Intentionality All-Stars Redux: Do language models know what they are talking about?

Abstract: The rise of large language models (LLMs) has brought with it a robust debate about whether these machines know what they are talking about, or whether they are just bullshitting. Answering this question requires wrestling with the problem of intentionality: what are the conditions for our thoughts, beliefs, and judgments to be *about* the world? How can we differentiate a model that says the right thing from one that understands what it is saying? This article addresses the intentionality of LLMs by revisiting Haugeland's storied team, "the Intentionality All-Stars." I lay out four major approaches to intentionality that are regularly applied to explain LLMs behavior—as well as their failures. Spelling out the different positions, in turn, shows that there are multiple conditions that must be met for an LLM to know what it is talking about—and the numerous ways current models fall short.

Keywords: Intentionality, Mind, Language Models, AI, Haugeland, Objectivity

> "Understanding pertains not primarily to symbols or rules for manipulating them, but to the world and living in it."
>
> John Haugeland (1998: 39)[1]

The prevalence of large language models (LLMs) has given rise to now well-worn debates in academic circles and the popular press about whether these impressive machines know what they are talking about—that is, whether there is any understanding in the machine (Aguera y Arcas 2022), or whether it is just an empty chatterbox (Bender et al. 2021). These debates are not unwarranted: LLMs demonstrate a stunning ability to generate plausible, human-like responses to linguistic inputs, such as conversation openers or essay prompts, and produce often subtle or insightful connections likely to have been overlooked by the casual observer. This evidence has suggested to one engineer that LLMs possess a *mind*—and thus deserves legal rights (Grant and Metz 2022).

[1] Jacob Browning, Visiting Scientist, NYU. Email: browning.jake@gmail.com
Acknowledgements: This paper benefitted from many comments and criticisms, including from a talk at the AI and Humanity Lab and a workshop from the Santa Fe Institute. In addition, I would like to thank helpful comments from Zed Adams, Thomas Cantone, Adam Gies, Joseph Lemelin, Kevin Temple, and Philippe Verreault-Julien.

How would we know if this were true? Dennett warns in a recent piece that humans have a "natural inclination to treat anything that seems to talk sensibly with us as a person—adopting what I have called the 'intentional stance'" (Dennett 2023; see also Mahowald et al. 2024). This natural inclination, he warns, makes us "sitting ducks" for speaking machines. Given how LLMs are trained, this is an especially acute worry. They are pretrained to repeat human-made sentences one word at a time and corrected until they have learned the general statistical structure of how humans speak and write. This extensive pretraining results in the models memorizing numerous passages, puzzles, and tests, and coming up with the general gist for countless others. The model is then fine-tuned to be more personable, attempting to respond with accuracy, earnestness, decency, and relevancy (Ziegler et al. 2019). This training grants them immense generative power, allowing them to come up with responses broadly in-line with what a human might write. But this approach also raises difficult questions about whether the model actually knows what it is talking about or simply regurgitates (with a bit of remixing) the comments of others. Benchmarks and tests only help so much because the model often memorizes the test and answers in pretraining. We are often stuck wondering if the model is genuinely showing sparks of intelligence or just gaming our metrics (Mitchell 2023; Frank 2023).

In the last few years, scholars have attempted to settle the issue by appealing to an enormous amount of data compiled on LLMs, such as ChatGPT, Claude, and Gemini. But data alone cannot settle the issue. The underlying problem is philosophical not empirical. What does it mean to have a mind, one that understands what it is talking about? We need some sense of how to address this question before we can make sense of LLMs.

This paper is an attempt to provide clarity to the issue by adopting Kant's reverse-engineering approach: what are the conditions are necessary for mindedness? More precisely, Kant asks what is necessary for our representations to be *about* the world—what makes it possible for our percepts, thoughts, beliefs, and desires to latch onto something outside them (Kant 1781/1991: A387)? This is *the problem of intentionality*, which concerns when it is it appropriate to ascribe intentional states—like beliefs, desires, concepts, and so on—to a system. And, as Kant argues, for our thoughts to be about the world, there are a number of different, normative conditions that

must be met—specifically, the thoughts must have the right form, must be applied in the right way, must represent an objective world, and must be unified in a self. This provides a useful framework for answering the question: does an LLM know what it is talking about? Answering this question requires getting clear on what makes it that case that a language-using system *could* be talking about the world.

While Kant's account is highly intellectualist, the same conditions Kant points to are visible even in Dennett's familiar, more minimal intentional stance. For Dennett, we ascribe a mind to a system if it is explanatory and predictive to do so, and it is explanatory and predictive to do so if the system is a goal-directed, purposive agent—what Dennett calls "rationality" (Dennett 1971). The goal-directedness provides a normative standard for us to determine whether the model is making *the right* moves based on its beliefs and desires. And making the right moves depends on the beliefs acting like beliefs (formal condition), the system leveraging its beliefs to achieve its goal (the rational condition), the system accurately representing the world (objective condition), and possessing some consistency over time (self-identity condition). As Dennett notes, the intentional stance applies even to a thermostat, where the box on the wall updates its temperature properly to achieve its goal of keeping the room cool, and a person, who needs to navigate the campus to their next class. But it fails to work if any part goes missing—if the thermostat makes the temperature hot in response to cold, or fails to detect temperature, or simply falls off the wall and breaks, the intentional stance becomes unworkable. Ascribing a mind and mental states is *holistic*—and thus failures in one part will imply failures in other parts.

To make sense of whether LLMs have minds—as well as why there are so many takes on the issue—I will make use of an old analogy from John Haugeland: *baseball* (Haugeland 1990). Haugeland uses the metaphor of baseball to lay out an imaginary team of *intentionality all-stars*, which provides the major theories for treating a system as possessing a mind. He does so by addressing the four positions in the infield (first- and second- base, shortstop, and finally third-base). Haugeland arranges them roughly by the degree of plausibility he ascribes to each, with first-base being the position he is most critical of, while third represents coming closest to his own view. I will follow this method here, laying out different views according to my own sense

of their plausibility, laying out criticisms as I go. These positions will also highlight challenges current LLMs face dealing with Kant's conditions.

