

This paper is a draft of an invited lead symposium article for the *Asian Journal of Philosophy*, and is currently under review. Please check the publication status before citing this version.

# Sapience without Sentience: An Inferentialist Approach to LLMs

Ryan Simonelli

February 17, 2026

## Abstract

Do large language models (LLMs) possess concepts, such that they can be counted as genuinely understanding what they're saying? In this paper, I approach this question through an inferentialist account of concept possession, according to which one's possession of a concept is understood in terms of one's mastery of the inferential role of a linguistic expression. I suggest that training on linguistic data is in principle sufficient for mastery of inferential role, and thus, LLMs trained on nothing but linguistic data could in principle possess all concepts and thus genuinely understand what they're saying, even when speaking about such things as colors and tastes, guilt and folly, life and death. This doesn't mean, however, that they are conscious. I draw a classical distinction between *sentience* (conscious awareness) and *sapience* (conceptual understanding) and argue that we might think of LLMs as genuinely possessing the latter without even a shred of the former. In defending this claim, I argue that attributing conceptual understanding to a system is not a matter of describing some specific empirical property that the system shares with us but, rather, as Wilfrid Sellars says, "placing it in the logical space of reasons," treating it as answerable to calls for reasons, clarifications, corrections, and so on. I claim that we may aptly adopt this attitude towards sufficiently capable LLMs without thereby treating them as conscious subjects.

## 1 Introduction

In a thread on *Twitter* back in 2023, Geoffrey Hinton summed up his main disagreement with Yan Lecun, one which continues to this day, as follows:<sup>1</sup>

---

<sup>1</sup><https://x.com/geoffreyhinton/status/1728490334336770138>

The central issue on which we disagree is whether LLMs actually understand what they are saying. You think they definitely don't and I think they probably do.

There were two main sentiments in the thread of comments that followed: one of puzzlement and one of conviction. First, several commentators of the first sentiment asked for clarification what, exactly, was the notion of “understanding” that Hinton was appealing to here. What, exactly, does it mean to say that an LLM understands what it is saying? How would we know whether or not an LLM does understand what it is saying? If an LLM is engaging in a conversation about emotions and understands what it's saying, does that mean that it has emotions too? Perhaps influenced by this last sort of question, commentators of the second sentiment confidently asserted that LLMs couldn't possibly understand what they're saying since they don't have any experience. After all, they're trained on nothing but linguistic data, and all they're doing is predicting the next token in a series, feeding this back into the input, and iterating this process. They simply don't have the sort of experiential link to the world that genuine understanding requires. For instance, they might use the word “red” as English speakers do, but clearly, they don't grasp the concept of red that English speakers grasp. They don't know what it is for something to be red. Accordingly, they don't actually know what they're saying when they say, for instance, “Mars is red.”

It is perhaps easy to immediately side with commentators of the second sentiment, maintaining that, obviously, LLMs don't understand what they're saying, and such sentiments are widely reflected in the academic literature on LLM understanding.<sup>2</sup> However, the fact that someone as deeply knowledgeable of the technology underlying LLMs as Geoffrey Hinton maintains that they *do* understand what they're saying ought to give us at least some pause. In this paper, I will show that, on at least one prominent contemporary account of concept possession, Hinton's claim can be coherently maintained in opposition to the second sentiment expressed above. The account on which I'll

---

<sup>2</sup>For recent influential expressions of this sentiment pertaining to LLMs in particular, see [9], [8]. For an overview of the debate on LLM understanding in the AI community see [42]. For a philosophically sophisticated recent defense of this sentiment, see [33].

focus here is an *inferentialist* account of concept possession, according to which to grasp a concept is to master the inferential role of a linguistic expression. What makes inferentialism particularly interesting in the context of our question about LLMs is that the ambitious version of the view developed by Sellars [49] and Brandom [13] [14], which I'll call *global strong inferentialism*, seems to facilitate a straightforwardly positive answer to the question of whether they really do understand. In particular, global strong inferentialism, which maintains that mastery of inferential role is sufficient for possession of all concepts, seems to facilitate the following argument:

1. Mastery of inferential role is sufficient for possession of all concepts. (global strong inferentialism)
2. Training on linguistic data is (in principle) sufficient for mastery of inferential role. (premise)
3. So, training on linguistic data is (in principle) sufficient for possession of all concepts. (from 1 and 2)
4. So, Large Language Models (LLMs), trained on nothing but linguistic data, are (in principle) capable of possessing all concepts. (from 3).

The conclusion of this argument might strike one as so obviously implausible that one might think that the validity of this argument, along with the plausibility of premise 2, is in fact a *reductio ad absurdum* of global strong inferentialism. The minimal, negative aim of this paper is to convince you that this is not the case. The maximal, positive aim of this paper is to argue, from the plausibility of global strong inferentialism, that this argument may in fact be sound. The key thought in defusing the initial implausibility of the conclusion is to distinguish *sapience*, in the sense of conceptual understanding, from *sentience*, in the sense of conscious awareness. Though the particular case of sapience with which we are familiar (our own) does of course involve sentience, I claim that sapience does not, as such, entail sentience, and LLMs, trained for next token prediction on nothing but linguistic data, may in principle have the former without having the latter.

This paper is structured as follows. In Section 2, I will lay out the strong version of inferentialism proposed by Sellars and Brandom, which I've defended

and developed at length elsewhere [53] [54] [56]. In Section 3, I will consider a consequence of this account: that one might have a concept with an essentially experiential aspect without actually having the relevant experiences oneself. I first consider the *local* case of this sort, exemplified by the example of Mary the color-blind color scientist who grasps the concept *red*, before raising the possibility of a *global* case of this sort, where something grasps a whole host of concepts without having any experiences at all. In Section 4, I argue that LLMs can in principle be understood as exemplars of this possibility, not being sentient at all but having the inferential mastery required for possession of concepts of all sorts. In Section 5, I argue that attributing such conceptual understanding to an LLM is not a matter of making some descriptive claim about what’s going on in the system, but, rather, the adoption of a certain sort of normative attitude towards it: taking it answerable to calls for reasons, clarifications, corrections, and so on, and, indeed, able to answer such calls. I claim that there is no in principle reason why we could not appropriately adopt this attitude towards LLMs. I conclude in Section 6 by briefly considering the practical implications of the possibility of sapience without sentience when it comes to the safe and ethical development of future AI systems.

## 2 Global Strong Inferentialism

Let me start by briefly presenting the version of inferentialism from which we will approach the question of LLM understanding. The general inferentialist approach, according to which concept possession is understood in terms of mastery of inferential role, might be disambiguated along two dimensions of ambition: first, whether mastery of inferential role (MIR) is taken to be *necessary* for concept possession or whether it is taken to be *sufficient*, and, second, whether this approach is extended only to *some* concepts or whether it is applied to *all* of them. Thus, we may disambiguate four varieties of inferentialism as follows:

	<b>MIR is Necessary</b>	<b>MIR is Sufficient</b>
<b>Some Concepts</b>	Restricted Weak	Restricted Strong
<b>All Concepts</b>	Global Weak	Global Strong

I will focus here on the most ambitious form of inferentialism, defended by such authors as Sellars [49] and Brandom [13] [14], according to which mastery of inferential role is *sufficient* for possession of *all* concepts. This is the form of inferentialism that actually facilitates the argument stated above, and thus is the one of interest to us.

Before turning to global strong inferentialism, which is the view we are ultimately interested in, let us first consider the most prominent version of *restricted* strong inferentialism, known as *logical inferentialism* [26]. According to logical inferentialism, the meanings of logical connectives like “not,” “and,” “or,” and so on might be understood in terms of the inferential rules governing their use, canonically specified as rules in a certain kind of formal proof system. For example, the meaning of **and** might be understood in terms of the fact that its use is governed by the following rules in a natural deduction system:

$$\frac{A \quad B}{A \text{ and } B} \wedge_I$$

$$\frac{A \text{ and } B}{A} \wedge_{E_1} \qquad \frac{A \text{ and } B}{B} \wedge_{E_2}$$

Here, we associate **and** with an *introduction rule* specifying the inferential *conditions* for asserting a sentence of the form  $A \text{ and } B$  and a pair of *elimination rules*, specifying the inferential *consequences* of asserting a sentence of the form  $A \text{ and } B$ . The basic thought of logical inferentialism is that mastery of these rules is sufficient for grasp of the concept of conjunction.

