Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/COLI.a.582

Squibs and Discussions

Anthropocentric bias in language model evaluation

Raphaël Millière[1], Charles Rathkopf[2]

[1] University of Oxford
raphael.milliere@philosophy.ox.ac.uk

[2] Forschungszentrum Jülich
c.rathkopf@fz-juelich.de

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 (auxiliary oversight), and dismissing LLM mechanistic strategies that differ from those of humans as not genuinely competent (mechanistic chauvinism). Mitigating these biases requires an empirical, iterative approach to mapping cognitive tasks to LLM-specific capacities and mechanisms, achieved by supplementing 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, as we lack an appropriate methodological framework to answer it. The rigorous methodology of experimental psychology (e.g., controlled experiments, counterbalanced stimuli) provides an excellent foundation for investigating the cognitive capacities of LLMs. However, we cannot transfer its paradigms whole cloth, since inferences from these paradigms in human studies rest on background assumptions about cognitive architecture that may not hold for LLMs.[1] As a result, naively applying these methods 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, conversely, entails evaluating LLMs by human standards without justification, and refusing to acknowledge genuine cognitive competence that differs in substantive ways from our own. We aim to clarify this challenge, offering a path forward for resolving disputes and advancing our understanding of LLMs' capacities.[2]

Action editor: {action editor name}. Submission received: 18 June 2025; accepted for publication: 02 August 2025.

1 See Ivanova (2025) for a related discussion.

2 We do not aim to discuss whether LLMs *should* be designed to have human-like capacities, but how to fairly assess and compare their capacities to human cognition.

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/COLI.a.582

2. The performance/competence distinction

The distinction between competence and performance (Chomsky 1965) plays a crucial role in cognitive science. Competence if often defined as the system's internal knowledge underlying a particular capacity, and performance as the observable behavior of a system exercising it (Firestone 2020). For a competent system, then, performance is the external manifestation of competence. This formulation allows for cases where performance and competence 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 this distinction to LLMs is inherently anthropomorphic, by relying on a concept of knowledge proper to humans. We can avoid this concern with an alternative formulation. 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 fair testing conditions.[3]

It is widely recognized that there is a double dissociation between performance and competence, and that this distinction 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 inferences about competence from performance are drawn in both directions in human experimental psychology, they are drawn in only one direction when studying LLMs: from bad performance to lack of competence. Without adequate justification, this asymmetry is suggestive of anthropocentric bias. We suspect that this reflects the reasonable desire to temper the opposite bias – *anthropomorphism*, the tendency to ascribe human-like capacities to LLMs without sufficient evidence.[4] 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

A pattern of reasoning is biased when it lacks adequate justification. Since justification can fail in different ways, each failure corresponds to a distinct variety of anthropocentric bias with its own methodological challenges. We outline a taxonomy of these biases and strategies to mitigate them.

3.1 Auxiliary oversight

The first kind of bias, which we call *auxiliary oversight*, is the tendency to overlook the persistent possibility that an LLM's performance failures on a task designed to measure competence C may be caused by *auxiliary factors* rather the lack of competence C. This bias arises from a failure to account for the different cognitive profiles of LLMs

3 This is meant to exclude ideal conditions where a system simply gets "lucky"—for example, where the correct answers in a test happen to be high-frequency tokens. Rather, fair testing conditions are those where cognitive mechanisms for the core competence can operate without impediment from auxiliary factors.

4 For an example of the strong anti-anthropomorphism sentiment, which we suggest may motivate this counter-reaction, see Bender (2022).

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/COLI.a.582

and humans, which can lead to flawed experimental designs in two ways. First, by providing LLMs with less support than humans for the same task, creating an unfair comparison. Second, by designing tasks with components that are trivial for humans but non-trivial for LLMs due to architectural differences. In both cases, the 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 student described above: the student's nervousness negatively impacts her performance on the test, even though she has the relevant knowledge and would have otherwise done well. Applying this concept to LLMs might seem anthropomorphic, 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.