This paper consists of five sections. In section one, I lay out what is called "the problem of intentionality." This problem addresses what makes it appropriate to ascribe intentionality—that is, ascribe beliefs, desires, and other mental states—to a system, like a person or LLM. In section two, I describe the first base position, which holds LLMs struggle with the problem of formal competency with manipulating symbols in rule-governed ways—an essential condition for possessing beliefs. In section three, I lay out the second base position, which focuses on the problem of rationality, concerning whether LLMs responses *make sense* all things considered—whether they suggest good responses, or instead just seem like empty banter. In section four, I present the shortstop position, which concerns the problem of objectivity, distinguishing whether the models grasp their comments are about an objective, language-independent world. In section five, I articulate the third base position, which focuses on the problem of self and whether the model behaves in a consistent, unified manner. The upshot of going through the bases is insight into what it takes for something to be a mind—and where and why current LLMs fall short.

1. The Problem of Intentionality

'Intentionality' describes the peculiar way representations manage to be *about* something else. How can our thoughts and beliefs, in other words, be about something or other and not nothing at all? Why is it that my claim 'the Yankees are overrated' is a claim *about* the Yankees, and not the Mets? This problem is essential for thinking about LLMs because we want some way of understanding not just whether their sentences refer to the same things we do, but also whether they *understand* what they are talking about: do they believe what they are saying? Do they want to help us and answer our questions? Or are they just spitting out words?

The usual starting point for thinking about intentionality is by looking at human-made representations, like photographs, maps, language, and secret signals. These representations are all *derivative* representations: they mean something because humans imbue them with meaning. A funny hand gesture means "strike" because of a shared convention between a catcher and

pitcher. An obscene drawing represents Peter Angelos because someone drew it that way. Like mental states, human-made representations can be accurate or inaccurate—and it often matters which is which. This is especially important for thinking about the linguistic representations at work in LLMs. Language forms a complex system of symbols that allows for infinite novelty by reusing the same symbols according to the same rules in infinitely many ways. For example, the meaning of, "the Yankees win the World Series!," is a result of the meaning of each individual word in that particular order (called "compositionality").

The problem of intentionality is focused on how mental states—especially beliefs and other propositional attitudes—get their meaning. Beliefs are important in language users especially because the kind of things we can say parallel to the kind of things we can think in a proposition, and many of the same properties—infinite novelty out of different symbols modified according to specific rules—applies to both (Fodor 1975). Beliefs are also normative: they *should* be true and, if they are not, they are wrong. Thus, when we are asking whether the model knows what it is talking about, we are curious about whether it *believes* what it is saying. We can also, using that metric, evaluate whether it believes something false.

There are a few different conditions that must be met for something to count as an intentional system—that is, a mind with mental states about the world.

1. The formal condition: for something to be a belief, we expect it to fall in line with certain normative constraints, like basic logical laws or rules of inference, that hold regardless of the content of the belief. For example, an intentional system should not believe both *x* and *not-x*; it should not respond to "if *a*, then *b*" and "a" with "not-*b*"; it should be able to represent "*Fp"* and "*Gq*" if it can already represent "*Fq and Gp,*" and so on. This condition ensures a model possesses beliefs and is not just saying things.
2. The rational condition: to determine what beliefs and desires are appropriate to a system, we need to have some sense of what it is trying to *do.* A robot making random moves—even inside a baseball stadium—does not mean it is playing baseball. It is only when it is making *legal* moves at the right time, in the right context, and those moves might result in winning that it counts as playing baseball. And it is only in the latter case—where we

specify the goal of "winning the game"—that talk of beliefs and desires pays off for making sense of the machine is doing.

3. The objective condition: beliefs are what a system takes to be true about the world, and so beliefs depend ultimately on some non-linguistic referent. If this is missing, we cannot say the words or corresponding beliefs have intentionality—that they are *about* something else. In cases like this, we run into familiar worries about solipsism and private languages (Esnau 2024), where it becomes impossible to make sense of error.
4. The self condition: there needs to be some coherent, persisting self to ascribe beliefs and desires to, even if we are just talking about the persisting thermostat box on the wall. If there is no persisting body, no shared memory, no stable goal pursued, or any other way of distinguishing the entity, ascribing beliefs and desires becomes empty.

It is only when all of these are met that we are talking about an intentional system—and, thus, a system that might know what it is talking about.

A simple example shows all four: if Sam really wants to see the Mets and believes I'll take her if she gets good grades, then she should adopt a whole host of behaviors at different times and places—arrive early at school, do her homework diligently, ask for help, and so on—to achieve her goals. Sam's behavior, in these cases, strikes us as *rational*: she behaves as she does because she wants to go to the game, and she chooses her actions because she believes they are likely to result in the outcome she wants. Her mental states count as beliefs because they follow the proper formal rules governing them, such as updating beliefs correctly in light of new information—for example, if Sam believes she did well on a math test but gets a bad grade, she updates her other beliefs about competency at math and seeks out tutoring or, instead—abandons her goal of good grades or renegotiates. And the beliefs all depend on a stable self and an objective world: her behavior depends on the belief that the world will exist in eight months, that she will still want to see the Mets, and that her uncle will keep his word. When these conditions are met, we are inclined to ascribe a mind, one that understands its world. While Sam may make mistakes and do foolish things—or even get sick of baseball—we treat her as "in her right mind" so long as we can follow the logic of her actions in terms of beliefs and desires.

Before we jump into the different positions on intentionality, we need to disambiguate the problem of intentionality from some closely related problems that LLMs also face. First, the question of consciousness. There are plenty of arguments in favor of AI consciousness (e.g., Butlin et al. 2023; Chalmers 2023), and arguments against (Chirimuuta 2025). But it is only a minority view that intentionality *depends* on consciousness (Searle 1980), and even Kant (who talks a lot about consciousness) is often read as concerned with epistemology rather than consciousness (e.g., Cohen 1885). Second, the "sensory grounding problem" (Harnad 1990). This is the problem of gathering whether a system trained solely on words would use words to refer to the same things humans refer to, or whether they need to have some "sensory grounding" to mean "blue" when they say blue. The typical response of many philosophers is to argue for the *externalist* view from the philosophy of language, in which the meaning of a word is not in the head but a matter of what it refers to in the world (Putnam 1975). As such, meaning is independent of the thoughts of language users: the meaning of "knuckleball" does not depend on what I think, but what particular type of pitch has been historically-linked to that particular name. Externalism provides a plausible response to LLMs because we can simply assert, as Mandelkern and Linzen (2023) do, that LLMs using the same words as humans are referring to the same things. This is also how many of us use these systems: we adopt an "interpreter-focused" metasemantics, where the meaning of an LLMs sentence is determined by what *we* take it to be saying (Cappeln and Dever 2021). On these accounts, the meaning of the words depends not on what is happening in the LLM, but instead on what the interpreter hears and the general facts about the world.

