Minds and Machines (2025) 35:46
https://doi.org/10.1007/s11023-025-09751-y

LLMs and the Logical Space of Reasons

Mirco Sambrotta[1]

Received: 12 December 2024 / Accepted: 22 October 2025 / Published online: 5 November 2025

Abstract

Can Large Language Models (LLMs), such as ChatGPT, be considered genuine language users? Can they truly understand the meanings of natural language? This paper adopts an inferentialist perspective, arguing that grasping the meaning of an expression is nothing but grasping the inferential role the expression plays. But roles are conferred by *rules*. An expression's contentfulness consists of its use being governed by *inferential rules*. Meaningful items incorporate norms of inference, which they are subject to. Thus, grasping meanings is mastering this bundle of rules. The paper questions whether LLMs, despite their ability to generate coherent text and perform inference-like tasks, follow such rules. However, unlike humans, LLMs are trained to detect *statistical patterns*. While it might be suggested that LLMs 'know' which words typically follow others based on statistical patterns, the norms they learn to master differ fundamentally from the norms of inference that govern language use. The paper concludes that the current version of LLMs, despite their advanced language processing capabilities, do not genuinely grasp or understand conceptual content and should, therefore, be viewed as simulations of language users rather than true participants in the logical space of reasons.

Keywords Large language models · ChatGPT · Inferential roles · Inferential rules · Rule-following

✉ Mirco Sambrotta
mirco.sambrotta@gmail.com

[1] Institute of Philosophy, V.V.I., Slovak Academy of Sciences, Bratislava, Slovakia

1 Introduction

A notable feature of the current generative AI boom is the machine processing of natural language by Large Language Models (LLMs). LLMs are neural networks, or better, current transformer-based neural natural language processing (NLP) systems,[1] such as GPTs and ChatGPTs,[2] which made available for the first time an AI with human-level performance on a wide range of cognitive tasks. One such achievement is language generation, which appears to approach human-level performance in many contexts.

These models can generate text by processing prompts (e.g., a short passage of text, often a single sentence), and autonomously produce coherent continuations, potentially achieving flawless performance across various conversational topics based solely on linguistic input. Unlike previous text generators like SCIgen, LLMs seem able to generate well-formed and meaningful texts as outputs often indistinguishable from those of humans. There is indeed compelling evidence indicating that they can even outperform many human writers. Ultimately, language processing in LLMs by generative AI easily goes undetectable by human scrutiny, passing in this way the well-known Turing Test (Turing, 1950) for determining whether a machine can demonstrate human intelligence Elkins and Chun (2020).

But can they really be considered genuine language users? The question is whether artificial neural nets can deal with and process conceptual contents. If not, should LLMs be treated (like automatic grammar and spell-checkers or translators) as producing written output without rising to the level of language user? If so, what else would they need to rise to the level of language user?

The difference, in the end, lies in the presence or absence of understanding, for meaning and understanding are co-ordinate concepts: meanings are, in the first instance, what one understands, and grasping a meaning is a kind of understanding. Are LLMs truly capable of understanding the meanings of natural language? Are there NLP systems that can genuinely grasp the content of our concepts, or should we regard their linguistic performance as merely a simulation of understanding—that is, nothing more than sophisticated syntactic manipulation?[3] But what, after all, does understanding really amount to?

The term "understanding" often evokes an internalist or mentalistic picture of hidden cognitive episodes, suggesting private mental states or processes accessible only to a conscious agent.[4] In the case of LLMs, however, such internal episodes may well

[1] The terms "transformer" or "transformer-based" refer to a new generation of NLP architectures pioneered by Vaswani et al. (2017).

[2] GPT (Generative Pre-trained Transformer) uses the decoder part and BERT (Bidirectional Encoder Representations from Transformers) uses the encoder part. In both cases, the resultant language models can then be used for various NLP tasks (Natural Language Processing).

[3] See Bender and Koller (2020) for a complaint of this kind, specifically applied to the claim that language models capture meaning or possess understanding.

[4] This conception has a long lineage in philosophy: beginning with Descartes, who grounded understanding in the inner operations of a thinking substance, i.e., *res cogitans* (1641/1984), and continuing through modern accounts. For example, Fodor's *Language of Thought* hypothesis conceives of understanding as the manipulation of internal mental representations (1975); Chomsky identifies linguistic understand-

be inapplicable, since these systems appear to lack any first-person perspective or awareness. This might suggest dispensing altogether with the notion of understanding from the outset when characterizing LLMs' language performance, as opposed to human language use.[5] Nevertheless, there are compelling reasons to retain this notion in discussions of language use—especially when comparing humans and LLMs. It allows us to distinguish genuine linguistic competence from mere simulation and preserves the possibility of asking whether a system truly grasps meaning, rather than simply producing outputs that appear appropriate. The concept provides explanatory utility, enabling discussion of semantic grasp and language use in ways that go beyond surface-level behavioral performance. Accordingly, appealing to the notion of understanding facilitates comparative analyses between humans and artificial systems, highlighting both the capacities that a system approximates and the gaps that remain before it can be considered a genuine language user. The aim of this paper is to provide a framework for retaining the notion of understanding without conceiving it as an inner state, thereby avoiding the pitfalls of internalist or mentalistic accounts.

2 Inferential Roles and Inferential Rules

Linguistic understanding is intimately connected to the directedness distinctive of (at least some of) our linguistic utterances (and psychological states). Why? Because language cannot be made sense of without appealing at the same time to the idea of a kind of contentfulness that is distinctive of at least some of our linguistic utterances (and psychological states). We can then pretheoretically specify the content of a linguistic utterance by saying, for instance, that it is of or about something, represents something in the world, has a peculiar kind of link to the world, and so on. Therefore, language cannot be made sense of without appeal to intentionality. There cannot be language in the absence of intentionality. Accordingly, understanding hinges on intentionality.

But what exactly is intentionality? Brentano defined intentionality as "reference to a content, a direction upon an object" (1970). Similarly, Searle emphasizes that: "… if a state S is intentional then there must be an answer to such questions as: What is S about? What is S of? What is it an S that?" (1983). More recently, Bender and Koller take (linguistic) meaning as "…the relation between a linguistic form and communicative intent" (2020, p. 5185). That is, according to them, the meaning of a word is its communicative intent.[6] In language processing, we are guided by the pursuit of certain intentions, which we express in linguistic expressions. Understanding meaning is then the ability to map an expression onto its intent.

ing with the possession of an internalized grammar encoded in the mind/brain (1965; 1986); and Searle insists that genuine understanding requires intentional mental states with intrinsic meaning, as illustrated in his well-known Chinese Room argument (1980).

[5] I am grateful to an anonymous referee for their insightful comments on this matter.

[6] This view is somehow reminiscent of Grice's (1957), which understands linguistic meaning in terms of the speaker's meaning, and the speaker's meaning in terms of the intention of a speaker to induce a belief in the audience by an utterance accompanied by the audience's recognition that the utterance was produced with that very intention.

