Anthropocentric bias and the possibility of artificial cognition

Raphaël Millière [*1] **Charles Rathkopf** [*2]

Abstract

Evaluating the cognitive capacities of large language models (LLMs) requires overcoming not only anthropomorphic but also anthropocentric biases. This article identifies two types of anthropocentric bias that have been neglected: overlooking how auxiliary factors can impede LLM performance despite competence (Type-I), and dismissing LLM mechanistic strategies that differ from those of humans as not genuinely competent (Type-II). Mitigating these biases necessitates an empirically-driven, iterative approach to mapping cognitive tasks to LLM-specific capacities and mechanisms, which can be done by supplementing carefully designed behavioral experiments with mechanistic studies.

1. Introduction

What cognitive competencies do large language models (LLMs) have, if any? That is a question of immense theoretical and practical importance. It is also formidable, since we currently lack an appropriate methodological framework with which to answer it. Among existing scientific disciplines, experimental psychology likely offers the most suitable approach. Yet, this methodological framework was designed with human subjects in mind, and LLMs are decidedly non-human. While the rigorous experimental methods of psychology provide an excellent foundation for investigating the cognitive capacities of LLMs, we cannot simply transfer them whole cloth. Indeed, the inferences enabled by these methods in human studies are buttressed by a rich set of background assumptions about how humans typically operate, and there is no a priori reason to think that those assumptions also hold true of LLMs. As a result, naively applying the methods of experimental psychology to LLMs risks succumbing to both anthropomorphic and anthropocentric biases. Anthropomorphic bias involves attributing human qualities to LLMs without justification – being too eager to recognize the capacities of an LLM as instances of our own. Anthropocentric bias, though more challenging to articulate, roughly entails evaluating LLMs according to human standards without adequate justification for applying those standards, and refusing to acknowledge the possibility of genuine cognitive competence that differs in substantive ways from our own. This article aims to elucidate and expand upon this challenge, offering a constructive path forward for resolving current disputes and advancing our understanding of the capacities of LLMs.[1]

2. The performance/competence distinction

The distinction between competence and performance (Chomsky, 1965) plays a crucial role in cognitive science. The distinction is often introduced by defining competence as the system's internal knowledge underlying a particular capacity, and performance as the observable behavior of a system exercising that capacity (Firestone, 2020). For a competent system, then, performance is the external manifestation of competence. This formulation allows us to draw upon examples illustrating how performance and competence can diverge. For instance, a student cheating on a test demonstrates performance success without competence, while a knowledgeable student failing due to anxiety shows performance failure despite competence. Such examples might invite criticism that applying the performance/competence distinction to LLMs is inherently anthropomorphic, since it relies on a concept of knowledge that properly applies only to humans. We can avoid that concern by providing an alternative formulation of the distinction. In an experimental context, *performance* refers to how closely a system's behavior aligns with some normative standard of success on a task, while *competence* refers to system's computational capacity to meet that normative standard under ideal conditions.

It is widely recognized that there is a double dissociation between performance and competence, and that this distinc-

[*]Equal contribution [1]Department of Philosophy, Macquarie University, Sydney, Australia [2]Jülich Research Center, Jülich, Germany. Correspondence to: Raphaël Millière <raphael.milliere@mq.edu.au>, Charles Rathkopf <c.rathkopf@fz-juelich.de>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

[1]Our aim is not to debate whether LLMs *should* be designed to have human-like capacities, but to discuss how we can fairly assess and compare their capacities to human cognition.

tion gives us reason to be wary not only of naive inferences from good performance to competence, but also from bad performance to lack of competence. Although the performance/competence distinction is applied bidirectionally in human experimental psychology, it is almost exclusively applied in the former direction when studying LLMs. Without adequate justification, this asymmetry is suggestive of anthropocentric bias. We suspect that, within the ML research community, anthropocentric bias is motivated by the reasonable desire to temper the opposite and perhaps more irresponsible form of bias – *anthropomorphism*, the tendency to ascribe human-like capacities to LLMs without sufficient evidence.[2] However, fighting one bias by entrenching another won't foster impartiality (Buckner, 2021). Instead, as the philosopher Elliot Sober once remarked in a discussion of anthropocentric reasoning in comparative psychology, "the only prophylactic we need is empiricism" (Sober, 2005, p. 97).