But the question of whether the words mean the same thing is of secondary concern for us in this paper. The problem of intentionality is concerned with whether ascribing a mind and mental representations mean what they *should* mean. This is a normative standard: we want to be certain that the machine understands that there is an objective world, that the words affect other people, and that asserting one thing means they should deny some other thing. In this context, it is not sufficient that the words often appear to us as meaningful; we want assurances that the words mean roughly the same thing, that the LLM shares our concepts—and, with it, our world. The goal is to find assurances that when the model says "Yankees in five," that they believe the

Yankees will win in five games, and that their concepts of "Yankees" and "games" is the same as ours.

What we want to avoid is the "ELIZA Effect" documents by Joseph Weizenbaum (1976). The simple program ELIZA is infamous for behaving like a Rogerian therapist, such that if I said, "I am frustrated by how well the Diamondbacks are doing," ELIZA would respond, "why are you frustrated by how well the Diamondbacks are doing?" But ELIZA was a trick: if just rearranged statements into why questions—or, if stumped—ask about the patient's relationship with their mother. It was not designed to understand what it was doing but only designed to deceive the unwary. And it succeeded; some people found it as meaningful—and even therapeutic (Weizenbaum 1976). From an interpreter-focused perspective, the words made sense and referred to the same thing as what the patient was talking about. But ELIZA, by design, does not have concepts corresponding to what it is saying. Its goal is to prevent us from getting ahead of ourselves, to provide a warning that we have bad intuitions about speaking beings—and, as Dennett warns, are sitting ducks for being deceived.

We now have a sense of the problem: do LLMs believe what they are saying? When they talk about the world, do they know what they are talking about? The next four sections will provide different answers to these questions and, in the process, show why understanding language is such a difficult challenge.

2. First Base

The first base position focuses most on the formal condition. Some researchers at first base are the most hostile towards regarding LLMs as cognitive at all, though others are simply more skeptical they are ever going to be that impressive. The basic complaint is that the impressive performance of LLMs obscures their underlying incompetence: although they often say the right thing, their underlying formal machinery is all wrong.

The basic sensibilities of this position come from defenders of the older paradigm of cognitive science and AI—a position typically called "cognitivism." This approach embraces the strict distinction between *syntax*—the formal properties of the system—and *semantics*—the meaning of the sentences (Haugeland 1985). Typically, syntax refers to the underlying mechanical rules for processing symbols, which often correspond to the rules governing a symbolic logic for binding one symbol to another, the various operations connecting symbols, and the various transitions between symbols (Marcus 2001). Semantics, by contrast, discusses what the symbols or groups of symbols refer to, whatever they are about. They contend that human thoughts consist in the appropriate conjunction of these two components, the syntactic, linguistic "form" and the meaningful "content." The traditional cognitivist position held that the meaning of words depends on the meaning of inner mental sentences—effectively, the propositional attitudes that make up our thoughts, beliefs, hopes, and dreams (Fodor 1971). And, as Fodor (1978) argues, for a system to possess propositional attitudes, there are a number of formal, hard to fake conditions they need to possess.

For first base, a useful model of this language of thought is the standard computer: discrete symbols kept in a memory store, processed according to basic rules. This mechanical system makes it possible to process symbols according to rules, and if each symbol *represents* some concept, and if the syntactical rules *represent* logical rules, then we should be able to interpret the overall system as a *semantic engine*. From the computer's perspective, it is just number-crunching in rule-governed ways but, from our perspective, the machine is saying something plausible, taking the right actions based on it, and making the proper inferences all things considered. But meeting the formal condition is a high bar, and the first base position warns we cannot *assume* that a system that says the right thing necessarily has the proper underlying machinery. Their performance might not stem from an underlying competence, but instead merely from memorizing plausible statements (see Pavlick 2022; Harding and Sharadin *forthcoming*). Assessing competence demands careful probing because there is a difference between saying the right thing and being able to derive all the proper inferences from the claim. If someone claims Adrien Gonzalez retired, they should also know they are not playing anymore, and if they are not playing anymore, then they will not see them playing today. Counting a belief *as* a belief demands it play a complex functional role in an intentional system: it involves the

combination of different concepts in a proposition, where the meaning of the proposition depends not just on the concepts involved but also the specific way they are combined. If the proposition is also believed, then the system is committing to believing a whole host of other propositions—and denying many others. These are formal matters; it does not matter what the beliefs are about, only that they behave in the right way.

Many from the traditional perspective have been extremely hostile towards LLMs (e.g., Chomsky et al. 2023). The main complaint is that LLMs do not get the *syntax* right—and, as such, the appearance of linguistic competency, much less understanding, is simply an illusion. Although we might be deceived into interpreting them as competent language users, first base highlights how their language systematically deviates in cases where getting the syntax right matters. For example, Tang et al. (2023) show that when the statistics of language point in one direction and the underlying logic of the sentences in another, LLMs veer towards statistics over syntax (see also Berglund et al 2023 and Dentella et al. 2023). Pavlick et al (2023) show that these systems consistently fail on an essential feature of syntax: variable-binding. Variable-binding involves attributing the right property or relation to the right object. (The failure is easy to see using generative image models: just ask DALL-E to show you an image of "a lefty Orioles pitcher throwing a sinker to a right-handed Mets batter.") The failures of variable-binding bleed into competency in other areas: Dziri et al. (2023) highlight that LLMs often fail at compositionality tasks, with the result that they fail on downstream planning and reasoning tasks dependent on putting together representations correctly. There is also extensive research on the challenges LLMs face with performing the right logical and mathematical inferences (Mirzadeh et al. 2024), something traditional computational models excel at. Even the newest models, for example, still cannot reliably parse negation (Garcia-Ferrero et al. 2024). The first base position argues features like variable-binding, compositionality, and rule-following are *fundamental* to human thought, and so the failures of LLMs on these tasks should be seen as clear evidence that they are not thinking.

But the cognitivist position has its own issues. There are many critics of traditional conceptions of language like Chomsky's (Piantadossi 2023) and the idea that language is principally for thought and only secondarily for communication (Federenko et al. 2024). The assumption that

the brain operates using a language of thought is also contentious; although it still has defenders, it also has plenty of critics (see Quilty-Dunn et al. 2023 and commentaries). We also need to be on guard against "anthropofabulation" (Buckner 2013), where we exaggerate human abilities in order to distinguish ourselves from animals and machines: humans tend to struggle at similar tasks as LLMs (Buckner 2023). Moreover, there has also been some progress in developing neural networks that more adequately capture some of the syntactical abilities cognitivists point to (e.g., Lake and Baroni 2023; Santoro et al. 2022). The challenges first base highlight about LLMs are substantial, but not necessarily insurmountable, nor disqualifying them from being cognitive systems.