However, auxiliary factors can and do influence LLM performance. To understand their role, it is helpful to recall that every mechanistic explanation involves a degree of idealization: it picks out a subset of the many causal influences on behavior. In doing so, it implicitly assumes that the remaining factors will either remain stable or exert only minor influence. When that assumption fails, performance may degrade in ways that do not reflect a lack of underlying competence. Sometimes, failure results from a missing enabler – a subsystem or context the subject relies on but does not receive. For example, in the mirror test for self-recognition, an animal may fail not because it lacks the capacity for self-representation, but because it lacks motivation to remove a visible mark. In other cases, failure stems from interference by competing mechanisms, as when a bilingual child struggles with a vocabulary test in one language due to lexical interference from the other. Finally, performance may be limited by how much processing the subject is permitted or able to perform – as when a student asked to solve a math problem mentally fails, but succeeds with pen and paper. These scenarios violate the idealized assumptions built into our explanatory models. Auxiliary factors are not part of the competence being tested, yet their influence can mask or suppress it. Ignoring them risks conflating failure of performance with failure of cognition.

In LLMs, we can distinguish at least three kinds of auxiliary factors. The first and most familiar kind are auxiliary task demands. Hu and 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

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/COLI.a.582

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 auxiliary task demands 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. In a similar vein, Hu et al. (2024) re-examined claims that LLMs fail at basic grammaticality judgments and found that apparent failures were an artifact of an experimental design that conflated linguistic competence with the auxiliary demand of metalinguistic reasoning. Once this demand was removed in favor of a more direct probabilistic evaluation, the models' performance was aligned with that of human subjects. These cautionary tales illustrate 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 are challenging 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 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 about analogical reasoning. Webb, Holyoak, and Lu (2023) showed that LLMs can match or surpass average human performance on various novel analogical reasoning tasks, including letter-string analogies such as $[ABC] \rightarrow [ABE], [MNO] \rightarrow [?]$. However, Lewis and Mitchell (2024) found that performance deteriorates when using a variant with a permuted alphabet. They interpret this drop as evidence that LLMs lack general competence in analogical reasoning. In response, Webb, Holyoak, and Lu (2024) argue that solving the permuted task requires counting letter indices, which LLMs cannot do effectively without tools like a Python interpreter (see Chang and Bisk 2024). Counting, they contend, imposes an auxiliary task demand that undermines the inference from poor performance to a lack of analogical competence. While humans may not assign numerical indices explicitly, they still must track *how many times* to apply successor or predecessor operations. The permuted alphabet task requires counting positions relative to the provided key—an

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/COLI.a.582

auxiliary demand that is trivial for a human with the key in view but difficult for an LLM that cannot reliably count without tools. This asymmetry in the impact of counting-related demands could explain the performance gap on counterfactual letter-string tasks.

Another kind of auxiliary factor is what we call *test-time computational bottlenecks*—variable constraints on the computations a model can perform at inference time. This is distinct from an auxiliary task demand, which imposes an additional, conceptually separate task. A computational bottleneck, by contrast, is a resource limitation on the primary task itself. This is most apparent in the context of *test-time scaling*, a family of techniques for dynamically allocating compute during inference. The simplest example involves prompting a language model to generate a "chain of thought" (CoT) before answering (Wei et al. 2023). A helpful analogy is the difference between solving a math problem with and without pen and paper. Prompting an LLM for a direct answer is loosely analogous to asking for a mental calculation, while allowing it to generate a CoT is loosely analogous to providing pen and paper for intermediate steps. When asked for a direct answer, Transformer models are fundamentally limited; theoretical work shows they cannot solve many sequential reasoning problems in a single forward pass (Merrill and Sabharwal 2024). Prompting them to "think step by step" triggers the generation of intermediate tokens that scaffold reasoning and substantially improve performance (Kojima et al. 2023). In theory, each additional decoding step expands the model's expressive power (Merrill and Sabharwal 2024). In practice, mechanistic analyses confirm that models use CoT tokens actively, drawing on multiple circuits to extract information from intermediate steps and assemble a final answer (Dutta et al. 2024). The stark difference in performance with and without CoT shows that latent capabilities may go unexpressed when models are prompted to answer directly.