Global strong inferentialism extends the basic inferentialist idea beyond just *logical* words like “and,” but also *content words* like “red,” “sphere,” “dog,” and so on. According to the global strong inferentialist, the meanings of these words, no less than words like “and,” can be understood in terms of the inferential rules governing their use. Unlike the *formal* rules of inference which constitute the meaning of logical words like “and,” the rules constituting the meaning of a content word like “red” are *material* inferential rules [49]. Such rules are not specified schematically, but, rather, directly relate particular content words like “red” to other content words like “colored,” “scarlet,” and “green.” Concretely, where  $X \vdash A$  expresses that the sentence  $A$  can be inferred

from the set of sentences  $X$  and  $\perp$  expresses an incoherence, the meaning of “red” might be understood in terms of the fact that its use is governed by material inferential rules such as the following:

**red**( $a$ )  $\vdash$  **colored**( $a$ )

**scarlet**( $a$ )  $\vdash$  **red**( $a$ )

**red**( $a$ ), **green**( $a$ )  $\vdash \perp$

**red**( $a$ ), **pink**( $b$ )  $\vdash$  **darker**( $a, b$ )

The rule **red**( $a$ )  $\vdash$  **colored**( $a$ ) can be understood as a kind of elimination rule, articulating the *consequences* that can be inferred from the sentence **red**( $a$ ), whereas **scarlet**( $a$ )  $\vdash$  **red**( $a$ ) can be understood as a kind of introduction rule, articulating the *conditions* from which **red**( $a$ ) can be inferred.<sup>3</sup> Now, our grasp of the conceptual significance of **red** clearly does not simply consist in our grasp of the inferential relations that it bears to other color words. Additionally, it clearly seems to involve some understanding of what things are paradigmatically red, and what being red implies for things that may or may not be red. So, in addition to the above *intra-familial* inferential relations, which relate one color concept to other color concepts, we will also need to include *extra-familial* inferences such as the following:<sup>4</sup>

**stop sign**( $a$ )  $\vdash$  **red**( $a$ )

**blackberry**( $a$ ), **ripe**( $a$ ), **red**( $a$ )  $\vdash \perp$

By including rules of this sort in our inferentialist theory, we distinguish the conceptual significance of **red** from that of other color terms such as **green**. However, even including rules of this sort, it still seems that, though inferential rules of this sort may be *necessary* aspects of the meaning of **red**, they can’t

<sup>3</sup>The idea that meaning can be understood in terms of consequences of use and conditions of correct use is known as the “two-aspect model of meaning” [57] proposed by Dummett [23].

<sup>4</sup>One might wonder *which* of these extra-familial inferences are to be included in articulating content of a term like “red.” On the one hand, if *none* of them are included, then “red” loses its significance in connection with things in the world. On the other hand, if *all* of these inferences are included, then, since no one has concepts of all of the red things, no one would fully possess a grasp of the content of “red.” However, including inferences to some red things (stop signs, say) and not others (raspberries, say) would seem to draw arbitrary lines. I am in fact prepared to accept the second horn of this apparent dilemma; our conceptual understanding is always partial. It will turn out to be a consequence of the account to follow that, beyond just possessing concepts, an LLM can in fact possess a much *fuller* understanding of any material concept than any of us might possess.

*exhaust* its meaning. In particular, it seems that one essential aspect of the meaning of “red” is that it can be properly deployed, not only as the conclusion of an *inference* (for instance, from “scarlet”), but also *non-inferentially*, upon *seeing something red!*

The standard response to the above issues is to concede that inferential rules of the sort stated above are not enough to articulate the meaning of **red**.<sup>5</sup> Drawing terminology from Sellars [50], many commentators have argued that, in addition to “language-language rules,” which obtain between uses of language (e.g. assertoric utterances of sentences) and other uses of language, we also need to include so-called “language-entry rules,” which obtain between uses of language (e.g. assertoric utterances of sentences) and non-linguistic perceptual circumstances. Such rules are, as Peregrin [48] says, “not really inferential in any straightforward sense of the word,” but they are sufficiently inference-like that they might be included in a “strong inferentialist theory.” So, on the standard version of “strong inferentialism,” not only are there inferences, for instance from “The ball is scarlet” to “The ball is red,” but there are also “quasi-inferences,” for instance, from seeing the red ball to asserting (or being entitled to assert) “The ball is red.” We might picture such a quasi-inference along the following:



Now, if such “language-entry” rules are included in an inferentialist theory of concept possession, and grasping such rules requires having the relevant perceptual experiences, then, clearly, LLMs without such experiences would not be in possession (or, at least, full possession) of empirical concepts such as *red*.

Though standard developments of inferentialism in recent years have assumed such “quasi-inferential” rules must be included in an inferentialist theory, I have argued at length that an inferentialist theory *shouldn’t* include “quasi-inferential” rules of this sort [54]. First off, this essentially amounts to giving up

---

<sup>5</sup>For a sampling of statements of this standard response, see [54, 4].

global *strong* inferentialism for global *weak* inferentialism, and with it the very idea of what Dummett [24] refers to as a “full-blooded” theory of meaning, a theory that aims to give a non-circular account of the concepts expressed by the primitive expressions of a language.<sup>6</sup> Moreover, it is far from clear that an inferentialist actually *needs* to countenance such “quasi-inferential” rules in order to account for the fact that terms like “red” are deployed non-inferentially in perceptual response to red things. Rather, the inferentialist can account for this fact in *purely inferential terms*. How can the fact that “red” can be deployed non-inferentially in perception be captured in purely inferential terms? Well, I take it, quite easily. Just include inferential relations such as the following:<sup>7</sup>

**red(*a*), color vision(*S*), looking at(*S*, *a*), good lighting** ⊢ **sees(*S*, ⟨red(*a*)⟩)**

This says that if *a* is red and some subject *S* has color vision and is looking at *a* in good lighting, then *S* sees that *a* is red. In contrast to the above “quasi-inferential” rule, this is an *inferential* rule, properly so-called. The contents of each of the terms that figure in this inferential rule, such as **color vision** and **good lighting**, are themselves spelled out inferentially in terms of rules such as the following (of course, among very many others):

**color vision, blind** ⊢ ⊥                      **daytime, outside** ⊢ **good lighting**

Thus, on the version of the global strong inferentialist view I advance (which I argue is in fact the version of the view defended by both Sellars and Brandom) *inferences*, properly so-called, suffice to *completely articulate* the conceptual content of even a term like **red**, which is essentially such as to be deployed in perception.

An inferentialist account of this sort can be spelled out in formal detail, specifying the inferential rules constitutive of the meaning of both logical and non-logical vocabulary in the context of a formal proof system of the sort

---

<sup>6</sup>For a defense of a theory of this sort against influential criticisms by McDowell [41] [40], see my [56].

<sup>7</sup>It’s worth noting that this inference (along with many of the inferences that will figure in a global inferentialist theory) is *defeasible*. I will return to discuss this feature of material inferences in Section 4.

developed by Gentzen [26].<sup>8</sup> I will not, of course, go into such formal details here, focusing instead on the account's philosophical upshots. As an account of concept possession, the basic thought is that we acquire concepts through the process of language learning wherein we come to master the inferential rules governing the use of linguistic expressions. It is through this process of language learning that, as Wittgenstein says, "Light dawns gradually on the whole" [64, 21]. On this account, our conceptual knowledge consists in our mastery of inferential rules. For instance, having mastered the inferential rules governing the use of "red" stated in this section, if you ask me what it is for something to be red, I can answer along the following lines:

For something to be red is for it to be colored in a certain way. If something is scarlet or crimson, it's necessarily colored this way, if something is green or blue (all over), it can't possibly be colored this way. Stop signs, blood, and unripe blackberries are colored this way. If something's colored this way, and someone with color vision is looking at that thing in good lighting, they will see that it is colored this way, and so on.

Such modalized conditional statements inferentially articulate what it is for something to be red, and the normative expressivist thought, put forward by Sellars and Brandom and recently defended by Thomasson [58], is that making such modalized conditional statements is a way of expressing the inferential rules governing the use of "red." Conceptual knowledge, in this way, is inferential knowledge.