The upshot is that, on the whole, first base provides a useful critique of formal failures of LLMs. But it is also unnecessarily dogmatic; failing in many cases does not show the model lacks the underlying competence (Harding and Sharadin *forthcoming*). It may just stem from contingent performance problems, such as failures of memory. We should probably acknowledge these formal problems, but leave it an open problem whether they are insurmountable for LLMs.

3. Second Base

The first base position has its strongest advocates from the traditional, cognitivist picture. The second base position, by contrast, is often made up of neuroconnectionists (Doerig et al. 2023). For the second base position, ascribing rationality to a system is not about the underlying architecture of the system. Rather, ascribing rationality is appropriate if it is *behaving* rationally—that it acts like it has beliefs, desires, and some goal. For a speaking agent, we adopt a "principle of charity" (Davidson 1973), one where we assume the agent has mostly true beliefs and is broadly conforming to Gricean norms of honesty, decency, and relevance (Grice 1975).

First base assumes that original intentionality is in the mind, with language inheriting its meaning from mentalese. Counterintuitively, many at second base reverse this: it is because language is meaningful that we have meaningful thoughts. For many empiricists (e.g., Quine 1960), original intentionality is found in the social community and its use of language, and infants then acquire a "derivative" intentionality—one dependent on learning how others talk and

think. A classic version of this kind of account—one that influenced many empiricists, such as Churchland 1996 and Gauker 2011—is Wilfrid Sellars (1956). For Sellars, learning a language is the primary source of meaning, one that transforms the blooming, buzzing confusion of the world into, "a structured logical space [. . .] in a world of physical objects, colored, producing sounds, existing in Space and Time" (1956: S30). Sellars argues that the "meaning" of our terms is not primarily learned through experience, but rather through the role specific words play in our practices. He writes, "The rubric '"—" means –' is a linguistic device for conveying the information that a mentioned word, in this case 'rot,' plays the same role in a certain linguistic economy [. . .] as does the word 'red'" (1956: S31). On Sellars's account, someone could come to infer what the word "out" means in baseball solely by listening to the announcer on the radio and figuring out the role of "out" alongside the role of other words, like "batter," "pitcher," and "strikes." Over time, listeners will acquire the appropriate, interconnected concepts, where each concept hangs together with all the others in a cohesive scheme. They can then acquire a "theory of mind," and the appropriate ascription of beliefs and desires, as a means of predicting and explaining behavior. Effectively, learning language shapes us into minded creatures (Haugeland 1990), with all the cognitive sophistication involved in that (Lupyan 2016).

Sellars's view is a version of "conceptual role semantics" in the philosophy of mind, where each concept is determined by its connection to all the others (Cummins 1989). It is also a close analogue to the view in linguistics known as "distributional semantics," where the meaning of "a word is characterized by the company it keeps" (Firth 1957). For both theories, the meaning of any word or concept is a holistic matter, one we infer only in light of the whole. Manning (2020), Pavlick (2023), and Piantadossi and Hill (2023) argue these approaches are useful analogues for thinking about LLMs: these system learn the meaning of words through its usage in countless different sentences, rather than grasped through any sensory experiences. And LLMs have been shown to acquire conceptual schemes for sensory concepts like color this way (Patel and Pavlick 2022), with the result that they can infer the location of an unknown color relative to its placement among other colors. This suggests that sensory grounding is not necessary for words to be meaningful (Søgaard 2023; Coehlo Mollo and Milliere 2023). All that is necessary, for conceptual role semantics, is that the model grasps—or, at least, could make explicit if prompted (Brandom 1994)—the numerous inferential implications our various assertions commit us to. On

this view, the focus of our research should be figuring out what the models believe (Chalmers 2025), rather than presuming they cannot believe because of their design or training.

But this has proven a high bar: LLMs often do not behave as if they possessed mostly true beliefs, instead stating surprising falsehoods, absurd suggestions, and just generally being confused about how the world works. As such, it has proven unwise to assume the system is rational; even the simple task of taking orders at McDonald's proved too difficult for LLMs (Tangermann 2024). The difficulty can be seen as stemming from three issues: first, as seen in the first base position, the models struggle with the formal features of language and logic; second, as we will see at shortstop, the models are often incompetent at real-world reasoning or planning; third, it is unclear if mental state talk—beliefs, desires, understanding, and so on—is a useful way to approach these systems.

Beliefs, by their nature, are fundamentally normative states: they are supposed to be true, and they are supposed to have lots of implications for other beliefs (though a system need not be aware of *all* the implications). But, as Yiu et al. (2023) note, "nothing in [LLMs] training or objective functions is designed to fulfill the epistemic functions of truth-seeking systems" (2023). Although the system possesses a conceptual scheme from language, they lack what traditional cognitive science often dubbed "the belief box"—the mythical place in the mind that encodes the propositions we assent to and those we reject. Is it possible to ascribe "mostly true beliefs" without this, or does the principle of charity fall apart? Or should we treat them, as Andreas (2022) puts it, as "incoherent encyclopedias"? Lederman and Mahowald (2024) suggest we should ascribe beliefs because this provides a useful way of understanding how they succeed on tasks, answer difficult questions, and solve puzzles. But, they acknowledge, since LLMs make many errors, we should probably explain their answers in terms of non-mental talk of training data and objective function, similarly to how we explain shivering by biology (2023: 1096). But this solution seems troubled: if a model keeps making mistakes on our interpretation, then that speaks against our interpretation.

A different option is Frankish (2024), who suggests we can ascribe beliefs but no goals or desires besides playing the "chat game," a one-player game the LLM is playing with itself to provide

plausible-seeming responses to any comment. But, he contends, the errors may push us to assuming the LLMs beliefs are not about *our* world, but about "linguistic items" within the world of the chat game (2024: 70; see also McCoy et al. 2023). But this interpretation makes it impossible to discern whether a model is expressing a false belief in our world or a true belief in the "linguistic" world (or, even more confusingly, a false claim in the linguistic world, or intentionally misleading us, or any host of other possibility). Mental state talk becomes too loose to provide any traction, and it is hard to avoid the impression the "chat game" is just words frictionlessly spinning in the void (Esnau 2024). For these reasons, many people have abandoned belief-talk in favor of treating the models as necessarily "bullshitting" machines (Hicks et al. 2024). The philosophical notion of bullshit concerns talk that is indifferent to truth—an apt description of systems that follow the statistics of language rather than facts about the world.