The influence of test-time computational bottlenecks is demonstrated more starkly by "large reasoning models" (LRMs), such as OpenAI o1 or DeepSeek R1, which are specifically trained to benefit from test-time scaling. RLMs are language models that undergo an additional post-training stage, which typically includes supervised fine-tuning on curated examples of CoT reasoning, as well as reinforcement learning on verifiable rewards in formal domains. At inference time, the computational budget for these models can be dynamically controlled by modulating the number of "thinking" tokens they generate before providing an answer. For example, Muennighoff et al. (2025) found that increasing the "thinking" token budget would increase performance on difficult math and science benchmarks. In this case, it would be premature to conclude that performance failures with a low "thinking" token budget shows that the model lacks competence in the task domain. For the very same model and task, performance can switch from failure to success simply by letting the model generate a longer CoT.

The notion of computational bottlenecks extends beyond the generation of external intermediate tokens to scaffold computations. Experimental language models designed with an intrinsic capacity for test-time scaling *in latent space* provide an even clearer illustration of the phenomenon. For example, Geiping et al. (2025) introduce a recurrent architecture that allows the model to "think" by iteratively refining its internal state in a continuous latent space, without ever needing to verbalize intermediate steps. They found that their model's performance on reasoning benchmarks improved when they increased the number of latent iterations at test time. For the very same problem, the model might fail with a few latent iterations but succeed with dozens. This demonstrates a purely computational bottleneck, disentangled from the auxiliary demand of articulating a chain of thought in natural language. To illustrate this phenomenon, we ran the latent reasoning model from Geiping et al. (2025) on examples from the

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/COLI.a.582

Winogrande commonsense reasoning benchmark. The task involves filling in the blank in sentences like "I tried to make mini lamps by using glow sticks in mason jars, but had to get larger jars because the ___ were too big." Initially, with only a few latent steps, the model assigns a higher log-likelihood to the incorrect completion, "jars". However, as the number of latent steps increases, the log-likelihood for the correct completion, "glow sticks", increases to become the dominant prediction.[5] The model's initial failure is not evidence of its inability to solve the problem, but rather the consequence of a task-independent computational bottleneck. Its eventual success reveals a latent competence that is only expressed when sufficient computational resources are allocated at test time.

Finally, 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 algorithm does not guarantee 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 circuits that interact to solve the task and potentially interfere with each other. For instance, a Transformer might contain one circuit implementing a general algorithm alongside another implementing a more heuristic 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 solve the task is present.

The distinction between mechanistic interference and auxiliary task demand lies in the source of the failure: a task demand is an extrinsic feature of the experimental setup, while mechanistic interference is an intrinsic property of the model's internal processing, where two or more circuits produce conflicting outputs. One might object that this is an overly charitable interpretation; perhaps the interference simply reveals that the model's competence is fragile. However, the mechanistic separability of the competent circuit from the interfering process suggests otherwise. The point is not that the model is infallible, but that a general, competent mechanism exists and its function can be isolated, even if its output is sometimes suppressed. These findings illustrate that a model's performance on a task can be disrupted by auxiliary computations that are conceptually distinct from the core competence required for the task.

3.2 Mechanistic chauvinism

Mechanistic chauvinism is the tendency to assume that even when LLMs match or exceed average human performance, any substantive difference in strategy counts as evidence that the model's solution lacks generality. In slogan form: all cognitive kinds are *human* cognitive kinds.[6] Whenever an LLM arrives at a solution through a different

5 Figure available at `https://github.com/raphael-milliere/latent-test-time-scaling`.

6 This bias resembles what Buckner (2013) calls "anthropofabulation"—a combination of (i) human-centric definitions of psychological terms with (ii) an exaggeration of human cognitive abilities. Mechanistic chauvinism differs from anthropofabulation in that it focuses on process rather than semantics or

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/COLI.a.582

computational process than humans use, the mechanistic chauvinist concludes that the LLM lacks genuine competence, regardless of how well it performs.