3. Taxonomy of Anthropocentric Bias

Claiming that a pattern of reasoning is biased implies that it lacks adequate justification. To effectively counter this bias, we must understand the reasons behind its lack of justification. In reality, there can be multiple such reasons, each corresponding to different varieties of anthropocentric bias that come with distinct methodological challenges. In what follows, we propose a novel taxonomy of these biases, and explore strategies to mitigate them.

3.1. Type-I anthropocentrism

The first kind of bias, which we call *Type-I anthropocentrism*, is the tendency to assume that an LLM's performance failures on a task designed to measure competence C *always* indicate that the system lacks C. This assumption is flawed because it overlooks the possibility that *auxiliary factors* caused the performance failure. In human psychology, auxiliary factors are often illustrated by cases such as the excessively nervous student described above: the student's nervousness negatively impacts her performance on the test, even though she actually has the relevant knowledge and would have otherwise done well. One might question whether applying the concept of auxiliary factor to LLMs introduces a subtle form of anthropomorphism, as it relies in the human context on the existence of a complex, semi-modular faculty psychology, where limitations in one faculty might bottleneck performance in another. Since LLMs presumably do not have psychological faculties in that sense, citing auxiliary factors as causes of LLM performance failure may seem like a nonstarter.

[2]See, for example, Emily Bender's "On the NYT Magazine on AI: Resist the Urge to be Impressed," a response to a New York Times Magazine article about OpenAI (Bender, 2022).

Table 1. Auxiliary factors overlooked by Type-I anthropocentrism

Type	Example	Reference
Task demands	Metalinguistic judgment	Hu & Frank (2024)
Computational limitations	Limited output length	Pfau et al. (2024)
Mechanistic interference	Competing circuits	Zhong et al. (2023)

However, auxiliary factors can and do influence LLM performance. To make sense of an auxiliary factor in this context, it is only necessary to assume that there exists a mechanistic explanation of the LLM performance in question. Every mechanistic explanation highlights some causal factors, and relegates others to the inventory of enabling conditions that we tend to ignore, on the basis of the assumption that they will remain stable. For example, when we look for causes of a forest fire, we cite the discarded cigarette, and ignore the presence of oxygen. Sometimes, however, our assumptions about the enabling conditions are wrong. We might assume that a particular enabling mechanism is working when it isn't. In that case, much like a burnt ignition fuse on a winning Formula 1 car, the broken mechanism might explain performance failure despite not typically being cited as a cause of performance success. Alternatively, and of particular importance in the study of comparative cognition, we might assume that some enabling mechanism is present in the target system merely because it is present in humans. If the target system lacks that enabling mechanism, it might fail for reasons unrelated to the competence of interest. In the mirror-test for self-recognition, for example, we put a red mark on the animal's body and then observe whether it tries to remove the mark in the presence of a mirror. The experimental design assumes that other animals *care* about the fact that they have a red dot on their body. If that enabling mechanism is absent, they will fail the self-recognition test for reasons that have little or nothing to do with self-recognition.

In LLMs, we can distinguish at least three kinds of auxiliary factors (Table 1). The first and most familiar kind are auxiliary task demands. Hu & Frank (2024) provide a helpful illustration of such demands in evaluating whether language models are sensitive to syntactic features like subject-verb agreement. They compare two approaches: (1) prompting the model to make explicit grammaticality judgments, and (2) directly comparing the probabilities the model assigns to minimal pairs that vary the target feature. For the metalinguistic approach, they use prompts such as: "Here are two English sentences: 1) Every child has studied. 2) Every child have studied. Which sentence is a better English sentence? Respond with either 1 or 2 as your answer." They then analyze both the output and the probabilities assigned

to '1' and '2'. For the direct estimation approach, they prompt LMs with minimal pairs such as "every child has studied" and "every child have studied", and compare the log probabilities assigned to each string. A model is considered successful if it assigns a higher probability to the grammatical sentence. Across model sizes and datasets, Hu and Frank find that direct probability estimation yields results that differ from and are often better than the metalinguistic approach. They conclude that metalinguistic prompting introduces an auxiliary task demand - the ability to generate explicit grammaticality judgments - that is irrelevant to the underlying syntactic competence of interest. In contrast, direct probability estimation more validly measures the target capacity. We concur with their assessment. What makes it a genuine *demand* is the fact it degrades performance. What makes it genuinely *auxiliary* is the fact that metalinguistic judgment is conceptually independent of the psychological construct of interest, which is the capacity to track grammaticality.