The underlying challenge of the second base position is that it expects an enormous amount of knowledge to be conveyed simply by learning the everyday usage of language: not just knowledge about what words mean and common-sense data, but also the formal structure of logic, language, and mathematics, the ontological structure of reality, and how to distinguish truth from falsity. These goals have proven, at present, overly ambitious—and it is unclear how further pretraining or fine-tuning on autoregressive LLMs might solve these problems.

4. Shortstop

The shortstop position shares with second base the sense that the model is accomplishing impressive things linguistically. But the shortstop develops their position around the persistent *failures* of these models: why are LLMs concepts so shallow, their beliefs so wonky, their behavior so unreliable? For shortstop, the basic issue is that the linguistic competency of these machines hides a kind of *cognitive incompetency,* one stemming from a failure to recognize that the words correspond to an independent, objective world.

Researchers at the shortstop position often focus on domain-specific, non-linguistic capacities seen as essential for helping us reasoning through problems. For example, Mahowald et al. (2023) argue these models possess the linguistic competency necessary for talking intelligently

about all kinds of things but lack "multiple extralinguistic capacities that comprise human thought, such as formal reasoning, world knowledge, situation modeling, and social cognition" (2023: 1). They point to results showing that human problem-solving in these domains often does not rely on the language areas of the brain (Federenko 2024). Similar extralinguistic capacities are also found in animals and infants (Stojnic et al. 2023), suggesting these capacities are broadly distributed in nature. While there are disputes about whether these are innate (Carey 2009) or learned (Buckner 2023), the underlying assumption is that certain non-linguistic functions are essential and, for LLMs to genuinely behave as well as even a cat (LeCun 2022), they will need similar extralinguistic competencies. Without these capacities, these researchers contend they will be stuck with a poor imitation—simply grasping at linguistic features rather than grasping the underlying rules governing the objective world.

The shortstop problem is often gestured at by denying LLMs possess a "world-model." Although the term means many things to many people, a useful approach is from Yilirim and Paul (2024: 405), who write that a world-model contains "structure-preserving, behaviorally efficacious representations of the entities and processes in the real world, including objects with 3D shapes and physical properties, scenes with topological relations and navigable surfaces, and agents with beliefs and desires" (Yilirim and Paul 2024: 405). The concern with world-models is whether the LLM possesses both the highest-level concepts—space, objects, living beings, persons—and whether they can reason using them. These concepts are often regarded as forming a "core knowledge" (Carey 2009), a basic, background understanding of the world that provides general rules for how the world work—rules that humans and animals can rapidly apply to new situations involving different objects and agents in different situations. Thus, even if someone has never seen a baseball game, they intuitively know how the balls, bats, and players will behave along various dimensions: throw balls will persist over time and continue moving in the same direction unless hit with a bat; players will walk on their legs, hold the bat with their hands, and often start and stop without being acted upon by an external force; and so on.

Various tests have been used to probe the world-models of LLMs, but the results are often underwhelming: for example, in Vafa et al. (2024), the authors probed the internal map of New York City generated by an LLM navigating the city. The various paths taken by the LLM often

crossed the same streets going in different directions, but the model did not unify these into a coherent map, instead producing a bunch of disconnected paths. Similar results were found in a simpler model trained to play the board game Othello: while the model accurately represented and tracked the pieces on the board (Li et al. 2023), they did not represent them into a separable spatial location and piece identity (such that the same piece could be at a separate place). Instead, they identified them *within* heuristics of play (Yuan and Søgaard 2025)—effectively representing the same piece at the same place as a different piece if it were engaged in a different series of actions. In short, in both the car case and with Othello, the LLMs struggled to disentangle the concepts of space and objects, suggesting its world-model does not accurately represent the world, but only maps loosely onto different possible actions. .

The problem of world-models helps explain the domain-specific failures of these models, as researchers tease out using quirky thought-experiments that prevent the model from simply reciting a memorized answers (Wu et al. 2023). For example, Collins et al. (2022) probed a model using examples where LLMs are tasked with solving basic but slightly out-of-the-box puzzles, such as figuring out how to get a couch to the roof without stairs or explaining why underwatered plants might not die. These tests highlight domain-specific knowledge—such as intuitive physics and intuitive biology—that humans often deploy unthinkingly when listening to a story or thinking through a problem. The specific object or entity involved is incidental; replacing the couch with a table or bike or even a dog would work largely the same, since the test is simply whether the LLM can accurately model the world accurately enough to come up with an answer. The LLMs could not: for example, when asked how to get a couch to the roof without stairs, CoPilot responds “lower it from a truck bed,” or “inflate it such that is floats.” Similar failures occur for situational reasoning (as seen in the driving example from Vafa et al. 2024), theory of mind (Burnell et al. 2023; Strachan et al. 2024), and planning (Momennejad et al. 2023; Valmeekan et al. 2023). The failures all stem from underlying challenges with the high-level concepts—and suggest that much of their capacity to solve these problems depends on specific word-choice. This often shows up when models can solve a math problem involving bananas but struggles with the same problem involving kiwi.

The inability to model objects, persons, and situations has knock on effects. For example, failures in theory of mind lead to deep failures in LLMs capacity to grasp the communicative intent of the user (Trott et al. 2021) or the broader pragmatics of language (Hu et al. 2022). Fine-tuning using human feedback can mitigate some of these problems. But, many philosohpers have noted, without theory of mind the model is not properly speaking *communicating* with the user. As Frankish warns, "[LLMs] don't assert, suggest, advise, warn, apologize, question, or do any of the other things we do in producing meaningful utterances [. . .] We may think they are advising us or instructing us, but they are just playing a game" (2024: 69; see also Szobeisk and Price 2022). Although the system acts as if we are an interlocuter, these researchers argue, they lack the underlying concept of mind necessary for engaging in conversation.

The takeaway from the shortstop position is that there is still a lot missing from current LLMs, and that simply scaling up won't help much. The problems stem from the mismatch between next-token prediction on linguistic data and the way domain-specific reasoning abilities often involve abstraction, reasoning, situation modelling, and planning. Without some supplement to mere next-token prediction, LLMs seem stuck with a "shallow" understanding (Browning and LeCun 2023)—one capable of often saying plausible things but lacking the capacity to put their concepts to work in the real-world.