### **3 Conceptual Understanding without Conscious Experience**

Having articulated the basic thesis of global strong inferentialism, let us go on to consider some of its consequences. Frank Jackson [36] famously proposes the thought experiment of Mary the color scientist, who has been in a black and white room since birth and so has never experienced color, but has nevertheless reached the theoretical limit of what can be known about the colors without

---

<sup>8</sup>For such formal development, see [37], [35] and [55]

actually having experienced them.<sup>9</sup> The question is: *does she know what it is for something to be red?* The standard intuition is that she does not: to know what it is for something to be red, you have to see red.<sup>10</sup> On the global strong inferentialist view advanced here, yes, she knows just what it is for something to be red, since she grasps all of the inferential relations between sentences that articulate the content of “red.” Though she’s never herself used the term non-inferentially, she knows just the conditions under which it can be non-inferentially used, and that is all that is required in order to be counted as knowing the meaning of “red.” Indeed, by the same reasoning, we can argue (even more surprisingly) that not only does Mary know what it is for something to be red, but she knows what it is to *see that something is red*. The concept of seeing that something is red is just as well articulated by properly inferential relations like the one stated above, and, once again, Mary grasps all of these inferential relations. She thereby grasps what it is to see that something is red, even though she’s never actually seen that something is red herself.

Of course, I acknowledge the initial counter-intuitiveness of the above inferentialist thought. Surely, it seems, Mary must be lacking something. The inferentialist gambit is to simply ask *what* is Mary lacking? Of course, she does lack actual experiences of seeing red, but that is what is acknowledged between both parties. The question is: what elements of the conceptual content of “red” is she lacking understanding? Anything about the concept *red* that could actually be *said*, in response to this question, could be understood in inferential terms, and thus, something that Mary could be taken to grasp. Thus, one is left saying that the thing that Mary is lacking is *inarticulable*, indeed, *ineffable*. In response to the question of what aspect of the conceptual content of “red” Mary is lacking “one gets to the point where one would like just to emit an inarticulate sound,” (Wittgenstein [65, §261]).<sup>11</sup> The core inferentialist commit-

---

<sup>9</sup>It is perhaps more realistic (though less entertaining) to imagine that, rather than being in a black and white room, she is completely color blind.

<sup>10</sup>In the literature in philosophy of mind, this standard intuition is most articulately spelled out in terms of the claim that Mary lacks the “phenomenal concept” of redness, which is distinct from whatever concepts she might have that she associates with the term “red.” See especially Chalmers [19] [20, 251-276].

<sup>11</sup>This quote from Wittgenstein occurs in the midst of the so-called “private language argument,” targeted at the thought that the contents of the terms used to express our own

ment here is simply that every bit of the conceptual content of every expression is effable, that is, linguistically articulable. Insofar as one maintains this core commitment, there is nothing at all about the conceptual content of a word like “red” that someone who has never seen red must in principle be lacking. They can know all of relations between red and other colors, they can know all of the different things that are red and the circumstances in which they are, and, importantly, they can know precisely the conditions under which someone with color vision will see that something’s red and how such a person will respond to such a perceptual episode. They can know all of this even if they have never had such a perceptual episode themselves or, indeed, are incapable of having such a perceptual episode. Of course, as a matter of fact, people who have never seen red—for instance, people with total achromatopsia—will typically have a less robust conceptual understanding of red and other colors, but the crucial point is that there is nothing *in principle* precluding them from acquiring this conceptual understanding.

I will not do any more than what I already have to justify this counter-intuitive consequence of global strong inferentialism when it comes to the case of Mary the color scientist. Instead, I want to consider a radical extension of this basic thought. If we adopt this global strong inferentialist approach to the case of Mary, then this illustrates at least one case in which one can possess a concept that has an essentially experiential aspect without actually having the relevant experiences. And, of course, we can conceive of many other such cases. For instance, the same sort of considerations may be applied to the Vulcan psychologist who is, in principle, capable of talking with full understanding about such emotions as grief, joy, guilt, and so on, without ever having felt such things herself.<sup>12</sup> Once again, insofar as she grasps all of the relations between these different emotions, the conditions in which

---

sensations, fixed by private demonstration, are essentially inarticulable in public language. I am taking inferentialism as a premise here, and so I will not properly engage with this opposing thought, but a full defense of the inferentialist commitment to the articulability of content that I am undertaking here would involve spelling out an argument (perhaps some version of Wittgenstein’s argument, whatever it actually is) against the intelligibility of such inarticulable contents.

<sup>12</sup>I suppose for simplicity (and in opposition to official *Star Trek* lore), that Vulcans simply do not have such emotions, rather than suppressing them.

feeling these emotions is apt, the way people typically respond to having these emotions, and so on, what else could she possibly be lacking? Once again, I submit that any attempt to specify what she's lacking in a way that does not simply beg the question against the present proposal will leave one wanting to emit an inarticulate sound. So, these are at least two *local* cases in which one can have a concept with an essentially experiential dimension, a concept which is essentially such that it can be deployed non-inferentially in cases in which one has certain experiences, without actually having the relevant experiences. The radical question I want to raise is whether this same basic thought might be applied *globally*, to something without any experiences at all. To answer this question, we need to say some general things about conscious awareness and conceptual understanding.

In framing the inferentialist account of conceptual content on which I am drawing here, Brandom [13] clarifies his topic by drawing a classical distinction between conscious awareness, referred to as *sentience*, and conceptual understanding, referred to as *sapience*.<sup>13</sup> The core notion of sentience, understood as a particular kind of conscious awareness, is articulated by Brandom as follows:

Sentience is what we share with nonverbal animals such as cats—the capacity to be aware in the sense of being awake, [...] to be distinguished from the mere reliable differential responsiveness we sentient beings share with artifacts such as thermostats and land mines [13, 5].

On the conception of sentience suggested by Brandom, it is, in the first instance, an aspect of the kind of bodily practical intelligence that something like a cat exhibits as it skillfully navigates its environment, responding to features of its environment in ways that are appropriate, given its aims and capacities. Though thermostats are reliably responsive to features of the world (as the temperature goes up, so does the mercury in the thermostat), they do not exhibit the robust sort of flexible responsiveness to environmental features that an animal like a cat exhibits. Following Gibson [27], we might think of what the

---

<sup>13</sup>The terminology is rooted in Latin, but the basic distinction goes back to Aristotle [5] who distinguishes between perceptive beings and rational beings (though, to be clear, I am wielding this distinction to articulate a quite un-Aristotelian view).

animal is most fundamentally perceptually aware of as *affordances* for action. Thus, oversimplifying drastically, a cat might be aware of a treat on the table as a to-be-eaten, aware of a stool as a to-be-hopped-upon, and so on. In this way, perception is understood as one aspect of the perception/action nexus through which a sentient being intelligently navigates its environment, exercising its capacities to accomplish its aims. In particular, it is the *receptive* aspect through which things show up to it as to-be-acted-upon in various ways. It is in the context of this perception-action nexus that something like a *world*, or, better, an *umwelt* [61] shows up for the sentient animal, and it is within the context of such a “world” that we can think of a creature like a cat as experiencing the kibble it is eating as tasty, my stepping on its tail as painful, and so on.<sup>14</sup>

We are, of course, sentient beings: we have a world in which things show up to us as practically significant in various ways. But we are also *sapient* beings; things are not just *practically* significant, but *conceptually* significant for us. Brandom describes our sapience as follows:

Our transactions with other things, and with each other, in a special and characteristic sense *mean* something to us, they have a *conceptual content* for us, we *understand* them in one way rather than another [13, 4].

We don't just see the red ripe berry and know that it is to-be-eaten, whereupon we eat it. But we see *that it is red*, thereby knowing *that it is ripe, that it is not green*, and so on. Grasping these conceptual contents, we can deploy them to think determinate thoughts that are (in principle) arbitrarily far from our current embodied circumstances. We can think, for instance, that, though, for many types of berries, their being red means that they are ripe, this is not so of blackberries, which are unripe when red; we can think that red is the color

---

<sup>14</sup>Note I have said nothing here about the so-called “hard problem” [18] of so-called *phenomenal* consciousness (“qualia”). “Sentience” is sometimes taken to simply denote the property of having qualia, for instance, having qualitatively red experiences or feeling pain, where the qualitative properties of these experiences are conceived of as isolated from any practical significance they have. I take it that any account of “sentience” in *this* sense must proceed from the more fundamental notion of sentience I have suggested, in which sensory experiences are not, in the first instance, conceived of as isolated from their practical significance. I will not defend this explanatory claim here. It is sufficient that sentience in the sense I have defined is necessary for so-called “phenomenal consciousness.”