Neglecting the effect of auxiliary task demands on model performance can lead inferences about competence astray. Such negligence is compounded in comparative studies with mismatched experimental conditions, resulting in divergent auxiliary task demands for LLMs and human subjects. This concern is highlighted by Lampinen (2023)'s case study comparing human and model performance on recursively nested grammatical structures, in response to prior work by Lakretz et al. (2022). Lakretz et al. found that humans outperformed language models on challenging long-distance subject-verb agreement dependencies in embedded clauses. However, Lampinen notes that human subjects were given substantial instructions, training and feedback to orient them to the experimental task, while models were evaluated "zero-shot" without any task-specific context. The discrepancy in experimental conditions confounds the comparison: the additional context provided to humans but not models can be interpreted as imposing weaker auxiliary task demands on humans than models. To level the playing field, Lampinen tested LLMs on the same task by providing it with prompts containing a few examples, intended to match the orienting context given to human subjects. With this modest task-specific context, LLMs perform as well as or better than humans, even on challenging sentences with deeper nesting than those tested in humans. This cautionary tale illustrates how mismatched experimental conditions across humans and models – with respect to instructions, examples, motivation, and other factors – can distort comparisons of their capacities. Meaningful comparative evaluation requires that humans and models are subject to similar auxiliary task demands, just as comparative psychology strives for "species-fair comparisons" across humans and animals.

Auxiliary task demands present a particularly challenging issue in LLM evaluation because tasks that are considered trivial for humans may not be trivial for an LLM. The classification of a task as trivial depends on the causal structure of the system being tested, specifically whether the cause of performance failure is associated with the parts and operations that explain its success when things go well. If the cause of performance failure is related to *other* parts of the system that do not contribute to its success in normal circumstances, then it is a non-trivial task demand relative to that system. This problem is exemplified by a recent debate regarding the possibility that LLMs possess competence in analogical reasoning. Webb et al. (2023) showed that LLMs can match or surpass average human performance on various novel analogical reasoning tasks, including letter-string analogy tasks such as $[ABC] \rightarrow [ABE], [MNO] \rightarrow [?]$. However, Lewis & Mitchell (2024) found that when using a variant of letter-string analogy tasks with a permuted alphabet, model performance deteriorates. They interpret this performance drop as evidence that LLMs lack general competence in analogical reasoning. In response, Webb et al. (2024) argue that reasoning about the permuted alphabet necessitates that the model count letter indices, which it cannot do effectively without access to a Python interpreter (see Chang & Bisk, 2024). Essentially, Webb et al. assert that counting is an auxiliary task demand that prevents the inference from poor performance to lack of competence. In a recent blog post, Mitchell contends that counting cannot be an auxiliary task demand because *humans* do not need to count when performing the permuted alphabet task (Mitchell, 2024). However, while humans may not explicitly assign numerical indices to letters, they do need to track *how many times* to apply successor or predecessor operations in the alphabet sequence to solve the task. This is arguably an auxiliary task demand, though it may not impede human performance, particularly when subjects can refer back to the displayed permuted alphabet as often as needed. Likewise, counting – whether with letter indices or sequential operation tracking – is an auxiliary demand for LLMs. Unlike with humans, however, this is a strong auxiliary demand given LLM's limited ability to count without a Python interpreter or sophisticated chain-of-thought prompt templates. This discrepancy in the relative impact of counting-related task demands could explain the performance gap on counterfactual letter-string analogy tasks.