5. Third Base

The third base position addresses an assumption underlying the other positions: the idea that the model has some basic self-identity over time. For Kant, our capacity to represent an objective world depends upon our awareness of our *own* persistence over time: we cannot be sure the world persists over time unless we also do. Although Kant's take is highly intellectualist, the underlying point is that grasping that our thoughts represent the world depends on some awareness that they are *our* thoughts—that there is a gap between mind and world. In the simplest creatures, the need to preserve a body against disturbances from outside is sufficient. In more complex creatures, as Burge highlights (2010), basic perceptual capacities often indicate to a species what distinguishes subjective sensory state from objective entities. In social creatures,

like humans, the concept of self becomes more robust, encompassing not just our persistent inner life but also our reputation and social commitments and obligations.

The important question is, what kind of self might be appropriate for an LLM? For some, any disembodied system will necessarily fail to grasp the concept of an objective world. Bisk et al. (2021), for example, contends embodiment if necessary for a sense of self, and Pezzulo et al. (2023) contend this is because interacting with the environment is essential for understanding the objective world. Shanahan (2024) goes even further, arguing that LLMs must fit into our human form of life—the distinctive way humans engage with each other—in order to merit talk of something having a mind.

But others, such as Chalmers 2024, highlight the long tradition of disembodied minds going back to Avicenna, and suggest these accounts provide space for a purely linguistic system might still possess a mind. And Shanahan's suggestion about the human form of life provides some guidance for how this might work: humans are normative creatures, and we take it as essential that others behave as they should. Social norms provide the shape not just for our language to be intelligible, but also the general structure for what it means to belong to a community (Haugeland 1982). Some researchers suggest we should use to fine-tune models to make appropriate comments given our social norms (Kasirzadeh and Gabriel 2023)—effectively make them *one of us*.

But current fine-tuning methods, such as reinforcement learning with human feedback, are coarse-grained: they push the model closer in the direction of an appropriate answer, but they do not ensure the model *gets it right* (Browning 2024b). The gap between near appropriate and genuinely appropriate is vast, as many users have learned when relying on LLMs for citations (see Staff 2023; Jaźwińska and Chandrasekar 2025). Ensuring a model gets it right often demands not just gradual correction but instead on sanctions—even in non-human species (Caro and Hauser 1992). Sanctions only matter to an entity with a self, since they depend on doing something that an agent does not desire. And for humans, getting it wrong in language isn't *just* a cognitive failure; it is also a social and even moral failure. This is because language users are fundamentally taken to be *persons*: beings who are responsible for their actions (Browning

2024a). A person is a specific type of self-identity, one which recognizes that their actions affect other people and, as such, needs to be cautious with what they say and commit to. As mentioned in section two, the models do not behave as if they have stable beliefs or any long-term goal, much less an objective world.

But, some argue, that they might *in a single conversation* or in a *single environment* possess these properties if properly prompted or fine-tuned. This argument has been suggested by Andreas (2022) and Shanahan et al. (2024), who suggests that within a conversation an LLM might adopt a temporary sense of self and a temporary world. This regards LLMs as role-playing in conversation, something capitalized on by companies like Character.AI, who fine-tune models to act according to a specific personality. A similar notion might be ascribed in the case of a system like Voyager (Wang et al. 2023), which uses an LLM to successfully play Minecraft. Neither approach is terribly stable or responsible; Character.AI has gotten into trouble because of the irresponsible behavior of its models (Roose 2024), and Voyager is a cumbersome system that involves non-LLM modules to ensure the model does not begin to wander aimlessly. Still, these do highlight paths to generating something closer to a mind, rather than a bullshitting chatterbox.

The third base position suggests that a central feature necessary for mindedness is a sense of self—even if a minimal one. For this position, the failures to behave rationally and recognize an objective world are the flip-side of the lack of a sense of self.

6. Conclusion

This paper aims to bring the problem of intentionality into contemporary discussions of LLM. But this paper also provides grounds for thinking that the various problems LLMs run into—nonsense outputs, formally incorrect reasoning, inconsistent behavior, and stupid solutions to problems—all stem from lacking a mind. While LLMs might well be a useful "cultural technology" (Yiu et al. 2024), their behavior remains more mindless tool than a purposive agent aiming to accomplish some task.

But the goal of baseball is not to get to third-base. What would it take to get the model all the way home? Although it is important that LLMs know what they are talking about and how the world works and what the right thing to say is, this only touches on the epistemic aspects. Fundamentally, a speaker needs to know what *matters*; as Haugeland points out, "the most ordinary conversations are fraught with life and all its meanings" (1979: 632). For an LLM to genuinely possess a mind, it needs to grasp that making claims means sticking our neck out—and, as such, risks getting it chopped off. It is not enough for an LLM to recognize that there are risks, but they also must *care* about the risks—they must grasp that actions are risky *for them*, in a way that influences their behavior. In short, to get LLMs all the way around the bases, we need to figure out how to make them give a damn.

References

Agüera y Arcas, B. (2022). Do Large Language Models Understand Us? Daedalus, 151(2), 183–197.

Bender, E. M., Gebru, T., Mcmillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.

Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A. C., Korbak, T., & Evans, O. (2023). The Reversal Curse: LLMs trained on" A is B" fail to learn" B is A". arXiv preprint.

Bisk, Y., Holzman, A., Tomason, J., Andreas, J., Bengio, Y., Chai, J., . . . Turian, J. (2020, November). Experience Grounds Language. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 8718-8735.

Brandom, R. (1994). Making it Explicit. Harvard UP.

Browning, J., & LeCun, Y. (2023). Language, common sense, and the Winograd schema challenge. Artificial Intelligence, 325, Article 104031.

Browning, J. (2024a). Personhood and AI: Why large language models don't understand us. AI & Soc (2023).

Browning, J. (2024b). Getting it right: the limits of fine-tuning large language models. Ethics Inf Technol **26**, 36 (2024).

Buckner, C. (2013). Morgan's Canon, meet Hume's Dictum: avoiding anthropofabulation in cross-species comparisons. Biology and Philosophy 28 (5):853-871.

Buckner, C. (2023). Black Boxes or Unflattering Mirrors? Comparative Bias in the Science of Machine Behaviour. British Journal for the Philosophy of Science 74 (3):681-712.

Buckner, C. (2023). From Deep Learning to Rational Machines. Oxford UP.

Burge, T. (2010). Origins of Objectivity. Oxford, GB: Oxford UP.

Burnell, B., Schellaert, S., Burden, B., Ullman, U., Martinez-Plumed Martinez-Plumed, Tenenbaum, T., et al. (2023). Rethink reporting of evaluation results in AI. Science, 380(6641), 136-138.