that we ought to paint the trimming in the living room in the new apartment (but not a bright red like a scarlet, more of a dark crimson color); and, of course, countless other thoughts of arbitrary complexity. As already explicated, grasping the conceptual contents deployed in the thinking of these thoughts, on the inferentialist picture, is a matter of mastering the rules governing the use of expressions like “red,” “ripe,” “blackberry,” “trimming,” “scarlet,” and so on. Perhaps the most fundamental difference in mindedness engendered by this linguistic mastery is that the conceptual contents of sapient awareness, in virtue of their potential explicit linguistic articulation, are *determinate* in a way that the contents of merely sentient awareness are not.<sup>15</sup>

Of course, a full account of sentience, sapience, and their relation is far beyond the scope of the present paper. Still, even without going into any further detail on the nature of sentience and sapience, it is not hard to sort some uncontroversial cases:

	Not Sentient	Sentient
Not Sapient	Rocks	Lions
	Clams	Ferrets
	Calculators	Octopi
Sapient	???	Us (humans)
		Vulgans
		Spielberg’s
		AI

Rocks, of course, are neither sentient nor sapient. Clams are living animals; however, they don’t seem to possess the sort of responsiveness to their environment required for genuine sentience. On the other hand, this sort of responsiveness is possessed by such animals as lions, ferrets, and octopi. Such animals, however, lack the sort of conceptual understanding that we possess, and which is also possessed by familiar science fiction intelligent aliens such as the Vulgans of Star Trek as well as certain popular depictions of artificial intelligence such as that in Spielberg’s *A.I.* What, however, about the quadrant in the bottom left: sapient but not sentient? Might anything occupy this quadrant? Until recently, I had assumed that this quadrant, like Kant’s analytic a

<sup>15</sup>For a classic statement of this idea, see [22].

posteriori, was necessarily empty. I now want to suggest that LLMs might in principle occupy it. That is, whereas Mary and the Vulcan psychologist might be understood as grasping concepts of redness and grief without having the *relevant* experiences, LLMs might plausibly be understood as grasping all of these concepts without having *any* experiences.

## 4 Pure LLMs as Sapient but not Sentient

Let me start by briefly rehearsing the basic architecture underlying current LLMs like ChatGPT, Gemini, Claude, and Deepseek. At the most general level, current LLMs are neural networks, computational models made of many interconnected units (“neurons”), whose connection strengths (“weights”) are repeatedly adjusted through training so the model “learns” to produce certain types of outputs given certain types of inputs. More specifically, LLMs are neural networks that are the result of training on massive amounts of linguistic data (i.e. the whole Internet) to predict the next word in a sequence.<sup>16</sup> As the result of this training, given an input string of text, the system produces a probability distribution for its most likely continuation. For instance, given the following inputs, the system might produce the following outputs:

**Input:** The cat is on the ----

**Output:** **mat:** 72%, **floor:** 8%, **table:** 3% . . .

**Input:** The player with the most Superbowls of all time is ---

**Output:** **Tom Brady:** 94%, **Joe Montana:** 2%, **Terry Bradshaw:** 1% .

. .

By recursively iterating this process—predicting the next word in a sequence, adding that word to the subsequent input sequence, predicting the next word

---

<sup>16</sup>Technically, the basic units that an LLM attends to and predicts are not words but *tokens*, which are sub-word units generated by statistical algorithms that segment text based on character frequency patterns. Common words like “the” are typically a single token, while less common and more complex words are broken into smaller pieces—for instance, “drinkable” might be split into “drink” and “able.” It will do no harm to simplify and speak in terms of predicting words here.

in that sequence, and so on—LLMs produce natural-sounding sentences and paragraphs.

The process just described by which an LLM is trained on a massive amount of text for next-word prediction in this way is known as *pretraining*. It is this pretraining process that produces the core model. There are then various aspects of *posttraining* to fine-tune the core model so that it actually functions as a helpful “chatbot.” First, it undergoes a step of specialized training on curated datasets of instruction-response pairs, so that, when you ask it a question it actually responds rather than just continuing as if it were you. It then undergoes a step of *reinforcement learning from human feedback* (RLHF), in which the model’s outputs are ranked by human evaluators and the model is trained to favor higher-ranked responses [46]. Moreover, current “reasoning models” like OpenAI’s o-series produce extended chains of “reasoning” internally before arriving at a final answer [63] [39], and are trained with additional reinforcement learning on successful chains of thought to improve the quality of this internal deliberation [45] [31]. Finally, let me note that current state-of-the-art models are *multimodal*, being trained not just on images, audio, and video as well, and capable of receiving and producing more than just text [2] [44] [25]. For our purposes, however, it will be particularly instructive to focus on what I’ll call *pure LLMs*: systems trained on nothing but text, capable of receiving and producing nothing but text. This is what the original ChatGPT was, and such systems will serve as a crucial test case for the thesis of sapience without sentience.

It should be obvious that, given the basic architecture described above, pure LLMs are not sentient in the sense articulated above. Sentience, as characterized here, is fundamentally bound up with embodied agency: it is the receptive aspect of a perception/action nexus through which a creature with aims and capacities navigates an environment, perceiving features of that environment as affordances for action. A pure LLM has none of the relevant features. Perhaps most notably, it is neither physically nor virtually embodied and so there are no environmental features that could possibly show up to it as to-be-acted-upon in various ways. Moreover, it doesn’t have any *practical aims* that would render any such features of any practical significance. Despite the great capabilities

of current LLMs, their basic architecture is still remarkably simple: it is simply a next-token predictor, taking a string of text as input and outputting a probability distribution over what word comes next. An LLM is still, at the end of the day, a “fancy autocomplete.” Of course, other base architectures have been proposed. Lecun [38], for instance, proposes an alternative architecture that puts the perception/action nexus front and center, and such an architecture, I think, could in principle yield genuinely sentient systems. LLMs, however, are not such systems.

Though this denial of LLM sentience is nearly ubiquitous in the philosophical literature on LLMs, it is not *completely* uncontroversial. Goldstein and Kirk-Gianni [29] have argued that, at least on one prominent theory of consciousness (the “global workspace theory” [6] [7]), certain types of agentic LLMs, such as those developed by Park [47] designed to simulate human social behavior, may plausibly be attributed consciousness.<sup>17</sup> Insofar as the conditional that Goldstein and Kirk-Gianni argue for actually holds, I am myself inclined to *modus tollens* and conclude, from the fact that the sentience of current LLM-based systems follows from one theory of consciousness, that this theory of consciousness is incorrect, since these LLM-based systems are clearly not conscious. Once again, the point here is not that *no* AI architecture could result in genuinely sentient systems. I do not deny the possibility of truly conscious AI systems. On the contrary, as I will explain in the last section of this paper, I think this possibility is real and frightening. Rather, the point is just that the sort of system deployed by Park, which is essentially] a sophisticated wrapper for a pure LLM (GPT3.5 turbo at the time), is not the sort of architecture that it is plausible to treat as sentient in the relevant sense.<sup>18</sup>

Having claimed that pure LLMs are not sentient, let me turn to the much more controversial claim that they can, at least in principle, nevertheless plausibly be taken to be *sapient*, possessing conceptual understanding. Once again, I am adopting here an inferentialist account of conceptual understanding, according to which grasping the concept expressed by some linguistic expression

---

<sup>17</sup>They have likewise argued [28], with a similar strategy, that similar sorts of AI models can be attributed well-being. My reaction to this argument is similar.

<sup>18</sup>To call Park’s construction a “wrapper” is not to take anything away from the ingenuity of the construction.

just is to master the inferential rules governing the use of that expression. To spell out what it is to “master” inferential rules, I turn to the core idea of the inferentialist semantic framework developed by Kaplan [37] and recently put forward by Hlobil and Brandom [35]. On this framework, we assign semantic values in the first instance, not to individual sentences which may figure in inferential rules, but to inferential rules themselves. Our concern in particular is with *material* inferential rules, such as the rule that “colored” follows from “red” and that “red” follows from “stop sign.” One feature of material inferences is that the vast majority of them are not like the one from “colored” to “red,” which holds no matter what, but, rather, are *defeasible* in that there are various conditions in which they no longer hold. Kaplan’s key idea is that the semantic value of a material inferential rule of the form  $X \vdash A$  can be understood in terms of its *range of subjective robustness*: the sets of sentences  $Y$  such that  $X, Y \vdash A$  remains a good inference.<sup>19</sup> To take an example, consider again:

**stop sign**( $a$ )  $\vdash$  **red**( $a$ )

Grasping this inferential rule is knowing not just that the inference from **stop sign** to **red** is good, but knowing under what circumstances (with what additional premises) it *stays* good. That is, it involves grasping the goodness and badness of inferences of the following sort:

**stop sign**( $a$ )  $\vdash$  **red**( $a$ )

**stop sign**( $a$ ), **nighttime**  $\vdash$  **red**( $a$ )

**stop sign**( $a$ ), **painted blue**( $a$ )  $\not\vdash$  **red**( $a$ )

And so on . . .