In a nutshell, discursive intentionality is what is exhibited by language users, as concept users, who can say and understand that things are thus and so**,** who can make and understand claims or judgments that are about something, about objects in the world, and so forth.[7]

But how does intentionality arise? How does this kind of relationship between a linguistic item (i.e., a sign) and things outside of it come about? By virtue of what, or by what means, does its rapport with its environment arise? This is hardly a causal connection.[8] At the same time, it is hard to appeal to some specific subjective experiences. It is hard to maintain that specific subjective experiences are necessary for linguistic understanding and hence for intentionality. To understand language does not require any subtle ingredient such as *qualia* (or mysterious "inner light").

If intentionality, and accordingly the meaning of any concept (even concepts close to sensory experience such as "pain" and "red") neither derive from a causal relationship to the environment nor from some sensory experience resulting from interactions with the environment, then how does meaning establish its link to the world? Sellars offers the following answer:

"…it is by virtue of the fact that we draw inferences that meaning gets its connection with the world" (Sellars, 1962, p. 246).

Thereby, the meaning of an expression is not something that lies behind the expression itself (or to which the expression refers). Nor does it exist without its relations to other expressions, whose meanings are themselves determined only by their relations to other expressions. Rather, conceptual content consists of inferences to and from other concepts:

"It is only because the expressions in terms of which we describe objects… locate these objects in a space of implications, that they describe at all, rather than merely label" (Sellars, 1958, pp. 306–307).[9]

In particular, something qualifies as conceptual content just insofar as it stands in relations of *material consequence* and *incompatibility* with other such contents. Material consequence and incompatibility relations, by contrast to formal logical ones, thus articulate the contents of non-logical concepts. In turn, propositional contents (which are a principal species of conceptual content) are what can figure as both premise and conclusion in *material inferences*.[10] These complex relations of the

[7] It is plausible that the normal subject does not have the term "intentionality," but she has a battery of other terms— most notably, "about" and "directed"—wiplauth which she can express her concept(ion) of intentionality. Without such a conception, she would be unable, for example, to understand simple traffic signs: 'Traffic signs are about something: a sign on Interstate 95 that says "New York, next exit" is *about* the sorts of actions that have to be taken in order to reach New York from where the reader is at' (Kriegel, 2011, p. 55).

[8] Note that the question is not a causal one (i.e., how linguistic understanding is causally produced), but a conceptual one: what are the criteria we applyto determine whether a being understands?

[9] It is worth noting that already rationalists, like Spinoza and Leibniz had somehow developed an account of what it is for one thing to represent another in terms of the inferential significance of the representing.

[10] Sellars (1953) characterizes material inferences by two key features, which distinguish them and make them essential in understanding non-logical reasoning. Firstly, material inferences are *non-enthymemic*, meaning they are complete inferences whose validity does not rely on adding hidden or unstated premises. Secondly, material inferences are *independent of logical vocabulary* for their validity. This independence captures the way their validity stems from the content of the concepts themselves rather than from their

individual parts of the corpus to all other parts of the corpus then ground the meaning of the isolated parts in a given context.

This view clearly stems from the work of Wittgenstein, according to which the meaning of an expression is its role within our *language games*, "its use in the language" (1953, §43). But what exactly is Wittgenstein getting at? He likens the meaning of an expression to its role within our language games:

"Compare the meaning of a word with the 'function' of an official. And 'different meanings' with 'different functions'" (1969, §69).

Inferentialists add weight to this idea by suggesting that conceptual contents are articulated by the *roles* they play in material consequence and incompatibility relations (Brandom, 1994, 2000, 2007; Peregrin, 2014, 2018). Expressions acquire their content by taking part in the complex network of such inferential relationships, which thus makes it possible for discursive practice—namely, the practice of explicitly claiming and judging that things are thus and so. Such relationships equip the expressions with roles, which may be considered as their meanings:

"Any web of relationships among linguistic items equips the items with more or less complex roles, which may be considered as their 'meanings'" (Peregrin, 2021, p. 314).

The meaning of an expression is thus its role (i.e., its place) in the complexity of relations to other expressions. To grasp this meaning, hence to understand the expression, is nothing but to grasp the inferential role the expression plays.

As a result, being considered a competent language user hinges on the ability to navigate inference-making and drawing conclusions from given premises. In other words, to count as a language user, one must *know how* to make inferences and so draw conclusions from his premises. That is, one must know how to distinguish what constitutes evidence supporting a claim, as well as what else is ruled out as incompatible with it. In this way, mastery of a language ultimately boils down to a practical ability, to a skill: understanding an expression is no more than knowing how to skillfully employ it.

Do Large Language Models (LLMs), such as GPTs and ChatGPTs, possess the aforementioned know-how? Undoubtedly, it can be argued that they engage in inference-making, discern implications or contradictions between sentences, deduce conclusions from given premises, map out reachable steps and excluded derivations, and so on. Hence, there is a sense in which they could individuate information by its place in an inferential network.[11] However, it remains an open question whether this

formal structure. In addition to the two features proposed by Sellars, Brandom (1994, p. 2000) introduces a third crucial one. According to him, material inferences are *non-monotonic*, meaning that their conclusions are defeasible. That is, an additional clause may alter the truth of the conclusion. In an interview published in *Disputatio*, Brandom explicitly discusses the non-monotonicity of material inferences as a potential challenge for artificial agents (Frápolli & Wishin, 2018).

[11] Brandom distinguishes three forms of inferentialism: "weak," "strong," and "hyper-inferentialism" (1994, pp. 131–2; 2007). "Weak inferentialism" holds that the inferential connections among sentences are at least *necessary* for those sentences to have the contents they do; in effect, it claims that inferential relations are at least as primitive as referential relations. "Strong inferentialism" and "hyper-inferentialism" go further, claiming that inferential articulation is *sufficient* to determine content. Both treat inferential relations as conceptually more primitive than referential ones (i.e., the latter can be understood in terms of the former). The key difference between the strong and hyper-inferentialist theses lies in scope: the

provides sufficient grounds for the conclusion that LLMs can be classified as genuine language users.

In order to answer this question, it is then crucial to recognize that roles are conferred by *rules*.[12] Indeed, Wittgenstein's insight into language games suggests that they essentially operate as rule-based systems. Within these systems, expressions fulfill their specific functions/roles dictated by the rules that govern their usage. This underscores that the significance of linguistic elements relies on the norms that regulate their application in judgment.

Brandom's (1994, 2000, 2007) "normative inferentialism"[13] builds on this idea, proposing that an expression's contentfulness consists of its use or occurrence being governed by *inferential rules*.[14] Meaningful items incorporate norms of inference (i.e., material consequence and incompatibility relations), which they are subject to. Inferential roles—hence meanings—are conferred by such norms to which speakers commit themselves and not merely by, say, their dispositions to make inferences. Grasping the meaning of a word, understanding it, is thus mastering this bundle of rules. Then, the relevant question to ask turns out to be: Do LLMs follow inferential rules? Or do they instead merely exhibit certain regularities in their behavior?