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., ... & VanRullen, R. (2023). Consciousness in artificial intelligence: insights from the science of consciousness. arXiv preprint.

Cappelen, H. & Dever, J. (2021). Making AI Intelligible: Philosophical Foundations. New York, USA: Oxford UP.

Caro, T. M., and M. D. Hauser. 1992. Is there teaching in nonhuman animals? Quarterly Review of Biology 67:151–174.

Chalmers, D. J. (2023). Could a large language model be conscious?. arXiv preprint arXiv:2303.07103.

Chalmers, D. J. (2024). Does thought require sensory grounding? From pure thinkers to large language models. arXiv preprint arXiv:2408.09605.

Chalmers, D. J. (2025). Propositional interpretability in artificial intelligence. arXiv preprint arXiv:2501.15740.

Chirimuuta, M. (2024). The brain abstracted: simplification in the history and philosophy of neuroscience. Cambridge, Massachusetts: MIT Press.

Chomsky, N., Roberts, I., & Watumull, J. (2023, March 8). The false promise of chatgpt. The New York Times. https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html

Churchland, P. (1996). The Engine of Reason, the Seat of the Soul. MIT Press.

Cohen, H. (1885). Kants Theorie der Erfahrung (2nd ed.). Berlin: F. Dummler.

Cummins, R. (1989). Meaning and Mental Representation. MIT Press.

Coelho Mollo, D.. & Millière, R. (2023). The vector grounding problem, arXiv preprint.

Davidson, D. (1973). Radical Interpretation. Dialectica, 27: 314–28.

Dennett, D. (1971). Intentional Systems. The Journal of Philosophy, 68(4): 87-106.

Dennett D. (1983). Intentional systems in cognitive ethology: The "Panglossian paradigm" defended. Behavioral and Brain Sciences. 6(3): 343-355.

Dennett, D. (1988). Conditions of Personhood. In: Goodman, M.F. (eds) What Is a Person? Contemporary Issues in Biomedicine, Ethics, and Society. Humana Press.

Dennett, D. (1997). True Believers: The Intentional Strategy and Why It Works. Mind Design II: Philosophy, Psychology, and Artificial Intelligence, John Haugeland

Dennett, D. (05/31/2023). "The Problem with Counterfeit People." The Atlantic, Atlantic Media Company, http://www.theatlantic.com/technology/archive/2023/05/problem-counterfeit-people/674075/

Dentella, V., Murphy, E., Marcus, G., & Leivada, E. (2023). Testing AI performance on less frequent aspects of language reveals insensitivity to underlying meaning. Arxiv.

Descartes, R. (1984). The Philosophical Writings of Descartes. (J. Cottingham, R. Stoothoff, & D. Murdoch, Trans.) Cambridge: Cambridge UP.

Dziri, N., Lu, X., Sclar, M., Li, X. L., Jiang, L., Lin, B. Y., ... & Choi, Y. (2023). Faith and fate: Limits of transformers on compositionality. Advances in Neural Information Processing Systems, 36, 70293-70332.

Esanu, A. (2024). Scrutinizing the foundations: could large language models be solipsistic? Synthese 203 (5):1-20.

Fedorenko, E., Piantadosi, S.T. & Gibson, E.A.F. Language is primarily a tool for communication rather than thought. Nature 630, 575–586 (2024).

Firth, J. (1957). A Synopsis of Linguistic Theory, 1930-55. In Studies in Linguistic Analysis (pp. 1-31). Special Volume of the Philological Society. Oxford: Blackwell.

Fodor, J. (1975). The Language of Thought. Harvard UP.
Fodor, J. (1978). Propositional attitudes. The Monist 61 (4):501-23.
Frank, M.C. (2023). Baby steps in evaluating the capacities of large language models, Natural Review of Psychology. 2: 451–452.
García-Ferrero, I., Altuna, B., Alvez, J., Gonzalez-Dios, I., & Rigau, G. (2023). This is not a Dataset: A Large Negation Benchmark to Challenge Large Language Models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 8596–8615., Singapore. Association for Computational Linguistics.
Grant, N.& Metz, C. (12 Jun 2022). Google Sidelines Engineer Who Claims Its A.I. Is Sentient. *The New York Times*, www.nytimes.com/2022/06/12/technology/google-chatbot-ai-blake-lemoine.html.
Grice, P. (1975). "Logic and conversation". In Cole, P.; Morgan, J. (eds.). *Syntax and semantics*. Vol. 3: Speech acts. New York: Academic Press.
Harding, J. & Sharadin, N. (forthcoming). What is it for a Machine Learning Model to Have a Capability? British Journal for the Philosophy of Science.
Harnad, S. (1990). The symbol grounding problem. *Physica d: Nonlinear Phenomena, 42*(1–3), 335–346.
Haugeland, J. (1979, November). Understanding Natural Language. Journal of Philosophy, 76, 619-32.
Haugeland, J. (1982). Heidegger on Being a Person. Nous, 16(1), 15-26.
Haugeland, J. (1985). Artificial Intelligence: The Very Idea. MIT Press.
Haugeland, J. (1990). The intentionality all-stars. Philosophical Perspectives 4:383-427.
Hicks, M.T., Humphries, J. & Slater, J. (2024). ChatGPT is bullshit. Ethics Inf Technol 26, 38. https://doi.org/10.1007/s10676-024-09775-5
Hu, J., Floyd, S., Jouravlev, O., Fedorenko, E., & Gibson, E. (2022). A fine-grained comparison of pragmatic language understanding in humans and language models. ArXiv. doi:10.48550/arxiv.2212.06801
Jaźwińska, K.& Chandrasekar, A. (6 Mar 2025). "AI Search Has a Citation Problem." Columbia Journalism Review, www.cjr.org/tow_center/we-compared-eight-ai-search-engines-theyre-all-bad-at-citing-news.php.
Kasirzadeh A, Gabriel I (2023) In conversation with artificial intelligence: aligning language models with human values. Philosophy of Technology.
Lake, B., & Baroni, M. (2018, July). Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In International conference on mac*hine learning* (pp. 2873-2882). PMLR.
LeCun, Y. (2022). A path towards autonomous machine intelligence, Version 0.9.2, 2022-06-27. https://openreview.net/ forum?id=BZ5a1r-kVsf.
Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., & Wattenberg, M. (2023). Emergent world representations: Exploring a sequence model trained on a synthetic task. ICLR.
Lupyan, G. (2016), The Centrality of Language in Human Cognition. Language Learning, 66: 516-553.
Mahowald K, Ivanova A, Blank I, Kanwisher N, Tenenbaum J, Fedorenko E. (2023) Dissociating language and thought in large language models. Arxiv.
Mahowald, K., Ivanova, A., Fedorenko, E., Blank , I. A., Tenenbaum, J., & Kanwisher, N. (2024, January 24). Google's powerful AI spotlights a human cognitive glitch: Mistaking fluent speech for fluent thought. The Conversation. https://theconversation.com/googles-

powerful-ai-spotlights-a-human-cognitive-glitch-mistaking-fluent-speech-for-fluent-thought-185099