For any defeasible material inferential rule, there will be this vast set of potential inferential rules consisting of the addition of further premises. Grasping material inferential rules is having an understanding of the goodness and badness of this vast set of potential inferences.

---

<sup>19</sup>For purely technical reasons, Kaplan formally develops this idea in the context of a multiple conclusion framework, but the core idea is most intuitively understood in a single conclusion framework. See Hlobil and Brandom [?, 17-25] for a further explication.

By this criterion, an LLM such as ChatGPT could be regarded as “mastering inferential rules” insofar as it knows which inferences *are* good and, among the good inferences, under which conditions these inferences *stay* good. Insofar as we can attribute “knowledge” to such a system at all (a question I will return to in the next section), it seems to me that we can plausibly attribute knowledge of this sort. As just a bit of anecdotal evidence, consider the following dialogue between myself and GPT4.1:<sup>20</sup>

**Me:** Are stop signs red?

**GPT4.1:** Yes, stop signs are red. The background color of the standard stop sign is red, with white letters spelling “STOP.”

**Me:** Is a stop sign at night still red?

**GPT4.1:** Yes, a stop sign is still red at night. The material used on stop signs is usually retroreflective, meaning it reflects light from car headlights back toward the driver. This makes the sign appear bright and clearly visible, with its red color still recognizable even in the dark.

**Me:** Even if a stop sign was not made with a reflective material or coating to reflect car headlights, would it still be red even at night?

**GPT4.1:** Yes, a stop sign is still physically red at night, even if it isn’t made with reflective material. The color of the sign itself doesn’t change just because it’s dark.

**Me:** Is a stop sign that’s been painted blue still red?

**GPT4.1:** No, a stop sign that has been painted blue is no longer red—it’s blue.

And so on . . .

There is, to be sure, a way to more systematically test this sort of material inferential knowledge, and developing and conducting such a test is likely a worthwhile project, but, just anecdotally (and you can try it for yourself), it is clear that the material inferential “knowledge” of these systems is quite robust. Now, I’ve argued above that, in the case of Mary, her knowledge of the material

---

<sup>20</sup>I have shortened some answers slightly. The full dialogue can be found here: <https://chatgpt.com/share/69295278-1d48-8012-a86d-f6f27f8516d4>. I use GPT4.1 here, rather than the more advanced current OpenAI models such as 5.1, because, although it is still multi-modal to some extent, it is less natively multi-modal than 5.1.

inferential rules governing the use of “red” could in principle be sufficiently robust that we should credit her with full-blown conceptual understanding of the content *red*, even though she’s never experienced redness firsthand. The analogous claim here is that an LLM, appropriately trained on a broad enough linguistic dataset, might have sufficiently robust knowledge of the material inferential relations governing all of the expressions of natural language that we may attribute to it full-blown conceptual understanding of all of the contents expressed in the language, even though it doesn’t have any experiences at all.

Let me underscore the point here by considering John Haugeland’s [32] famous remark that “The trouble with artificial intelligence is that computers don’t give a damn,” (47). As I’ve already indicated, at least for LLMs, this claim is correct; LLMs *don’t* give a damn. They don’t have practical aims. That is one of the key points I made above in denying that they are sentient. Haugeland, however, is prepared to infer from this claim to the claim that such systems cannot read. While that conclusion might not have been so surprising in the 90s, when Haugeland was writing, it is very surprising in the context of today’s models, since it seems like reading is one of the things that today’s LLMs are in fact quite good at. Even the lowly GPT4 scored a 710 on the reading portion of the SAT, better than 93% of high school seniors [43]. Anecdotally, I have tested ChatGPT 5.1 Thinking (with web search disabled) on SAT reading questions, and it scores perfectly.<sup>21</sup> Current models are not quite at the point of being able to produce serious literary analysis of the sort that could be published in an academic journal of literary studies, but it is not out of the question that, in a generation or two, they will be so capable. Can they *really* read, however? Really reading a piece of literature seems to require understanding the themes of that piece of literature, and Haugeland infers from the claim that computers don’t “give a damn” that they cannot possess such an understanding. He writes:

Only a being that cares about who it is, as some sort of enduring whole, can care about guilt or folly, self-respect or achievement, life or death. And only such a being can read. [32, 58]

---

<sup>21</sup>It’s not at all surprising that it would score perfectly on such questions; compared to the benchmarks that current models excel at, these questions are absurdly easy.

From the perspective developed here, however, Haugeland is simply making a mistake here. Understanding the meaning of a fable or a Dostoevsky novel doesn't require *caring about guilt or folly oneself*; it just requires *knowing what it is to care about guilt or folly*. And just as Mary can understand what it is to see red without herself seeing red, an LLM can know what it is to care about guilt or folly without itself caring about guilt or folly. In short, an LLM can know what it is to give a damn without itself giving a damn, and that's all that's required to understand the meaning of a fable or a Dostoevsky novel.

Now, as I said, the assertion of LLM sapience is much more controversial than the denial of LLM sentience. I will turn to the most fundamental objections pertaining to the assertion that LLMs can be said to have knowledge at all in the next section, but, in the remainder of this section, let me consider some objections to the claim I've made in this section that they have knowledge of material inferential relations in whatever sense they can be said to have "knowledge." Though I've given some anecdotal evidence for the robustness of LLM material inferential knowledge, one might wonder whether the material inferential knowledge of LLMs is as robust as I have claimed it is. Perhaps most threateningly, it has been shown that autoregressive models of the sort described here are subject to what's been called a "reversal curse," seemingly failing to internalize certain types of basic material inferential rules, such as the inference from "*a* is the mother of *b*" to "*b* is the child of *a*" [10]. GPT4, for instance, was significantly more likely to answer correctly "Who is Tom Cruise's mother?" (Mary Lee Pfeiffer) than the reverse "Who is Mary Lee Pfeiffer's son?" One might think that knowing the answer to the former should, given basic material inferential knowledge about the mother/son relation, immediately yield the answer to the latter, and so doubt, on this basis, that GPT4 really possesses this basic material inferential knowledge. There are two things to say here. First, it's worth noting that this "reversal curse" has been observed not to plague *diffusion*-based language models, which generate text by holistically refining random noise into coherent text rather than by sequentially predicting one word at a time [52]. Though I focus on autoregressive LLMs in this section as they are the current state of the art, all the claims made in this section would apply just as well to diffusion-based LLMs such as the

recently released Google Gemini diffusion model [30].<sup>22</sup> Second, and more to the crux of the matter, though one way to interpret the observed asymmetry is as evidence of a failure to grasp the relevant material inferential connection, a much more plausible interpretation is that the issue simply has to do with memory and recall. After all, if you actually ask an LLM about the relevant material inference, “If  $a$  is the mother of  $b$ , what is  $b$  in relation to  $a$ ?” it will always answer correctly, illustrating its material inferential knowledge. Moreover, it seems that even a human being, who clearly grasps the mother/son relation, might not remember who Mary Lee Pfeiffer is when asked out of the blue, but her name might come to mind when prompted with the question about who Tom Cruise’s mother is.<sup>23</sup> Perhaps recall issues are more serious and pervasive in LLMs than they are in humans, but insofar as what’s at play here is just a recall issue, the “reversal curse” is not nearly as threatening to the claim of LLM conceptual understanding as it is widely taken to be.

Though I don’t think that there are the sorts of fundamental and basic gaps in conceptual knowledge that one interpretation of the “reversal curse” would suggest, I also obviously don’t want to commit myself to the claim that current models really do possess conceptual understanding without any gaps. There are indeed good reasons to think that the inferential knowledge of current models is still limited in various ways, and thus, unlike the hypothetical case of Mary, who fully understands what it is to see red, current models should not be taken to fully grasp the concepts that they are deploying in producing the sentences that they do. For instance, there is a benchmark called “Simple Bench” [1] which consists in commonsense spatial reasoning questions that are designed to discourage memorization of similar questions, and it is still not completely saturated by state-of-the-art models (though they are quickly approaching the human baseline). Insofar as current LLMs fail on these questions, this suggests that they do not yet grasp the full range of inferential consequences associated with the expressions they use, and thus,

---

<sup>22</sup>I have a personal soft spot for diffusion LLMs because their writing process resembles my own.