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. Indeed, we can easily imagine an LLM outperforming humans at addition while using a distinctly unhuman-like computational process.This concern is arguably reflected in ongoing debates about the linguistic capacities of LLMs. Current models are not only proficient at generating grammatically well-formed sentences, but can also match human performance on judgments of syntactic acceptability, often aligning closely with human intuitions (Hu et al. 2024). Yet, many linguists in the generative tradition, would argue that LLMs lack genuine syntactic competence (Chomsky 1965; Chomsky, Roberts, and Watumull 2023; Everaert et al. 2015). From this perspective, their success stems not from an internal knowledge of hierarchical syntactic rules, but from their nature as massively scaled statistical models trained to predict the next word. This conclusion, however, is challenged by a growing body of mechanistic evidence suggesting that LLMs do learn and represent hierarchical syntax internally, even if the computations that produce these representations differ from those posited in formal linguistic theories (Millière forthcoming). This debate thus exemplifies mechanistic chauvinism: if competence is defined by the instantiation of a specific computational architecture hypothesized in humans, then LLMs are deemed incompetent by fiat—thereby dismissing both their behavioral success and the alternative, learned mechanisms that support it.

3.3 An objection

We now anticipate an objection to our discussion of mechanistic chauvinism: since the only indisputable examples of language-driven cognition are human, and since the capacities of LLMs are acquired through training on human-produced data, isn't human cognition the most appropriate yardstick for studying LLMs? Our response is that it depends on how the yardstick is used. 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. To preserve the empirical character of debates about the capacities of LLMs, we must focus on the appropriate level of analysis. Following Marr's classic framework, this means focusing on the computational and algorithmic levels, where we ask both *what* the system computes and *which* algorithms enable it to

performance. Nevertheless, both forms of bias risk setting standards that no non-human system could meet.

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/COLI.a.582

perform that computation. Defining competence at the physical or implementational level—for example, by requiring a biological substrate—would make the question of whether an LLM could possess a human-like competence non-empirical by definition. Therefore, our investigation explicitly sets aside facts about physical implementation.

The investigation of cognitive capacities in LLMs is best viewed as an iterative process in which our conception of LLM competencies, and of the mechanisms that implement them, mutually inform and revise each other. Solving the challenge of mapping cognitive tasks to capacities often involves consideration of the underlying mechanisms (Francken, Slors, and Craver 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 an investigative process in which our characterization of cognitive tasks, our ontology of capacities, and our understanding of mechanisms evolve in tandem. In the case of LLMs, this process may lead us quite far from our initial starting point, and may ultimately produce a novel ontology of cognitive kinds, 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. However, they can also be wrong for the wrong reasons. Auxiliary factors can suppress performance in ways that obscure latent competence. While anthropomorphism has received substantial critical attention, anthropocentric bias has gone largely unexamined. We have argued that it poses a serious obstacle to the objective assessment of LLM capacities, and we offered a taxonomy of its forms to help guide future research.

While carefully designed behavioral experiments are the primary means of identifying performance failures, mechanistic interpretability techniques are uniquely suited to adjudicate between competing hypotheses about the causes of these failures. For instance, when behavioral evidence is ambiguous, intepretability methods can, in principle, help distinguish whether a failure is due to a lack of a competent circuit or due to an auxiliary factor (such as mechanistic interference) suppressing an existing competent circuit. As such, supplementing behavioral with mechanistic evidence is useful way to counteract auxiliary oversight.

The role of mechanistic interpretability in addressing mechanistic chauvinism 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. 2024; Huang et al. 2023). So there is a risk of anthropomorphic projection in mechanistic interpretability research. Nevertheless, there are examples of robust information processing mechanisms that differ from the ones humans typically use. Mechanistic interpretability research can help identify these mechanisms. When they are discovered, we should not dismiss them simply because they deviate from the human case. Doing so would foreclose the possibility of understanding artificial cognition on its own terms.

