of view of AI engineering this proves a practical insight that can inform future AI architecture, and perhaps there is space for potential conclusions to be drawn about human learning. More importantly however, Rothschild seems to venerate the empiricist view for human learning. It offers a proof of concept for how potentially exposure to language alone can result in sophisticated task performance, something alike reasoning. This challenges extreme nativist positions that deny the human capacity to 'learn' syntax or semantic structure from raw stimulus (or data) alone. While it does not prove anything, it still challenges these views, which is enough reason for a discussion to be had. Likewise, other studies support this idea such as Piantadosi (2024) which showcases how LLMs may refute the more extreme claims about the necessity for innate structures in language acquisition. Rothschild's claims can be seen as direct evidence for Piantadosi's claim

that ‘The rise and success of large language models undermines virtually every strong claim for the innateness of language that has been proposed by generative linguistics. Modern machine learning has subverted and bypassed the entire theoretical framework of Chomsky’s approach, including its core claims to particular insights, principles, structures, and processes’ (Piantadosi, 2024). However, while we can accept the previous two conclusions, in order to go a step further and discuss the impact of language on cognitive abilities, and use LLMs evidence, a discussion on what constitutes reasoning needs be had beforehand, and other fields considered and weighed before such a conclusion can be made. In this respect, LLMs provide a useful analogue that shows how a powerful learner can use symbolic, linguistically encoded structure as a scaffold for more complex inference, but does not show that “only AI systems with access to language perform well at general reasoning” in a way that reveals language’s distinctive role in human thought, something we need to be more careful about.

2475 words

Bibliography:

- Buckner, Cameron J. (2023) From Deep Learning to Rational Machines: What the History of Philosophy Can Teach Us about the Future of Artificial Intelligence’. New York, 2023; online edn Available at: https://doi.org/10.1093/oso/9780197653302.001.0001. [Accessed 24 Jan. 2026].

- Camp, Elisabeth (2009) ‘A language of baboon thought’ in The Philosophy of Animal Minds, edited by Robert Lurtz, Cambridge University Press.

- Cole. David (2024) ‘The Chinese Room Argument’, The Stanford Encyclopedia of Philosophy (Winter 2024 Edition), Edward N. Zalta & Uri Nodelman (eds.), Available at: https://plato.stanford.edu/archives/win2024/entries/chinese-room/. [Accessed 30 Jan. 2026].

- Fedorenko, E., Piantadosi, S.T. & Gibson, E.A.F. (2024) Language is primarily a tool for communication rather than thought. *Nature* **630**, 575–586 Available at: https://doi.org/10.1038/s41586-024-07522-w/. [Accessed 30 Jan. 2026].

- Fodor, Jerry (1985) ‘Fodor's Guide to Mental Representation: The Intelligent Auntie's Vade-Mecum’ in Mind, vol. 94, no. 373, 1985, pp. 76–100. *JSTOR*, Available at: http://www.jstor.org/stable/2254700. [Accessed 30 Jan. 2026].

- Piantadosi, Steven T. (2024) ‘Modern language models refute Chomsky's approach to language’, in From fieldwork to linguistic theory. Berlin: Language Science Press, pp. 353–414.

- Rothschild, Daniel (2024) ‘LANGUAGE AND THOUGHT: THE VIEW FROM LLMS’, in Oxford Studies in Philosophy Language, Lepore and Sosa (eds.).

- Seyfarth, Robert & Cheney, Dorothy. (2013) The primate mind before tools, language, and culture. 105-122. Available at: https://www.researchgate.net/publication/290986716_The_primate_mind_before_tools_language_and_culture. [Accessed 29 Jan. 2026].