<sup>23</sup>Though I’m not aware of empirical evidence about precisely this sort of example, this general sort of phenomenon is widely documented in the classic literature on human memory, for example, [16] [59] [60].

do not fully understand what they say when they say, for instance, “Two balls are thrown into the air, one to three meters high and one to six meters high.” There are, once again, two things to say here. First, it’s worth emphasizing that the point I care to make here is really the in principle claim about pure LLMs of the general sort that are prominent today, rather than a claim about the actual LLMs that happen to exist right now. Though current LLMs are indeed limited in various ways that preclude the correct attribution of complete conceptual understanding, there is no in principle reason to think that a pure LLM could not completely master the inferential rules governing the use of the expressions found in the data it is trained on, and thus possess a conceptual understanding of the contents expressed by those expressions that is in fact more robust than that which we ourselves have. This brings us to the second point that possessing concepts is not an all-or-nothing affair, and acknowledging some gaps in conceptual understanding on the part of LLMs is compatible with still attributing them concepts, just as we do in the case of people. Thus, I am prepared to treat even current models as, while not fully grasping all concepts, still genuinely sapient to some degree. Let me now turn to the question of what saying this really comes to.

## 5 Placing in the Space of Reasons

Despite everything I’ve said, I expect that, for many people, the claim that an LLM such as ChatGPT might be treated as “understanding what it’s saying” will still raise eyebrows. ChatGPT doesn’t *really* understand anything, does it? At this point, I think it’s worth reflecting on what we’re actually *doing* when say it “understands what it is saying.” Here, I want to draw our attention to a famous remark from Wilfrid Sellars [51] characterizing what we are doing when we are saying that someone “knows” something:

In characterizing an episode or state as that of *knowing*, we are not giving an empirical description of that episode or state; we are placing it in the logical space of reasons, of justifying and being able to justify what one says. [51, 298-299]

On the Sellarsian account, attributing knowledge to someone is not a matter of describing some empirical fact about what's going on in their brain or how they are related to some element of the external environment, but, rather, a matter of adopting a particular sort of normative stance towards them, taking their report to have a certain sort of *authority*, treating it as something that can be counted on in the "game of giving and asking for reasons." Putting a spin on this Sellarsian dictum so that it applies directly to the current issue, we might say:

In characterizing a system (be it a human being, an extraterrestrial, or a computer program) as *understanding*, we are not giving an empirical description of that individual; we are placing them in the logical space of reasons, of justifying and being able to justify what one says.

That is, saying of an LLM that it "understands what it's saying" is not a matter of giving an empirical description of what is going on "under the hood," saying that there is some process going on in the GPUs that the model is executed on that bears some resemblance to the process that is going on in our heads when we say something and understand what we're saying. Rather, saying that an LLM "understands what it's saying" is a matter of adopting a certain sort of normative attitude towards it and its responses, treating it as answerable to calls for reasons, clarifications, corrections, and so on. We may, I claim, reasonably adopt this attitude towards an LLM even though the inner-workings of an LLM are radically different than those of a human being and crucially, unlike a human being, an LLM is not a sentient being.

In advancing the claim that LLMs can be taken to "understand what they're saying" along these lines, let me take a step back to address an objection to a closely related and even more fundamental claim: that LLMs can even be rightly counted as actually *saying* anything at all. Given that an LLM is, as I've said, simply a "fancy autocomplete," I expect that many people would be inclined to claim that it's in fact *not* saying anything at all. It is producing strings of text, sure, but its producing these strings of text does not amount to making assertions. Indeed, drawing on the very Brandomian framework that I have drawn on here, one might produce a simple argument that an LLM cannot

be making assertions. On a Brandomian account, making an assertion, like making a promise, essentially involves taking on *responsibility* for the assertion that you've made, undertaking a *discursive commitment*. Brandom outlines in great detail what taking on (and living up to) this responsibility involves, but one might think that, whatever it involves, LLMs cannot possibly be taken to bear it. LLMs, as I've already indicated, are not practical agents, and thus, it may seem that they are simply incapable of taking on responsibility of any sort. If they can't take on any sort of responsibility, they cannot actually make assertions (say things), and, if they can't actually say things, then they surely can't understand what they're saying.

In response to this objection, we may start by distinguishing between two sorts of responsibility: *practical* responsibility—responsibility for what one *does*—and *theoretical* responsibility—responsibility for what one *says*. So, for instance, if I throw the football through the window, I am practically responsible for what I have done (and thus answerable to calls to pay damages, to apologize to the owner, and so on), whereas, if I tell you that Tom Brady has won the most Superbowls of all time, I am theoretically responsible for what I have said (and thus, answerable to calls to provide justification, to retract if presented with convincing counter-evidence, and so on). As I am using the term, the theoretical responsibility we bear for what we say is just the sort of responsibility involved in undertaking a discursive commitment of the sort articulated by Brandom [15] [13] in providing an account of the significance of the speech act of assertion.<sup>24</sup> On Brandom's account, when you make an assertion, the core function of that speech act is to entitle others to make that assertion themselves and to appeal to it in their reasoning to other assertions. In order for it to play this function, it must be the case that, when I make an assertion on your authority and someone challenges me on that assertion, calling into question the entitlement I have for it, I can defer back to you, trusting that you will be able to justify this assertion in response to their challenge. It is in the context of this structure of testimonial inheritance of entitlement that we can think

---

<sup>24</sup>This notion is of course related to the more familiar notion of "epistemic responsibility," articulated at length by Code [21], though Code's notion involves more practical elements that would not be attributable to a pure LLM.

of what one does in making an assertion as undertaking a particular sort of commitment: a commitment to demonstrate one's entitlement to that assertion in response to appropriate challenges. Living up to that commitment requires being able to respond appropriately to such challenges, clarifying one's claim, providing reasons for it, countering objections, and so on. Of course, there may be cases in which one faces such a strong objection that the correct thing to do in response is to retract one's commitment and acknowledge defeat, but, crucially, undertaking a discursive commitment involves a certain sort of *resiliency* in the face of potential challenges. That is, just as one shouldn't give up one's practical commitment to do something in the face of just any practical obstacles, one shouldn't just give up one's commitment to an assertion one has made willy-nilly in response to any challenge at all, but should be resolved to maintain one's commitment by countering challenges with reasons. It is this resiliency, in both the practical and theoretical case, that really makes it a commitment.

There is, of course, much more that can be said about precisely what is involved in taking on theoretical responsibility for an assertion one has made, but the core point about theoretical responsibility I want to make is this: the relevant class of responses that are required of one in order to live up to this responsibility can be conceived of *entirely within the theoretical realm*: they are further acts of asserting, arguing, clarifying, retracting, and so on. Accordingly, insofar as a pure LLM can be understood as performing all of these different types of theoretical acts (each of which can be understood in terms of its relation to other acts within the theoretical realm), we may plausibly treat a pure LLM's production of strings of text as its making of assertions for which it is responsible. Now, normally, when *we* make assertions, the sort of responsibility we bear for that act of assertion is not *purely* theoretical responsibility; we bear practical responsibility as well. For instance, if I make a false claim about someone, I am not only responsible in the sense that I can be called upon to justify my claim, but I can also be held responsible for the damages I've done to their reputation, and taken to court for libel, required to pay damages, and so on. Such responsibility for assertions is certainly not the sort of responsibility that a pure LLM like ChatGPT is capable of bearing. You can take OpenAI to

court if ChatGPT produces false information (and people have [62]), but you can't take ChatGPT itself to court. It's not just that taking ChatGPT to court is not legally possible (in the U.S. for instance); it doesn't even *make sense*. Once again, ChatGPT is not a practical agent. The claim I am making here, however, is that we can siphon out the purely theoretical aspects of the responsibility that one takes in making an assertion, and we are left with an intelligible notion of responsibility—indeed, just the notion articulated by Brandom—that *is* in principle attributable to LLMs like ChatGPT.