3 Traditional Computers and Black Boxes

Traditional computer systems rely on software states, which essentially encompass computational operations or processes. Those are programmed instructions explicitly stated—such as codes, algorithms, and software implementations. Can the practical capacity to follow rules be cashed out in terms of such computational operations?[15]

latter restricts itself to relations among sentences, whereas the former also takes into account the "inferential circumstances of appropriate application" and the "inferential consequences of application," thereby extending inferential relations beyond purely linguistic moves to include connections between sentences and states of affairs (e.g., via perception). Brandom himself endorses the strong version. Purely linguistic LLMs, relying exclusively on textual data, may be seen as aligning with hyper-inferentialism rather than strong inferentialism, whereas multimodal language models (MLLMs) could in principle approximate the latter more closely. Since this paper focuses on pure LLMs, it adopts the hyper-inferentialist framing, though the analysis may also extend to multimodal systems. I thank an anonymous reviewer for comments that helped clarify this point.

[12] The term "rules" is usually used to refer to what you get when you make *norms* explicit in the form of sentences, making them things one can say—things one can articulate and assert. To minimize confusion, I will use these terms interchangeably in the following discussion and will not emphasize this terminological distinction between norms as implicit guidelines and rules as their explicit, verbalized counterparts.

[13] The approach concerned is normative in that it holds that linguistic meaning is constituted by the material inferences one ought to or may (not) make. This contrasts with descriptive inferentialism, which considers meaning-constituting inferences as those that language speakers normally or typically make (or are disposed to make). For a naturalistic version of inferentialism, see notably Paul Boghossian (1993; 1994; 2003; 2012).

[14] See also (Peregrin, 2014) and Loeffler (2017, Ch. 4).

[15] For instance, *intellectualism* (a contemporary version of *Platonism*) sees every practice as underwritten by a rule or principle: something that is or could be made discursively explicit. The *computational theory of the mind* endorses this possibility of explicitly stating in rules all the implicit practical background skills (i.e., practices) necessary to institute those rules. This view is then shared by the program of *symbolic*

Software states can be understood as a device's *internal states*, that is, as a device's internal representations or interpretations. In other words, they are its "mental life." Can we then attribute to machines the ability to follow rules (such as the inferential rules governing language use) in light of such internal states?

Wittgenstein's analysis of rule-following nicely fits here. As he taught us, rule-following is not a matter of what is in the mind. We cannot, and need not, know which internal representation a pupil relies on when correctly continuing a series of numbers (Wittgenstein, 1953, §143–§185). He goes on to suggest that the rules governing human activity in general (and linguistic practice in particular) cannot be explicated by the Platonic tradition of reference to ineffable objects, nor by a subjective "interpretation" at the moment of each instantiation of the rule. Indeed, attempting to explain rule-governed activities by appealing to subjective interpretation at each instantiation of the rule inevitably results in infinite regress (Wittgenstein, 1953, §201). If the correct execution of an action each time requires consulting the guiding rule (i.e., the proposition expressing the rule), then the act itself of consulting that rule would in turn presuppose consulting the rule that guides the act of consulting the rule previously consulted, and so forth.[16] Thereby, rules neither determine actions nor determine or establish meaning. Rule-following (and accordingly language use) does not hinge on internal states, that is, it is not a matter of having inner representations of any kind.[17]

In a nutshell, according to the Wittgensteinian perspective advocated here, language use is a matter of following certain (social) rules governing the employment of expressions;[18] but rule-following is not a matter of what is in the mind:

"If God had looked into our minds he would not have been able to see there whom we were speaking of" (Wittgenstein, 1953, p. 217).

Consequently, rule-following cannot be accounted for in terms of mere computation either.[19] If a computer's output relies on its internal state for meaning, that internal state cannot possess meaning without additional states, thus initiating an infinite regress, wherein each state requires a further state for validation. To qualify as language users, human beings must exhibit the ability to follow the rules that govern language use regardless of their internal mental states. Likewise, to count as genuine language users, computers must exhibit the ability to follow the rules that govern

artificial intelligence, which endorses the possibility of explicitly codifying in programmable rules all the implicit practical background skills necessary to institute those very rules.

[16] In the same way, if a device's output inherited its meaning from the device's internal state, then the internal state could not itself consist in representation (and hence being meaningful) without presupposing further representation; we would thus again be faced with an infinite regress.

[17] Wittgenstein (1953, §158) also explicitly ponders the idea of assessing whether somebody has mastered reading by further investigating neural processes in the brain. Wittgenstein rejects the idea by emphasizing our knowledge of such matters, that is, the way we commonly assess whether somebody knows how to read. Surely, these would be the criteria to identify any neural processes jointly necessary and sufficient for mastery of reading, and not vice versa.

[18] Recall that, following inferentialist semantics, such rules are inferential rules.

[19] This perspective notably contrasts with Fodor's computational representational theory mind (1975), which is compatible with the possibility that computers use language in a manner that somehow engages an internal mental life.

language usage regardless of their internal states and mechanisms—such as code implementations, algorithms, or computational processes of any kind.

Nevertheless, contrary to classical von Neumann architectures (i.e., the traditional computer systems), the neural network form has turned out to be able to address topics and solve problems without requiring explicit programming and formulation of rules. This might give rise to epistemic challenges because it becomes difficult to predict and explain the behavior of autonomous devices or trace causal connections to individuals controlling the outcomes. That is to say, where behavioral output exceeds the purposes of initial programming, our efforts to decipher causes might not succeed. As a result, some neural networks may well remain "black boxes."

The same happens in sufficiently large models, where it is possible to confirm the existence of features for which the models have not been directly programmed. In this case, we are then unable to interpret the inner workings of LLMs, and "any attempt at a precise explanation of an LLM's behavior is doomed to be too complex for any human to understand" (Bowman, 2023). This inherent *opacity* in how some LLMs operate is precisely why they too are often referred to as "black boxes." The complexity and lack of transparency in their underlying processes make it difficult, even for those who design them, to fully understand or predict how they generate specific outputs.

Does this mean that we should engage with them through what Dennett (1987) calls "the intentional stance"?[20] If so, should we then conclude that LLMs are really following rules irrespective of their internal states and mechanisms?

4 The normativity of Norms

Rules cannot but be a *social matter*. That is, rules must be understood as a public affair, implicit in the social practice. But with social practice comes normativity. This means that norms are instituted socially through "reciprocal recognition": by attributing *normative statuses* to one another, exercising authority and taking responsibility. Brandom (2019, Ch. 2) endorses this Hegelian view, emphasizing that normative statuses exist only insofar as individuals adopt normative attitudes toward themselves and others, thereby situating and bounding up themselves within a web, network, and constellation of rights and obligation. Norms thus arise only when such attitudes are adopted and enacted. Unlike a mere habit, which is simply a regularity in behavior or a causal disposition, a norm-governed stance is evaluated in terms of correctness within a social practice. Accordingly, to speak of "roles" is to speak of the canons of authorization and prohibition sustained within the social community.