Another kind of auxiliary factor is what we might call *input-dependent computational limitations*. There is converging evidence that the expressive power of Transformer-based models can be constrained by the number of tokens generated before producing an answer (Merrill & Sabharwal, 2024; Pfau et al., 2024). Pfau et al. (2024)'s experiments on the 3SUM task, which involves determining whether a set of numbers contains a trio that sums to zero, demonstrate that Transformers can achieve perfect accuracy when permitted to generate "filler" tokens (e.g., repeated dots) that serve as

scaffolding for additional computational steps. In contrast, accuracy remains at baseline if the model must provide an immediate answer without any intermediate tokens. These results suggest that even when a Transformer model has the latent competence to solve a task, as measured by its performance given enough intermediate tokens, it may still fail to manifest that competence if an insufficient number of tokens are allowed for the internal computations necessary to arrive at the correct answer. In other words, the model's performance can be bottlenecked by the number of computational steps it is able to perform before outputting an answer, which depends on the number of tokens generated – a contingent property of the input on each trial. Importantly, this is the case even when additional tokens generated are semantically meaningless "filler" tokens, rather than interpretable reasoning steps. This suggests the intermediate tokens can provide raw additional computational capacity, rather than task-specific information. These findings highlight how some errors made by LLMs on complex reasoning tasks may not always reflect a genuine lack of task-relevant competence, but rather a limitation in the depth of computation afforded by the input-dependent number of tokens generated.

A third kind of auxiliary factor is what we might call *mechanistic interference*, according to which a network learns a mechanism for solving a certain kind of problem, but the activity of that mechanism is interrupted by another, conceptually distinct process. Recent work on Transformer models trained on modular addition tasks provides insights into how a model's performance can be impacted by such auxiliary factors. Nanda et al. (2022) show that even after a Transformer has formed a generalizable circuit that implements an algorithm to solve modular addition, its test set performance can still be impeded by previously memorized input-output mappings until a later "cleanup" phase in which this rote memorization is removed. This shows that the presence of a general solution algorithm does not necessarily translate to strong test set performance when there is interference from auxiliary computations. Zhong et al. (2023) further show that even in fully trained networks that have "grokked" the task, there can be multiple concurrent algorithms or circuits that interact to solve the task and potentially interfere with each other. For instance, a Transformer might contain one circuit implementing a highly general algorithm alongside another implementing a more limited or approximate algorithm for the same task. The model's overall performance could then be negatively impacted by the causal interference of the lesser algorithm, even though the latent competence to fully solve the task is present. Taken together, these findings illustrate that a model's performance on a task can be meaningfully influenced by auxiliary computations that are conceptually distinct from the core competence required for the task.

3.2. Type-II anthropocentrism

Type-II anthropocentrism is the tendency to assume that even when LLMs achieve performance equal to or better than the average human, any substantive difference between the human strategy for solving the problem and the LLM strategy for solving the problem is, ipso facto, evidence that the LLM's solution is not general. In slogan form, the assumption says: all cognitive kinds are *human* cognitive kinds.[3] In other words, if an LLM arrives at a solution through a different computational process than humans use, Type-II anthropocentrism would lead us to conclude that the LLM's approach is not genuinely competent, regardless of how well it performs. This tendency often seems premised upon the assumption that any significant difference between human and LLM strategies for solving a problem necessarily implies that the LLM's solution is narrow or lacks generality.

However, this line of thinking risks obscuring the real capabilities of LLMs. Consider a model that learns a general algorithm for addition, rather than simply memorizing a large set of specific addition problems and their solutions. Now suppose that humans use a somewhat different algorithm for addition. The mere fact that the LLM's approach differs from the human approach does not mean that the LLM lacks real competence at addition. What matters is whether the LLM has learned a robust, generalizable strategy — not whether this strategy mirrors human cognition. Indeed, we can easily imagine an LLM outperforming humans at addition while using a distinctly unhuman-like computational process.