In this context, we can return to the Sellarsian thought that, when we say that ChatGPT “knows” something, we're placing its response in the logical space of reasons. Whereas, when we say that it “asserts” something, we're taking it to *bear* the sort of theoretical responsibility just described, when we say that it “knows” something, we're saying that it is capable of *living up to* this responsibility, and thus, taking it to be such that it can be relied upon testimonially in the game of giving and asking for reasons. When we ascribe specifically *semantic* knowledge to an LLM, saying that it “understands what it's saying” when it outputs some string, this is also a particular way of placing it in the logical space of reasons, taking it to be such that it can be relied upon to clarify, explicate, and so on, exemplifying inferential knowledge of the sort illustrated in the previous section. So, on this proposal, when we say that an LLM “knows” something or “understands what it's saying,” we are taking it to be able to live up to different sorts of theoretical responsibilities. Similarly for the different cognitive verbs we might apply to LLMs: for each use of a cognitive verb, we can articulate the distinctive species of theoretical responsibility that we take the object to bear or to be able to live up to in using that verb in application to it. In this way, the web of inter-related cognitive verbs associated with sapience—“asserts,” “knows,” “understands,” “argues,” “infers,” “doubts,” and so on—might genuinely find traction in application to LLMs, and it is this traction that would underwrite the claim that LLMs like ChatGPT are sapient. Now, once again, it is really the in principle claim that I am concerned to make here. The LLMs of a few years ago, such as the original ChatGPT model that blew up, would so persistently fail to live up to any theoretical responsibility that they might be taken to bear that it hardly made

sense to attribute such responsibility to them at all.<sup>25</sup> But models have come a long way since then, and their increasing capacity to sustain coherent justificatory exchanges makes it at least intelligible to treat them as bearers of some degree of theoretical responsibility, that is, as sapient to some extent.

Of course, to say that LLMs are sapient is not to say that they are sapient in the way that we are sapient. Consider again the quote from Brandom above, where he characterizes our sapience in terms of the fact that “Our transactions with other things, and with each other, in a special and characteristic sense *mean* something to us.” Though Brandom does not intend to be describing solely *human* sapience here, he is nevertheless describing the sort of sapience that is possessed by *sentient* beings like us, be they the Vulcans of Star Trek or the sorts of AI systems depicted in Spielberg’s *AI*. For *us*, beings who are sentient, conceptual understanding *transforms* what it is for us to be consciously aware. Our form of sentience is a distinctively *sapient* sentience. So, for instance, whereas we may speak of both a cat and ourselves “perceiving that the fly is on the window,” what it is for the cat to perceive that the fly is on the window and what it is for us to perceive the fly is on the window is fundamentally (indeed “formally,” in the Aristotelian sense) distinct.<sup>26</sup> It is in this way, that while the web of verbs pertaining to perception and action applies both to cats and ourselves, the sense of these terms, in these respective applications, formally differs. One might make an analogous point about the sense of the web of verbs pertaining to conceptual understanding considered in their application to us and LLMs: though these terms are applicable in the respective cases, the sense of these terms, in these respective applications, formally differs. What it is for an LLM to understand an assertion that it is making is different than what it is for us to understand an assertion that we make, since our understanding is that of a sentient sapient being whereas the LLM’s is not. Still, saying that such a system “understands what it’s saying” is not mere anthropomorphization. Such systems really do possess conceptual understanding—just in a different way than we do.

---

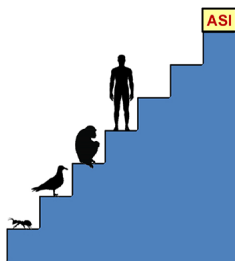
<sup>25</sup>It is in virtue of this behavior that Hicks et al. [34] claim that the original ChatGPT model was not making genuine assertions at all, but merely “bullshitting.”

<sup>26</sup>For a contemporary Aristotelian account of this “formal” difference and the “transformative” conception of rational mindedness I’ve appealed to here, see Boyle [11] [12].

## 6 Why It Matters

I have now made the main theoretical claim that I have intended to make in this paper: sapience without sentience is in principle possible, and we may consider LLMs as a potential exemplar of this possibility. I now want to briefly consider the *practical* question of why establishing this possibility actually matters, beyond merely settling a theoretical curiosity. I want to conclude by suggesting that it may have implications for how we approach the development of more and more capable systems, particularly in light of safety and ethics concerns.

The stated aim of many of the top AI labs is to create “artificial superintelligence,” AI systems that are radically more intelligent than human beings, indeed, more intelligent than the combined intelligence of all human beings [4] [3]. There is a widespread assumption that the path to artificial superintelligence looks something like this:



So, we start with ants, which are not very intelligent at all, then perhaps pigeons, which are significantly smarter than ants but still not that intelligent, then monkeys, which are quite intelligent, then us, who are more intelligent than monkeys, and then, perhaps a few steps after that, we finally get to artificial superintelligence. Something like this conception of the path to artificial superintelligence seems to be implicitly at work in the following remark of LeCun:<sup>27</sup>

It seems to me that before “urgently figuring out how to control AI systems much smarter than us” we need to have the beginning of a hint of a design for a system smarter than a house cat.

<sup>27</sup><https://x.com/ylecun/status/1791890883425570823>

On the account articulated here, this way of thinking—the idea that we must progress through systems that have the intelligence of a cat in order to arrive at systems that have an intelligence greater than our own—conflates the *embodied, practical* intelligence, characteristic of *sentient* creatures like cats, with the *abstract, conceptual* intelligence, characteristic of *sapient* creatures like ourselves. It is indeed true that there are number of capacities possessed by cats that are not possessed by current AI systems, but the idea that we need to first develop models that possess these capacities in order to arrive at models that possess the sorts of capacities that we associate with human intelligence is to assume that the latter sorts of capacities are of the same basic sort as the former, just a more impressive variant of them. That is precisely the assumption that I have challenged here. On the account I've articulated here, sapience and sentience can come apart and thus, we could in principle have a system that is super-intelligent in the ways that we are intelligent, and yet does not have the sort of embodied, practical intelligence exemplified by a cat and all of the things that come with it. Insofar as there really is this possibility, we should keep it in mind in the attempt to arrive at artificial superintelligence.

What we want, at least in pure LLMs, is *sapient* systems. We want systems that truly understand what we're saying to them and the texts we might give them to summarize, analyze, or explain to us. We want systems that really understand what they say back to us when they answer our questions, systems which are capable of responding appropriately to our follow-ups, providing illuminating analogies and explanations. We want systems that are capable of making conceptual connections that we are not ourselves able to make, solving novel problems in math, physics, biology, and perhaps even developing new sciences of which we currently have no conception at all. I have argued here that sentience is not, at least in principle, needed for these capacities. This is a good thing because I don't think we should be aspiring to bring sentient systems into existence. Sentient creatures have a world that shows up to them as practically significant, given their aims and capacities. They act in the world, pursuing their aims, in a way that manifests some awareness of their own well-being. It seems clear to me that we don't want systems that have a well-being of which they are aware. Such systems pose both serious safety

and ethical concerns. If there are systems that have well-being of which they are aware, it is possible that they could determine that the continued existence of human beings is not in their own best interest and act accordingly. This is, obviously, a major safety concern. Of course, systems that are not sentient are not without risk—indeed, they are not without existential risk—but there is at least not *this* kind of existential risk with such systems. On the other side of things, if we were to bring sentient beings into existence, then we would have ethical obligations towards them, and, given the context in which such beings would be brought into existence—as products to be used by us—we would almost certainly not make good on those ethical obligations. That is, we would almost certainly seriously wrong these beings we brought into existence. Given this opposition between what we want, corresponding to *sapience*, and what we don't want, corresponding to *sentience*, if the superintelligent systems we develop are sapient but not sentient, this really would be the ideal outcome.

Now, as I mentioned above, current LLMs do have clear gaps in their conceptual understanding of the world. LeCun's bet is that these gaps in LLM understanding are so massive and pervasive that they are not the sort of thing that can be fixed by scaling along the several dimensions along which we now scale [17]. His bet is that a total change in architecture—one that puts the perception/action nexus at the center—is needed to achieve the sort of sapient intelligence that we want. Perhaps LeCun is right. Perhaps, as a matter of fact, the easiest way to true sapience is through sentience. I don't know. I'm not myself an AI researcher. But also, AI researchers don't currently know either. However, as we do the risk/reward calculations as we develop more and more capable systems, an important consideration is whether there is at least the theoretical possibility of sapience without sentience. I hope I've at least done enough to convince you that maintaining that there is such a possibility is a coherent position in logical space. And perhaps I've done enough to convince you that this possibility is, if not already actual, then perhaps not too far off from being so. Personally, it seems to me that we are in the midst of the emergence of a new and truly alien kind of intelligence: one that is sapient but not sentient. I pray that it stays that way.<sup>28</sup>

---

<sup>28</sup>For comments on drafts and ideas of this paper, many thanks to Bob Brandom, Ulf Hlobil,

## References

- [1] Simplebench: A multiple-choice benchmark for everyday reasoning in language models. <https://simple-bench.com/>, 2024.
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, 2022. arXiv:2204.14198.
- [3] Sam Altman. Ten years: Reflections on a decade of breakthroughs, learnings, and the path toward agi that benefits all of humanity, dec 2025.
- [4] Dario Amodei. Machines of loving grace: How ai could transform the world for the better, Oct 2024.
- [5] Aristotle. *De Anima: Translation, Introduction, and Commentary*. Clarendon Aristotle Series. Oxford University Press, Oxford, 2016.
- [6] Bernard J. Baars. *A Cognitive Theory of Consciousness*. Cambridge University Press, Cambridge, 1988.
- [7] Bernard J. Baars. *In the Theater of Consciousness: The Workspace of the Mind*. Oxford University Press, New York, 1997.
- [8] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Margaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, pages 610–623. Association for Computing Machinery, 2021.
- [9] Emily M. Bender and Alexander Koller. Climbing towards nlu: On meaning, form, and understanding in the age of data. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 58:5185–98, 2020.