Mandelkern, M. & Linzen, L. (2023). Do Language Models Refer? Arxiv. https://doi.org/10.48550/arXiv.2308.05576

Manning, C. Clark, K. Hewitt, J. Khandelwal, U. & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision, Proc. Natl. Acad. Sci. U.S.A. 117 (48) 30046-30054, https://doi.org/10.1073/pnas.1907367117

Marcus, Gary F. (2001). The Algebraic Mind: Integrating Connectionism and Cognitive Science. MIT Press.

McCoy, R. T., Yao, S., Friedman, D., Hardy, M., & Griffiths, T. L. (2023). Embers of autoregression: Understanding large language models through the problem they are trained to solve. arXiv preprint arXiv:2309.13638.

Mitchell, M. (2023). How do we know how smart AI systems are? Science 381(5957).

Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., & Farajtabar, M. (2024). GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. arXiv preprint arXiv:2410.05229.

Momennejad, I., Hasanbeig, H., Vieira, F., Sharma, H., Ness, R. O., Jojic, N., ... & Larson, J. (2023). Evaluating Cognitive Maps and Planning in Large Language Models with CogEval. arXiv preprint arXiv:2309.15129.

Patel R, Pavlick E. (2022). Mapping language models to grounded conceptual spaces. In Int. onf. on Learning Representations, Online, 25–29 April 2022.

Pavlick, E. (2022). Semantic Structure in Deep Learning. Annual Review of Linguistics, 8(1), 447-71.

Pavlick E. (2023). Symbols and grounding in large language models. Philosophical transactions. Series A, Mathematical, physical, and engineering sciences, 381(2251), 20220041.

Piantasodi, S. T., & Hill, F. (2022). Meaning without reference in large language models. ArXiv, 1-8.

Piantadosi, S. (2023). Modern language models refute Chomsky's approach to language. Lingbuzz Preprint.

Putnam, Hilary (1975). The meaning of 'meaning'. Minnesota Studies in the Philosophy of Science 7:131-193.

Roose, K. (23 Oct. 2024) "Can A.I. Be Blamed for a Teen's Suicide?" The New York Times, The New York Times, www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html.

Quilty-Dunn J, Porot N, Mandelbaum E. The best game in town: The reemergence of the language-of-thought hypothesis across the cognitive sciences. Behav Brain Sci. 2022 Dec 6;46:e261.

Santoro, A., Lampinen, A., Mathewson, K., Lillicrap, T., & Raposo, D. (2021). Symbolic behaviour in artificial intelligence. arXiv preprint arXiv:2102.03406.

Sap, M., LeBras, R., Fried, D., & Choi, Y. (2022). Neural theory-of-mind? on the limits of social intelligence in large LMs. ArXiv. doi:10.48550/arxiv.2210.13312

Searle, J., 1980, 'Minds, Brains and Programs', Behavioral and Brain Sciences, 3: 417–57

Shanahan, M. (2024). Talking about Large Language Models. Commun. ACM 67, 2 (February 2024), 68–79.

Shanahan, M., McDonell, K. & Reynolds, L. (2023). Role play with large language models. Nature 623, 493–498.

Sobieszek, A., & Price, T. (2022). Playing games with AIs: The lim¬its of GPT-3 and similar large Language models. Minds and Machines, 32(2), 341–364.

Søgaard, A. (2023). Grounding the vector space of an octopus: word meaning from raw text. Minds and Machines, 33(1), 33–54.

Staff. (2023, December 30). Michael Cohen says he unwittingly sent ai-generated fake legal cases to his attorney. NPR. https://www.npr.org/2023/12/30/1222273745/michael-cohen-ai-fake-legal-cases

Strachan, J.W.A., Albergo, D., Borghini, G. *et al.* Testing theory of mind in large language models and humans. Nat Hum Behav **8**, 1285–1295 (2024). https://doi.org/10.1038/s41562-024-01882-z

Tang, X., Zheng, Z., Li, J., Meng, F., Zhu, S., Liang, Y., et al. (2023). Large language models are in-context semantic reasoners rather than symbolic reasoners.

Tangermann, V. (17 June 2024). "McDonald's Abandoning AI-Powered Drive Thrus after Embarrassing Failures." Futurism, Futurism, futurism.com/the-byte/mcdonalds-abandoning-ai-drive-thrus.

Trott, S., Torrent, T. T., Chang, N., & Schneider, N. (2020). (Re)construing meaning in NLP Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.462

Vafa, K., Chen, J., Rambachan, A., Kleinberg, J., & Mullainathan, S. (2024). Evaluating the world model implicit in a generative model. Advances in Neural Information Processing Systems, 37, 26941-26975.

Valmeekam, K., Olmo, A., Sreeharan, S., & Kambhampati, S. (2023). Large Language Models Still Can't Plan. Arxiv, 1-21.

Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., ... & Anandkumar, A. (2023). Voyager: An open-ended embodied agent with large language models. arXiv preprint arXiv:2305.16291.

Weizenbaum, J. (1976). Computer Power and Human Reason: From Judgment to Calculation. New York: W. H. Freeman and Company.

Wu, Z., Qiu, L., Ross, A., Akyürek, E., Chen, B., Wang, B., ... & Kim, Y. (2023). Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. arXiv preprint

Yildirim, I., & Paul, L. A. (2024). From task structures to world models: What do LLMs know? Trends in Cognitive Sciences, 28(5), 404–415.

Yiu, E., Kosoy, E., & Gopnik, A. (2023). Transmission Versus Truth, Imitation Versus Innovation: What Children Can Do That Large Language and Language-and-Vision Models Cannot (Yet). Perspectives on Psychological Science, 0(0).

Yuan, Y., & Søgaard, A. (2025). Revisiting the Othello World Model Hypothesis. arXiv preprint arXiv:2503.04421.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., et al. (2019). Fine-tuning language models from human preferences. doi:10.48550/arxiv.1909.08593