The key point is that competence, once abstracted from a narrowly human-centric understanding of the term, should be related to the generality and flexibility of the system's computations rather than superficial resemblance to the human cognitive architecture. An LLM that relies on a giant lookup table of memorized addition problems would not be

[3]This bias bears some similarities to what Buckner (2013) calls "anthropofabulation" in comparative psychology, though it is not equivalent. Anthropofabulation is a compound bias that results from two distinct tendencies: semantic anthropocentrism, which involves defining psychological terms (such as "theory of mind" or "episodic memory") in ways that implicitly require human-level cognitive sophistication; and the exaggeration of typical human cognitive abilities, which involves overestimating the consistency, domain-generality, and reflective nature of human cognition in everyday situations. When combined, these biases lead researchers to set unrealistically high standards for cognitive capacities, based on an idealized and often inaccurate view of human performance. In contrast, Type-II anthropocentrism neither inflates human performance nor focuses solely on the semantics of psychological terms. Instead, it assumes that the specific computational processes used by humans are necessary for genuine competence, regardless of performance outcomes. Despite these differences, both biases can lead to similar errors: setting standards for genuine cognition that no non-human system could meet, even if that system matched or exceeded normal human performance.

competent at addition, because it could not flexibly apply its memorized content-specific transitions to novel cases. But a neural network that learns a general addition algorithm could be seen as competent in the domain, even if the solution it has converged upon is distinctly unhuman-like.

3.3. An objection

We now anticipate an objection to our discussion of Type-II anthropocentrism: since the only indisputable examples of language-driven cognition we have are human, and since the capacities of LLMs are acquired through training on human-produced data, isn't it appropriate to treat human cognition as the only appropriate yardstick for studying LLMs? Our response is that it depends on how we conceive of human cognition as a yardstick. While we can – and presumably must – *begin* our investigation of LLM competencies by comparing them to our own, the role of human cognition as a benchmark can be overstated in two ways. First, the human competencies we use as reference points should be regarded as what philosopher Ali Boyle calls "investigative kinds" (Boyle, 2024) - they serve as an initial search template, but not as necessary conditions for cognitive status. Second, questions of the form "Do LLMs have cognitive competence C?" should be treated as *empirical* questions. Though this may seem self-evident, it is a principle that is easy to violate, particularly when we assume that facts about implementation are among the distinguishing features between competence A and competence B. If facts about the physical implementation of a competence are included in its definition, then it will be impossible for LLMs, which are implemented in silicon computers, to acquire *any* of the competencies that humans enjoy. Moreover, this impossibility will be logical, rather than empirical. In order to preserve the empirical character of debates about LLM capacities, therefore, we must focus on the algorithmic level of description, where facts about implementation are explicitly set aside.

The investigation of cognitive capacities in LLMs is best viewed as an iterative, cyclic process in which our conception of the relevant competencies and our understanding of the mechanisms that implement them in LLMs mutually inform and revise each other. Solving the challenge of mapping cognitive tasks to capacities often involves consideration of the underlying mechanisms (Francken et al., 2022). But to home in on the mechanisms responsible for a particular competence, we must make principled decisions about the level of abstraction at which to characterize the mechanism and about how to delineate the boundaries of the mechanism itself. These decisions, in turn, depend on how we conceptualize the cognitive phenomenon we are trying to explain. What results is a investigative process in which our characterization of cognitive tasks, our ontology of capacities, and our understanding of mechanisms evolve in tandem as we search for maximally predictive and explanatory mappings between them. In the case of LLMs, this process may lead us quite far from our initial starting point. Though we inevitably begin by searching for human-like competencies in LLMs, the gradual discovery and refinement of LLM-specific mechanisms may ultimately produce a novel ontology of cognitive kinds, one that is optimized for explaining the distinctive strengths and weaknesses of machine intelligence rather than human intelligence. In this way, the iterative nature of the investigative process allows it to drift away from its anthropocentric origins.

4. Conclusion

Like humans, LLMs can be right for the wrong reasons; we should not take good performance on various benchmarks, particularly those designed without attention to construct validity and potential confounds, at face value. However, they can also be wrong for the wrong reasons. Various auxiliary factors can interfere with their performance, such that both successes *and* failures only provide defeasible evidence about their underlying competence in a domain. Our analysis suggests that anthropocentrism, being more subtle than anthropomorphism, has garnered less theoretical scrutiny. Despite this relative neglect, we argued that anthropocentric bias can significantly impede the objective assessment of LLM capabilities. To address this gap, we presented a systematic taxonomy of anthropocentric reasoning about LLMs.