[20] Dennett's intentional stance is the level of abstraction in which we make sense of the behavior of an entity in terms of mental properties, such as beliefs and desires. We adopt the intentional stance when we attribute such mental properties to the entity in question, and use these attributions to make sense of its actions. This operates on a higher level of abstraction compared to other stances, such as the "physical stance" (which explains behavior based on the physical composition and laws governing an entity) or the "design stance" (which focuses on the intended function of the entity's design). The intentional stance is particularly useful in contexts where the behavior of an entity is complex and where attributing mental states can provide a clearer and more effective explanation than other approaches.

However, to say that social norms are instituted when we take or treat each other as responsible or authoritative is to say that norms are instituted when we take or treat performances as appropriate or inappropriate, correct or incorrect. Practically sanctioning or rewarding performances[21] is one way of treating them as correct or incorrect. By setting the standard for assessments of correctness, the norms one applies determine what one has made oneself responsible for in applying them.

Ultimately, it is important to note that, on this view, norms are not prior to practice; rather, it is the practice itself that grounds the normativity of norms. Consequently, norms cannot be accessed from outside the practices that institute them: they can only be grasped from within, through participation in the relevant social practices, which determine their content. Accessing norms requires attributing normative statuses to oneself and others, which involves adopting corresponding normative attitudes toward oneself and others. In doing so, one contributes to shape the norms themselves. Thus, participation both engages with and helps constitute the norms. Normative statuses, and hence the normativity of norms, arise only within social practice, and there is no perspective "from nowhere" from which to access them.

Brandom's inferentialism also rejects purely descriptive or causal accounts of *discursive norms*, such as naturalistic or empiricist approaches. Drawing on a view taken from Hegel by the American pragmatists (C. S. Peirce, William James, and John Dewey) and later on by Wittgenstein, Brandom holds that also norms governing linguistic performances are to be understood as implicit in social practices. They are products of our practical engagement in discursive activities, rather than of habits or dispositions: they are in force because we treat them as binding in reasoning and social interaction. As such, they are instituted only when individuals attribute normative statuses to themselves and others by adopting corresponding normative attitudes toward themselves and others.

Brandom (1994, Ch. 1) elaborates this view in terms of the normative attitudes of commitment, entitlement, and precluded entitlement to commitment. Making a claim involves undertaking a commitment to "demonstrate one's entitlement to the claim, if that entitlement is brought into question," where being entitled to a claim is being "entitled to make it" (Brandom, 1994, pp. 171–172). This roughly means that one is making a claim insofar as "He is making a move in a practice of giving and asking for reasons, in which one move has normative consequences for what others are obligatory, permitted, or prohibited" (Brandom, 2014, p. 354). Normativity is thus presupposed for the very formulation of claims. Claims imply a normative space where they can be criticized and justified—the normative web of giving and asking for reasons—which is constituted exactly by those very social-normative relations. Accordingly, "To understand an assertional speech act is to know how to keep score on the commitments the speaker has undertaken by performing that act" (Brandom, 2014, p. 357). That is, understanding a claim is taking up a stance in a network of related possible claims, which stand to one another in rational relations of material consequence and incompatibility. It follows that talking about "linguistic roles"—roles in practices of giving and asking for reasons—is talking about the canons of commitment and entitlement sustained within the speech community.

[21] For the sake of argument, I take sanction and reward to be any positive and negative corrective behavior.

We take each other to express meanings when hold each other committed and entitled. The relevant normative practice is therefore conceived as constitutive of conceptual and propositional contents, and "basic in the order of semantic explanation" (Brandom, 1994, p. 496; 1983, pp. 640–644). Meaningfulness is built into such practices, within which alone meanings get expressed. It emerges from practices that has the right normative and social structure (Brandom, 1994, Ch. 3), wherein individuals adopt instituting normative attitudes towards one another and, in turn, attribute normative statuses to each other. This view clearly stems from the work of Wittgenstein, according to which the capacity to make propositionally explicit claims and to have conceptually contentful thoughts is intelligible only in the context of implicitly normative social linguistic practices (1953, §199).

However, again, to say that discursive norms are instituted when we take or treat each other as committed and entitled is to say that social norms are instituted when we take or treat performances as appropriate or inappropriate, correct or incorrect. And, again, practically sanctioning or rewarding performances is one way of treating them as correct or incorrect. This is, norms that confer inferential roles (which constitute meaning) are set up by what is fixed as correct or incorrect inferences in social interactions: "…inferences are correct in the sense that they are accepted in the practice of a community," where communal acceptance is a matter of "actual practical attitudes" of assessment (Brandom, 1994, p. 137).

To sum up, we have established that language users—those who can say that things are thus and so—exhibit some sort of *intentionality*. This "discursive intentionality" is to be found in the *role* linguistic items play in *material consequence* and *incompatibility* relations. These roles are then conferred on linguistic items by rules of inference. However, *Inferential norms* are set up by what is fixed as correct and incorrect inferences in social interactions. Norms governing linguistic performances are therefore instituted when individuals adopt *normative attitudes* towards themselves and others and attribute *normative statuses* to each other—such as commitments, entitlements, and precluded entitlements to commitments. As a result, the relevant normative practice (of authority and responsibility) turns out to be constitutive of conceptual and propositional contents. Accordingly, *discursive intentionality* emerges from practices that have the right normative and social structure (Brandom, 1994, Ch. 3).[22]

5 LLMs and Language Acquisition

Pursuing this order of explanation leads to the conclusion that exercising authority and taking responsibility is exactly what one must do in order to count as grasping and understanding what, in social practice, shows up as conceptual and propositional contents, and thus to count as a competent language user. In particular, "To be a com-

[22] In turn, insofar as understanding hinges on intentionality, conceiving it as participation in socially instituted, norm-governed practices of inference, justification, and correction allows this Brandomian inferentialist framework to retain the notion of understanding without treating it as an inner state or hidden mental episode, thereby avoiding the pitfalls of internalist or mentalistic accounts.

petent speaker (competent concept user) is to engage in practices of giving and asking for reasons, that is, making inferentially articulated assertions and justifying them. To do so, one must attribute and acknowledge commitments and entitlements, and practically keep track of their inferential relations" (Brandom, 2014, pp. 359–360).

This ultimately implies that, to be considered genuine language users, LLMs should first be given some form of "personality" in their own right, where persons are agents individuated by their position in the normative relations into which they enter through adopting normative attitudes to one another: namely, the normative web of rights and obligations.[23] But can LLMs be held accountable in a normative sense for their output? Can we adopt *normative attitudes* towards them and attribute *normative statuses* to them?