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/COLI.a.582

References

Bender, Emily M. 2022. On NYT Magazine on AI: Resist the Urge to be Impressed.
Boyle, Alexandria. 2024. Disagreement & classification in comparative cognitive science. *Noûs*, n/a(n/a).
Buckner, Cameron. 2013. Morgan's Canon, meet Hume's Dictum: Avoiding anthropofabulation in cross-species comparisons. *Biology & Philosophy*, 28(5):853–871.
Buckner, Cameron. 2021. Black Boxes or Unflattering Mirrors? Comparative Bias in the Science of Machine Behaviour. *The British Journal for the Philosophy of Science*, pages 000–000.
Chang, Yingshan and Yonatan Bisk. 2024. Language Models Need Inductive Biases to Count Inductively.
Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA, USA: MIT Press.
Chomsky, Noam, Ian Roberts, and Jeffrey Watumull. 2023. Noam Chomsky: The False Promise of ChatGPT. *The New York Times*.
Dutta, Subhabrata, Joykirat Singh, Soumen Chakrabarti, and Tanmoy Chakraborty. 2024. How to think step-by-step: A mechanistic understanding of chain-of-thought reasoning.
Everaert, Martin B. H., Marinus A. C. Huybregts, Noam Chomsky, Robert C. Berwick, and Johan J. Bolhuis. 2015. Structures, Not Strings: Linguistics as Part of the Cognitive Sciences. *Trends in Cognitive Sciences*, 19(12):729–743.
Firestone, Chaz. 2020. Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences*, 117(43):26562–26571.
Francken, Jolien C., Marc Slors, and Carl F. Craver. 2022. Cognitive ontology and the search for neural mechanisms: Three foundational problems. *Synthese*, 200(5):378.
Geiping, Jonas, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R. Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. 2025. Scaling up Test-Time Compute with Latent Reasoning: A Recurrent Depth Approach.
Hu, Jennifer and Michael C. Frank. 2024. Auxiliary task demands mask the capabilities of smaller language models.
Hu, Jennifer, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. 2024. Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences*, 121(36):e2400917121.
Huang, Jing, Atticus Geiger, Karel D'Oosterlinck, Zhengxuan Wu, and Christopher Potts. 2023. Rigorously Assessing Natural Language Explanations of Neurons.
Ivanova, Anna A. 2025. How to evaluate the cognitive abilities of LLMs. *Nature Human Behaviour*, 9(2):230–233.
Kojima, Takeshi, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large Language Models are Zero-Shot Reasoners.
Lakretz, Yair, Théo Desbordes, Dieuwke Hupkes, and Stanislas Dehaene. 2022. Can Transformers Process Recursive Nested Constructions, Like Humans? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3226–3232, International Committee on Computational Linguistics, Gyeongju, Republic of Korea.
Lampinen, Andrew Kyle. 2023. Can language models handle recursively nested grammatical structures? A case study on comparing models and humans.
Lewis, Martha and Melanie Mitchell. 2024. Using Counterfactual Tasks to Evaluate the Generality of Analogical Reasoning in Large Language Models.
Makelov, Aleksandar, Georg Lange, Atticus Geiger, and Neel Nanda. 2024. Is This the Subspace You Are Looking for? An Interpretability Illusion for Subspace Activation Patching. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, OpenReview.net.
Merrill, William and Ashish Sabharwal. 2024. The Expressive Power of Transformers with Chain of Thought.
Millière, Raphaël. forthcoming. Language Models as Models of Language. In Ryan Nefdt, Gabe Dupre, and Kate Stanton, editors, *The Oxford Handbook of the Philosophy of Linguistics*. Oxford University Press, Oxford.
Muennighoff, Niklas, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. S1: Simple test-time scaling.
Nanda, Neel, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2022. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*.

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/COLI.a.582

Sober, Elliott. 2005. Comparative psychology meets evolutionary biology. In Lorraine Datson and Gregg Mitman, editors, *Thinking with Animals: New Perspectives on Anthropomorphism*. Columbia University Press.

Webb, Taylor, Keith J. Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, pages 1–16.

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

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.

Zhong, Ziqian, Ziming Liu, Max Tegmark, and Jacob Andreas. 2023. The Clock and the Pizza: Two Stories in Mechanistic Explanation of Neural Networks.