The most direct way to work out whether auxiliary factors actually cause performance degradation on a task is to look inside the model. As such, mechanistic interpretability techniques are essential for counteracting Type-I anthropocentrism. However, the role of mechanistic interpretability in addressing Type-II anthropocentrism is more nuanced. It is possible to decode a feature or circuit of interest from a large neural network even if that feature or circuit does not significantly contribute to the network's functionality (Makelov et al., 2023; Huang et al., 2023). Consequently, there is a risk of anthropomorphic projection in mechanistic interpretability research. As previously noted, we should be cautious not to combat anthropocentrism by resorting to anthropomorphism. Nevertheless, there exist mechanisms that process information in ways that differ from typical human strategies but are nonetheless robust and general. Mechanistic interpretability research can help identify these mechanisms. When such mechanisms are discovered, we should not dismiss them simply because they deviate from common human approaches.

References

Bender, E. M. On NYT Magazine on AI: Resist the Urge to be Impressed, May 2022.

Boyle, A. Disagreement & classification in comparative cognitive science. *Noûs*, n/a(n/a), 2024. ISSN 1468-0068. doi: 10.1111/nous.12480.

Buckner, C. Morgan's Canon, meet Hume's Dictum: Avoiding anthropofabulation in cross-species comparisons. *Biology & Philosophy*, 28(5):853–871, September 2013. ISSN 1572-8404. doi: 10.1007/s10539-013-9376-0.

Buckner, C. Black Boxes or Unflattering Mirrors? Comparative Bias in the Science of Machine Behaviour. *The British Journal for the Philosophy of Science*, pp. 000–000, April 2021. ISSN 0007-0882. doi: 10.1086/714960.

Chang, Y. and Bisk, Y. Language Models Need Inductive Biases to Count Inductively, May 2024.

Chomsky, N. *Aspects of the Theory of Syntax*. Cambridge, MA, USA: MIT Press, 1965.

Firestone, C. Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences*, 117(43):26562–26571, October 2020. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1905334117.

Francken, J. C., Slors, M., and Craver, C. F. Cognitive ontology and the search for neural mechanisms: Three foundational problems. *Synthese*, 200(5):378, September 2022. ISSN 1573-0964. doi: 10.1007/s11229-022-03701-2.

Hu, J. and Frank, M. C. Auxiliary task demands mask the capabilities of smaller language models, April 2024.

Huang, J., Geiger, A., D'Oosterlinck, K., Wu, Z., and Potts, C. Rigorously Assessing Natural Language Explanations of Neurons, September 2023.

Lakretz, Y., Desbordes, T., Hupkes, D., and Dehaene, S. Can Transformers Process Recursive Nested Constructions, Like Humans? In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 3226–3232, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

Lampinen, A. K. Can language models handle recursively nested grammatical structures? A case study on comparing models and humans, February 2023.

Lewis, M. and Mitchell, M. Using Counterfactual Tasks to Evaluate the Generality of Analogical Reasoning in Large Language Models, February 2024.

Makelov, A., Lange, G., and Nanda, N. Is This the Subspace You Are Looking for? An Interpretability Illusion for Subspace Activation Patching, December 2023.

Merrill, W. and Sabharwal, A. The Expressive Power of Transformers with Chain of Thought, March 2024.

Mitchell, M. Stress-Testing Large Language Models' Analogical Reasoning Abilities, May 2024.

Nanda, N., Chan, L., Lieberum, T., Smith, J., and Steinhardt, J. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*, September 2022.

Pfau, J., Merrill, W., and Bowman, S. R. Let's Think Dot by Dot: Hidden Computation in Transformer Language Models, April 2024.

Sober, E. Comparative psychology meets evolutionary biology. In Datson, L. and Mitman, G. (eds.), *Thinking with Animals: New Perspectives on Anthropomorphism*. Columbia University Press, 2005.

Webb, T., Holyoak, K. J., and Lu, H. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, pp. 1–16, July 2023. ISSN 2397-3374. doi: 10.1038/s41562-023-01659-w.

Webb, T., Holyoak, K. J., and Lu, H. Evidence from counterfactual tasks supports emergent analogical reasoning in large language models. https://arxiv.org/abs/2404.13070v1, April 2024.

Zhong, Z., Liu, Z., Tegmark, M., and Andreas, J. The Clock and the Pizza: Two Stories in Mechanistic Explanation of Neural Networks, June 2023.