What is needed for us to attribute them responsibility is for it to be in principle possible for us to assess their behavior (i.e., their attempts at rule-following) as right or wrong. In turn, what is needed for LLMs to count as responsible is for it to be in principle possible for them to improve in response to (output) failures. Hence, they should be able to learn from their mistakes. But this means that it must be in principle possible for us to teach them by communicating wrongdoing—or better, train them by sanctioning their wrongdoing, even punishing them.[24] Like our accountability practices toward fellow humans, we should be able to hold AI to account by rewarding or imposing sanctions.

Indeed, it is only by reciprocating normative attitudes that learning can take place. Teacher-student relation obtains when the student does something wrong (i.e., she does not follow the teacher's instructions), and the teacher asserts power in response to it: "If a child does not respond to the suggestive gesture, it is separated from the others and treated as a lunatic" (Wittgenstein, 1958, §30). Instead of the tool/user model, we have here a complex relationship between the teacher and the student, in which both share responsibility and authority. We praise or blame the skill and sacrifice of the teacher and laud or sanction the actions of the student as well. But do LLMs learn in this way? Do we train them *practically*?

Language is always learned in this way: one learns how to follow the rules of language in practice, through trial and error. Correspondingly, language is always taught by treating performances as correct or incorrect through practically sanctioning or rewarding them. The social practices that allow the infant to acquire their first language are thus of the same kind as the social practices that imbue strings of letters

[23] Of course, the idea of AI systems being considered persons largely depends on how we define the concept of a "person." For instance, it could be argued that unless and until AI systems function as agents in the legal sense, then, they cannot acquire any form of person-hood. If so, we can simply rely on the concept of agency, provided it refers to the ability to act within the scope of their actual authority. In any case, the key point here is that if AI systems begin to operate as more than tools or instruments—acting independently of the user's control or direction—this would imply an agency relationship that presumes some level of responsibility on the part of the AI system itself.

[24] Social sanction and reward, praise and blame, enable us to influence and regulate one another's behavior: they get us to act, or get us to refrain from acting, so as to adjust our various actions in relation to one another. For example, we may get upset and blame our roommate for eating our leftovers without asking. Likewise, we jail tax evaders and drug dealers, and so on. In both cases, we aim to deter their behavior and prevent similar occurrences from occurring in the future. The same should be possible with LLMs. For a similar view regarding AIs in general, see Allen and Wallach (2009) and Hellström (2013).

with meaning (and reference). Accordingly, the claim that LLMs of generative AI qualify as language users cannot rest solely on the superficial indistinguishability of their outputs from human texts; it must also consider the extent to which the mechanisms underlying their "language acquisition" resemble the processes by which we acquire language. LLMs can be expected to follow rules that govern the use of linguistic locutions (i.e., rules of inferences), and so mastering the use of words and grasping and understanding their meaning, only if they have learned to do so in practice, through social sanction and reward in response to trial and error. But, even if LLMs can be practically trained in this way, are the inferential rules the ones they learn to follow?

The current way in which language models are trained involves detecting statistical patterns. What they learn is to capture statistical distributions of sequences of words in a collection of texts, and to generate new texts drawn from the same distribution. In a nutshell, LLMs are trained to predict statistically likely sequences of words from massive datasets of text from the internet. They can then be fine-tuned to improve specific aspects of their performance (e.g., correctness).[25] This can be achieved through fine-tuning strategies such as Reinforcement Learning with Human Feedback (RLHF), supervised fine-tuning (SFT), and Quantized Low-Rank Adapters (Q-LoRA) method. For instance, in RLHF, crowdworkers' evaluations are used to train a reward model, which provides the reward function for fine-tuning the system itself via reinforcement learning so that outputs aligned with human preferences are rewarded and dispreferred ones penalized. To some extent, these processes can indeed be understood as practices of reward and sanction in response to trial and error**,** functioning as positive and negative corrective behaviors and thus as mechanisms for holding LLMs accountable. Nevertheless, the outcome remains fundamentally a model that captures the statistical distribution of word sequences in human language, reflecting patterns of token co-occurrence without genuinely engaging with the norms of inference that underlie language use and discursive practice.

If so, they turn out to be just sophisticated implementers of "surface statistics" (Li, 2023); mere repeaters guided only by probability. They merely assemble words into sentences based on statistical and probabilistic information about how to combine them, and thus they behave like a kind of "stochastic parrots" (Bender et al., 2021). That is, the prevailing view is that language models only operate stochastically with linguistic expressions, generating text according to pre-set stochastic patterns stored. We can then compare the performance of LLMs to, for example, the jargon-spouting students who try to imitate their professors but basically do not know what they are talking about (Browning & Lecun, 2022). As a result, the ability of an LLM to predict word sequences, which reflects a sophisticated pattern-matching process (e.g., essentially a form of syntactic manipulation), does not necessarily encompass the ability to follow the deeper, complex, inferential norms that govern discursive practice.

However, when I say that an LLM can only be "trained" through our interventions, I do not mean to exclude self-correction from the picture. LLMs can indeed exhibit limited forms of what looks like self-correction—for example, by revising an output within a conversational exchange. Yet such revisions do not reflect an autonomous

[25] Examples of fine-tuned AI sistems include LaMDA and Sparrow.

grasp of correctness. They amount only to the redeployment of statistical patterns internalized during *pre-training*, not to norm-sensitive responsiveness. Pre-trained LLMs (e.g., BERT, GPT-3, PaLM, Falcon) undergo an initial stage of training on large corpora via self-supervised learning (i.e., unsupervised learning) to predict the next word (token) in a sequence, acquiring statistical patterns of language. Their outputs reflect the statistical models they embody, and once this stage is complete, their weights—and hence their behavioral dispositions—are fixed: they cannot modify their underlying parameters in response to interaction. Apparent flexibility in such systems derives from prompting, where examples of desired outputs serve as contextual cues that shift probability distributions.

By contrast, fine-tuned models (e.g., LaMDA or Sparrow) are subject to additional alignment procedures**,** such as reinforcement learning from human feedback (RLHF), reward models, and supervised fine-tuning on diverse datasets. At this stage, genuine improvement over time is possible, but any change in the model's parameters depends entirely on these external evaluative signals. Without such interventions, the model cannot alter its behavior in a way that counts as correction or improvement, precisely because what qualifies as "error" can never be instituted autonomously by the model itself.

In Wittgenstein's terms, the very possibility of being right or wrong in applying a rule presupposes the availability of external correction: there is no such thing as following a rule *privately* (1953, §§198–202). From this perspective, even apparently self-corrective mechanisms ultimately rely on evaluative signals and reward structures derived from explicit human supervision. In this sense, LLMs' capacity for improvement remains rooted in and parasitic on our capacity to train and teach them, since the very normative standards of correctness against which improvement is measured are externally fixed and imposed rather than internally generated.[26]

Even so, though, LLM-driven chatbots remain trained to optimize the likelihood of word sequences in large corpora. Fine-tuning methods such as RLHF, SFT, or Q-LoRA can improve alignment with human expectations, but the responsiveness achieved through these methods remains tied to the statistical calibration of the model's weights rather than to the inferential norms that structure language use. At best, this results in a pattern-based conformity to discursive practice, which nonetheless falls short of genuine participation in the norm-governed space of reasons.