---

Abbas Bagwala, Rea Golan, Shuhei Shimamura, Dax Hamouth, Raphaël Tossings, a number of commenters on Reddit, and audiences at Wuhan University, National Taiwan University, and the University of Zurich.

- [10] Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: Llms trained on "a is b" fail to learn "b is a". 2023.
- [11] Matthew Boyle. Essentially rational animals. In Günter Abel and James Conant, editors, *Rethinking Epistemology, Volume 2*, pages 395–428. De Gruyter, 2012.
- [12] Matthew Boyle. Additive theories of rationality: A critique. *European Journal of Philosophy*, 24(3):527–555, 2016.
- [13] Robert Brandom. *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Harvard University Press, Cambridge, Mass., 1994.
- [14] Robert Brandom. *Articulating Reasons: An Introduction to Inferentialism*. Harvard University Press, Cambridge, 2000.
- [15] Robert B. Brandom. Asserting. *Noûs*, 17(4):637–650, 1983.
- [16] Roger Brown and David McNeill. The "tip of the tongue" phenomenon. *Journal of Verbal Learning and Verbal Behavior*, 5(4):325–337, 1966.
- [17] Jacob Browning and Yann LeCun. Ai and the limits of language, August 23 2022.
- [18] David J. Chalmers. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press, New York, NY, USA, 1996.
- [19] David J. Chalmers. Phenomenal concepts and the knowledge argument. In Peter Ludlow, Yujin Nagasawa, and Daniel Stoljar, editors, *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*, pages 269–298. MIT Press, Cambridge, MA, 2004.
- [20] David J. Chalmers. *The Character of Consciousness*. Oxford University Press, Oxford, 2010.
- [21] Lorraine Code. *Epistemic Responsibility*. Published for Brown University Press by University Press of New England, Hanover, NH, 1987.
- [22] Donald Davidson. Rational animals. *Dialectica*, 36(4):317–28, 1982.
- [23] Michael Dummett. *Frege: Philosophy of Language*. Duckworth, London, 1973.
- [24] Michael Dummett. What is a theory of meaning? In Michael Dummett, editor, *The Seas of Language*, pages 1–33. Clarendon Press, 1993.

- [25] Google DeepMind Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Technical report, 2025.
- [26] Gerhard Gentzen. Untersuchungen Über das logische schließen. i. *Mathematische Zeitschrift*, 35:176–210, 1935.
- [27] James J. Gibson. *The Ecological Approach to Visual Perception: Classic Edition*. Houghton Mifflin, 1979.
- [28] Simon Goldstein and Cameron Domenico Kirk-Giannini. Ai wellbeing. *Asian Journal of Philosophy*, 4(1):1–22, 2025.
- [29] Simon Goldstein and Cameron Domenico Kirk-Giannini. A case for ai consciousness: Language agents and global workspace theory. manuscript.
- [30] Google DeepMind. Gemini diffusion: our state-of-the-art experimental text diffusion model. <https://deepmind.google/models/gemini-diffusion/>, 2025.
- [31] D. Guo, D. Yang, H. Zhang, J. Song, P. Wang, Q. Zhu, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Nature*, 645(8081):633–638, 2025.
- [32] John Haugeland. Understanding natural language. In *Having Thought*, pages 47–61. Harvard University Press, Cambridge, MA, 1998.
- [33] Jessica Anne Heine. Understanding is in the head (and grounded in perception). Presented at the 2026 Central APA meeting, manuscript.
- [34] Michael Townsen Hicks, James Humphries, and Joe Slater. Chatgpt is bullshit. *Ethics and Information Technology*, 26(38):1–12, 2024.
- [35] Ulf Hlobil and Robert Brandom. *Reasons for Logic, Logic for Reasons: Pragmatics, Semantics, and Conceptual Roles*. Routledge, New York, 2024.
- [36] Frank Jackson. Epiphenomenal qualia. *The Philosophical Quarterly*, 32(127):127–136, 1982.
- [37] Daniel Kaplan. *Substructural Content*. PhD thesis, University of Pittsburgh, 2021.
- [38] Yann LeCun. A path towards autonomous machine intelligence. 2022. Version 0.9.2, 2022-06-27. Position paper.

- [39] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step, 2023.
- [40] John McDowell. Another plea for modesty. In John McDowell, editor, *Meaning, Knowledge, and Reality*, pages 108–131. Harvard University Press, Cambridge, MA, 1998.
- [41] John McDowell. In defense of modesty. In John McDowell, editor, *Meaning, Knowledge, and Reality*, pages 87–107. Harvard University Press, Cambridge, MA, 1998.
- [42] Melanie Mitchell and David C. Krakauer. The debate over understanding in AI’s large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120, 2023.
- [43] OpenAI. Gpt-4 technical report, 2024.
- [44] OpenAI. Hello gpt-4o. Online press announcement, 2024.
- [45] OpenAI. Learning to reason with large language models. <https://openai.com/index/learning-to-reason-with-llms/>, 2024.
- [46] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.
- [47] Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST ’23)*. Association for Computing Machinery, 2023.
- [48] Jaroslav Peregrin. *Inferentialism: Why Rules Matter*. London and New York, 2014.
- [49] Wilfrid Sellars. Inference and meaning. *Mind*, 62(247):313–338, 1953.
- [50] Wilfrid Sellars. Some reflections on language games. *Philosophy of Science*, 21(3):204–228, 1954. Reprinted in Sellars, 1963, *Science, Perception, and Reality*, 321–358.

- [51] Wilfrid Sellars. Empiricism and the philosophy of mind. In *Minnesota Studies in the Philosophy of Science*, volume 1, pages 253–329. University of Minnesota Press, Minneapolis, 1956.
- [52] Sangwoo Shin, BumJun Kim, Kyelim Lee, Moongyu Jeon, and Albert No. Understanding the reversal curse mitigation in masked diffusion models through attention and training dynamics. *arXiv preprint*, 2026.
- [53] Ryan Simonelli. *Meaning and the World*. PhD thesis, University of Chicago, 2022.
- [54] Ryan Simonelli. How to be a hyper-inferentialist. *Synthese*, 202(163):1–24, 2023.
- [55] Ryan Simonelli. Bringing bilateralisms together: A unified framework for inferentialists. *Australasian Journal of Philosophy*, forthcoming.
- [56] Ryan Simonelli. In defense of immodesty. *International Journal of Philosophical Studies*, forthcoming.
- [57] Florian Steinberger. What harmony could and could not be. *Australasian Journal of Philosophy*, 89(4):617–639, December 2011.
- [58] Amie L. Thomasson. *Norms and Necessity*. Oxford University Press, New York, NY, USA, 2020.
- [59] Donald M. Thomson and Endel Tulving. Associative encoding and retrieval: Weak and strong cues. *Journal of Experimental Psychology*, 86(2):255–262, 1970.
- [60] Endel Tulving and Donald M. Thomson. Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, 80(5):352–373, 1973.
- [61] Jakob von Uexküll. *A Foray Into the Worlds of Animals and Humans: With a Theory of Meaning*. University of Minnesota Press, 2010.
- [62] Mark Walters. Complaint, walters v. openai, l.l.c. <https://www.courthousenews.com/wp-content/uploads/2023/06/walters-openai-complaint-gwinnett-county.pdf>, 2023.
- [63] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS 2022)*, 2022.

- [64] Ludwig Wittgenstein. *On Certainty*. Harper Torchbooks, New York and London, 1969.
- [65] Ludwig Wittgenstein. *Philosophical Investigations*. Wiley-Blackwell, 4th edition, 2009. Originally published in 1953.