Recent empirical findings provide strong support for this diagnosis. Studies consistently show that LLMs fail at tasks requiring genuine logical reasoning, displaying significant limitations in reasoning even when their outputs appear inferentially coherent. Accordingly, some authors have argued that LLMs do not genuinely learn to reason, but instead fit statistical regularities that amount to superficial pattern recognition without internalizing logical structure.[27]

[26] Many thanks to an anonymous referee for helpful comments on this matter.

[27] For instance, gains from fine-tuning and Chain-of-Thought reasoning are largely task-specific and may not reflect faithful reasoning (Lobo et al., 2024), performance often reflects token-level biases rather than genuine inference (Jiang et al., 2024), and models fail on tasks requiring true logical reasoning rather than superficial pattern recognition such as 3-SAT (Hazra et al., 2024). Recent evaluations of mathematical reasoning show that even state-of-the-art models collapse under minimal problem variations, indicating reliance on training data patterns rather than genuine inference (Mirzadeh et al., 2024). Empirical stud-

The question motivating this paper, however, is not merely whether these limitations exist, but why they should compel us to conclude that LLMs do not engage in the reasoning processes involved in natural language use, despite their impressive performance on many linguistic tasks. Undoubtedly, future improvements are possible (e.g., through larger corpora, higher-dimensional models, more efficient implementations, or greater computational power). Yet the central issue remains: even in light of such performance gains, can reasoning in natural language be reduced to operations on statistical patterns (and semantics equate to distribution)? Or, at best, do LLMs' statistical operations merely simulate our reasoning practice (just as distribution only provides a useful but imperfect reflection of semantics)? The argument put forward here to support this latter conclusion is precisely that what ultimately distinguishes genuine reasoning from mere simulation is the presence of participation in normative inferential structures.

This distinction, again, can be illuminated by Wittgenstein's analysis, notably through the difference he emphasized between genuinely *following a rule* and merely acting in accordance with a rule-like pattern (1953, §§198–202). To follow a rule in the strict sense requires grasping the standards that determine correct and incorrect applications; it is essentially a normative activity. By contrast, one may simply *adhere to* (or *act in accordance with*) a rule in the sense of producing behavior that conforms to a pattern, without any sensitivity to such normative standards. Brandom also stresses this point in his critique of what he calls "regularism": the attempt to explain rule-following entirely in terms of behavioral regularities (1994, Ch. 1). Regularism collapses the distinction between being bound by a rule and merely displaying conformity to it.[28] This distinction, though, makes it possible to differentiate between genuine inferential capacity and mere regularistic adherence. In this light, LLMs at most can generate outputs that conform to pattern-based inferential relations, but lack any grasp or responsiveness to the normative significance of those relations. While their outputs often appear inferentially structured, this appearance results from regularistic adherence rather than genuine normative participation in the logical space of reasons.[29]

ies also highlight persistent weaknesses in consistency (Arkoudas, 2023; Saparov & He, 2023), planning (Valmeekam et al., 2022), and self-evaluation (Stechly et al., 2023). Theoretical analyses reinforce these concerns, showing that transformers cannot in principle solve certain reasoning problems such as Derivability, 2-SAT, and Circuit Evaluation (Peng et al., 2024) and performances deteriorate in "hard" regions where superficial statistical features are absent (Hazra et al., 2025). While some research points to emergent abilities in zero-shot tasks, which improve with model scale (Kojima et al., 2022; Srivastava et al., 2023; Wei et al., 2022), many conclude that LLMs exhibit a "Clever Hans" effect—mimicking the style of reasoning without internalizing its logic (Bachmann & Nagarajan, 2024; Zhang et al., 2023). I am grateful to an anonymous reviewer for prompting me to include these remarks.

[28] Brandom contrasts regularism with what he calls "regulism," which makes rule-following depend on explicitly formulating and consulting rules in each application. This reflects one aspect of Wittgenstein's rule-following considerations (1953, §§198–202), where he problematizes the idea that every act of rule-following must be guided by an explicit interpretation—an idea that generates an infinite regress of rules for interpreting rules (see §5 below). Brandom's inferentialist alternative avoids both extremes by treating rule-following as participation in practices structured by implicit normative statuses—practices of giving and asking for reasons that determine what counts as correct or incorrect (1994, Ch. 1).

[29] I thank an anonymous reviewer for helping to emphasize this point.

6 Proxy Assertion, Proto-Assertion, and Make-Believe

To recall, the main point advanced in this paper is that the behavioral dispositions of current LLM-driven chatbots do not provide a prima facie case for considering them proper asserters. To be sure, chatbots often produce what appear to be defenses and explanations for their putative assertions: they can respond to reasonable challenges, offer reasons for their claims, avoid blatant contradictions, and even retract statements when shown to be unsupported. In this respect, their behavior conforms to the profile of asserters to a significant degree. Nonetheless, according to the view defended here, assertion is a norm-governed practice. A putative speaker is capable of genuine assertion only if they stand in the appropriate relation to the constitutive norms of assertion. The objection, then, is that engaging in the practice of assertion requires the capacity to follow the relevant inferential norms, whereas LLM-driven chatbots lack such an ability. Their behavior may be largely consistent with those norms in the sense of exhibiting rule-like patterns, but they do not follow the norms themselves. Accordingly, what chatbots generate are not proper assertions, but sophisticated simulations of them. This reinforces the central claim of the paper: LLMs, despite their impressive linguistic performances, do not genuinely participate in the norm-governed space of reasons.

Some of the conclusions reached here overlap with recent proposals in the literature, but there are also important differences. Butlin and Viebahn (2023), for instance, argue that for an entity to be capable of assertion, it must satisfy two requirements: it must produce outputs with descriptive functions, and it must be capable of being sanctioned by the agents with which it interacts. According to them, fine-tuned language models may satisfy the first requirement but fail the second, and thus their outputs cannot count as assertions. While the present paper arrives at a similar conclusion, the arguments put forward differ in important respects. On the account defended here, such systems can be seen as meeting both requirements, yet this still does not suffice to qualify them as genuine asserters.

Their first point is that assertion is a species within the genus of descriptive representation. Descriptive representations are those that say that things are a certain way, and are therefore assessable as true or false (or accurate or inaccurate). Accordingly, for an output to have a descriptive function is for it to convey information to a consumer system so as to affect the consumer's behavior. Butlin and Viebahn maintain that LLMs modified through fine-tuning (e.g., LaMDA or Sparrow) are designed to produce correct linguistic outputs in response to a very wide range of inputs, which suggests that many of their outputs do indeed have the kind of function characteristic of descriptive representation. Nevertheless, they contend that, while a descriptive function is a necessary condition for assertion, it is not sufficient. There are many cases in which a system produces outputs with descriptive functions that nonetheless fall short of assertion. The case of the thermometer illustrates this point: although its outputs have descriptive functions, it would be mistaken to say that thermometers make assertions about the temperature, since stretching the notion of assertion to cover cases like this would render it too broad to be explanatorily useful. As a result, judgments about whether an AI system is capable of assertion cannot rely solely on its outputs; it is also essential to examine its architecture and training processes. On

this point, the present view aligns with theirs: just as one cannot determine whether an AI system is sentient merely on the basis of its outputs, we likewise cannot determine whether it is sapient on that basis alone.[30]

Their key idea, though, is that assertion should be understood as a norm-governed social practice and, as such, entails sanctionability: a system can only properly assert if it can be sanctioned for failing to meet the standards of this practice. Since, according to them, LLMs cannot be sanctioned for outputs that violate these standards, they cannot be regarded as asserters. I share this view too: assertion is a norm-governed social practice, and part of what it is to be subject to a norm is to be sanctionable upon violating it. I also concur with them that sanctions require the agent to be apt to change their behavior in response, thereby improving their performance (as is evident in the case of simple thermometers). However, they regard this condition as still insufficient. They claim that even a more sophisticated thermometer—capable of recalibrating in response to feedback—does not plausibly qualify as sanctionable (and thus apt to produce assertions). For sanctions to count as such, they must be experienced as negative or undesirable by the agent, and there is no clear sense in which systems of this kind possess interests of their own: nothing can be good or bad for them.

Within the framework outlined here, instead, this further requirement is not necessary for an entity to qualify as a target of sanction. If the more sophisticated thermometer is capable of modifying its behavior in response to our interventions— so that it can be influenced and trained by us, thereby learning from our guidance —it can be said to meet the criteria required for sanctioning.[31] Nevertheless, this does not mean that the sophisticated thermometer can be regarded as following the inferential norms of discursive practice and thus as producing genuine assertions. Similarly, insofar as LLMs can be fine-tuned, they can be considered sanctionable even without experiencing such interventions as bad or undesirable. But still, they cannot be said to be subject to inferential norms, undertake discursive commitments, or act for reasons, since the norms they follow are statistical norms that regulate the probabilistic calculation of word-level co-occurrence patterns in large corpora. While this allows LLMs to mimic language use by approximating regularities in linguistic behavior, they remain insensitive to the inferential norms that govern it, are not subject to them, and are hence not genuine asserters.[32]

[30] The view under consideration corresponds to behaviorist accounts in the philosophy of mind, which sought to analyze mental states exclusively in terms of observable behavior (Ryle 1949; Skinner 1953). According to such accounts, the manifestation of appropriate behavioral outputs could, in principle, suffice to warrant ascriptions of mentality. By contrast, the present view contends that behavioral outputs alone cannot ground attributions of sentience or sapience, as these require participation in normative practices.

[31] Crucially, I hypothesize that this occurs not as a result of explicit design or programming (see §3).

[32] This raises the question of whether these two elements in the account of sanctioning —the capacity to change behavior in response to reactions, and the element of undesirability—can come apart, and whether this might also occur in the case of humans. Addressing this question lies beyond the scope of this paper. However, a possible line of response consistent with the view advanced here might be to treat them as two sides of the same coin. An agent's having interests of their own manifests itself precisely in social practice, in their responsive behavior within normatively regulated activities (such as learning and training processes). In such contexts, where Dennett's "physical stance" and "design stance" prove ill-suited to explain their behavior, appeal to the "intentional stance" may be appropriate (see §3). Does it follow that

However, some authors have argued that simply rejecting the thesis that LLM-driven chatbots can assert, while acknowledging that their outputs strongly resemble assertions, is still unsatisfactory. As Williams and Bayne (2024, p. 11) note, any adequate account should not merely deny that chatbots assert, but should also explain why their outputs appear so much like genuine assertions, even if they are not. Several responses to this dilemma have been proposed in the literature. One influential strategy is to treat machine assertion as a form of *proxy assertion*. In a proxy speech act, one speaker illocutes on behalf of a distinct principal, and it does not follow that the proxy thereby performs a speech act on its own behalf. Nickel (2013) develops this view, suggesting that machines may qualify as "speech actants to a substantial degree" (p. 495), while maintaining that "ultimate responsibility for artificial speech does not lie with machines, but either with persons or companies, or with nobody at all" (p. 500). Yet the proxy assertion strategy faces significant difficulties. Arguably, one of the most pressing challenges lies in the fact that, in standard cases of proxy assertion, there is a clear principal—the speaker whose assertion is voiced through the proxy. By contrast, in typical interactions with LLM-driven chatbots, no such principal can be identified. Therefore, who ultimately bears responsibility for the speech acts performed by NLG systems—the engineers who designed them, the programmers who implemented them, the manufacturer, or some other party? Green and Michel (2022) discuss cases in which the principal is clearer, such as AI systems deployed by a police department, where the institution implicitly endorses the system's outputs (e.g., issuing fines).[33] As Williams and Bayne (2024, p. 13) observe, though, while such cases may avoid some of the concerns raised, they do so only by substantially narrowing the scope of proxy assertion, thereby limiting its usefulness and appeal as a general solution. The view advanced in this paper avoids these difficulties by simply maintaining that LLM-driven chatbots do not produce assertions at all, not even proxy assertions. Nevertheless, the dilemma raised by Williams and Bayne remains.

For their part, Williams and Bayne propose a different resolution to the dilemma, suggesting that chatbots are best understood as *proto-asserters*. On their account, the capacity for assertion should be conceived as a graded phenomenon, with current-generation chatbots occupying an intermediate stage. They describe proto-asserters as "located somewhere in the space between beings that are fully-fledged asserters and those that are non-asserters" (2024, p. 24). Crucially, they take "proto-assertion to be distinct from (full-blooded) assertion, standing to it roughly as toddling stands to walking" (p. 24). The notion of proto-assertion thereby marks the theoretically important stage through which children also pass on their way to becoming full participants in the socially and metacognitively complex practice of assertion. Like children, chatbots are neither complete non-asserters nor genuine asserters, but rather proto-asserters: they display some of the characteristic features of assertion (though only partially and to varying degrees). Among these features, Williams and Bayne (2024, p. 23) highlight the norms governing assertion. In line with the analysis pre-

AI systems have emotions or consciousness? This ultimately depends on how these concepts are defined, which warrants further investigation.

[33] A similar position is defended by Freiman and Miller (2020, p. 428).

sented here, the authors characterize these norms as inferential norms. Accordingly, drawing on Piantadosi and Hill (2022), they contend that LLM-driven chatbots may exhibit a certain sensitivity to the inferential roles these norms fulfill. Nevertheless, insofar as their grasp of these norms remains partial, they qualify as proto-asserters rather than fully competent asserters. This, in turn, implies that they may have acquired some degree of semantic understanding. Analogous to the developmental trajectory of children, chatbots appear capable of acquiring this capacity, albeit only to a limited extent.

Contrary to this view, the arguments put forward here aim to show that LLM-driven chatbots can be classified neither as asserters nor even as *proto-asserters*, for the norms they follow—if any—are not the norms that structure the material inferences constitutive of conceptual content (i.e., the norms governing assertional practice in Brandom's sense). Instead, they are norms of a different kind: namely, the statistical norms that regulate the probabilistic calculation of word-level co-occurrence patterns in large corpora. This reflects the way such systems are trained and fine-tuned, whether through Reinforcement Learning with Human Feedback (RLHF), supervised fine-tuning (SFT), or Quantized Low-Rank Adapters (Q-LoRA). While these training processes can make chatbots more regularly aligned with human communicative expectations, they do not render them participants in the norm-governed space of reasons. At most, what results is a kind of syntactic or formal competence: a surface-level conformity to language-use regularities that does not give rise to genuine engagement with the inferential norms underlying discursive practice, but merely simulates it.

Why so? In Wittgensteinian terms, there is no hidden fact of the matter that makes it the case that one is following a rule; rule-following shows itself in the practices through which it is acquired and sustained. Human speakers acquire the norms of assertion by being trained into practices of giving and asking for reasons. By contrast, LLM-driven chatbots are trained into a very different kind of practice: one governed by optimization over probabilistic word co-occurrence in large corpora. To the extent that they are "corrected" (through fine-tuning techniques such as RLHF, SFT, or Q-LoRA), their responsiveness is to statistical optimization norms rather than to the inferential norms governing language use.[34] But still, this does not appear to resolve Williams and Bayne's dilemma: how can any adequate account explain why LLM outputs so closely resemble genuine assertions, even if they are not?

[34] Would it be possible to incorporate inferential roles into computational models? Addressing this question lies beyond the scope of the present paper. However, Williams and Bayne's focus on children's participation in the practice of assertion and their acquisition of illocutionary capacities provides valuable insights that align closely with the approach adopted here. A promising first step may be to explore language acquisition corpora (particularly those reflecting child-directed speech) rather than corpora of grown-up communication. Although smaller and more costly to collect, these corpora offer high-quality data: children's speech often makes explicit what is otherwise implicit, revealing the conceptual and inferential structures that guide language use—structures that adults typically assume or express only indirectly. Investigating these corpora can therefore provide a richer foundation for modeling sentence-level semantics. The practical challenges are considerable, and models developed within this framework may not match the performance of current models in the short term. Nevertheless, pursuing this strategy is worthwhile, as it has the potential to yield models that better capture what natural language meaning does, or what humans do with their semantic knowledge.

One possible way forward, consistent with the conclusions reached here, is Mallory's (2023) account of chatbot speech as a form of make-believe—at least to the extent that simulations can themselves be understood as a type of make-believe.[35] They argue that our interaction with chatbots is a kind of prop-oriented make-believe. On this view, "the texts produced by artificial agents are best conceived as props in games of make-believe, in which those agents function as fictional characters" (p. 1089). Interacting with a chatbot such as ChatGPT is therefore not a matter of engaging with a genuine asserter, but of imaginatively participating in a practice akin to fiction: the bot's apparent linguistic behavior prescribes certain imaginings on the part of the user, who treats its outputs as if they were assertions made by an agent. In this sense, interpreters engage with chatbot outputs as props in a game of make-believe, produced by a fictional speaker. As a result, this account avoids the risk of taking our interaction with chatbots as unintelligible—either by dismissing chatbot outputs as wholly meaningless or by denying their linguistic character altogether. Nor does it imply that users are deluded in engaging with these systems. Rather, people do not literally believe they are conversing with agents, but instead imaginatively make-believe that they are. Accordingly, the practice of gaining knowledge from chatbot testimony can be understood as a specific instance of the more general phenomenon of gaining knowledge from fiction. This, in turn, reinforces the claim defended here that chatbot speech is best understood as a simulation of assertional practice, rather than as genuine participation in the logical space of reasons.

7 Conclusion

As Ryle (1949) pointed out, only abilities that become manifest as "acquired" count as know-how. For this, it is required that we learn those abilities (vs. innate abilities) to count as knowing how to exercise them (vs. mere habits). Intelligence manifests itself in learning.

In light of what is suggested in this paper, it could then, perhaps, be argued that an LLM "knows" which words typically follow others based on statistical patterns. But even if we allow this, what they master are totally different norms from the norms of inference which contentful items are subject to**.** The latter are material inferences (i.e., material consequence and incompatibility relations), which mostly have the feature of non-monotonicity. On the contrary, although it might argue that statistical inferences are non-monotonic,[36] and thus LLM reasoning (if any) may exhibit non-monotonicity, statistical inferences are typically not considered material inferences.

[35] For discussions of how simulations be understood as a form of make-believe, see Walton's approach to fiction (1990) for the foundational theory of prop-oriented make-believe, and Toon (2012), Levy (2015), and Friend (2019), among others, for applications of this framework to scientific models, including Frigg (2021) for an illustration of how computational systems can function as props in imaginative engagement.

[36] In probabilistic reasoning, a conclusion is typically regarded as warranted if its probability, given the premises, exceeds a certain threshold. This inference relation can be considered non-monotonic insofar as the addition of new premises may lead to a revision of one's probability assignments, with the result that a conclusion that previously exceeded the threshold now no longer exceeds it. For a detailed discussion on non-monotonic logic and statistical inference, see Kyburg (1990), and Kyburg & Teng (2001; 2006).

While statistical inferences might reveal patterns of use or associations, they do not typically involve reasoning about content in the way material inferences do. As a result, they merely map, to some extent, onto the reasoning processes involved in natural language use, without genuinely engaging with the material inferences that underwrite discourse practice.

Machine language performance should therefore still be regarded as a mere simulation of grasping and understanding conceptual and propositional content (if that makes any sense at all). But if we agree that grasping and understanding conceptual and propositional contents is what one needs to count as a genuine language user, LLMs are still quite far from that. Therefore, although the future holds promise for assessing AI as genuine language users, the current landscape still presents considerable challenges in realizing this vision.

Acknowledgements I wish to express my sincere appreciation to the anonymous referees for *Minds and Machines* for their careful reading and constructive feedback. I also owe a special debt of gratitude to the members of the Institute of Philosophy v.v.i., Slovak Academy of Sciences (Bratislava), and to the members of the research project 'Metaphysics of categories: modalities and cognitive factors' at the Department of Philosophy I, University of Granada, for the insightful comments on this paper.

Author Contributions The author conceived the study and wrote the manuscript.

Funding Open access funding provided by The Ministry of Education, Science, Research and Sport of the Slovak Republic in cooperation with Centre for Scientific and Technical Information of the Slovak Republic. This work was supported by the Slovak Research and Development Agency under Contract No. APVV-22-0323 (project “Philosophical and methodological challenges of intelligent technologies”).

Data Availability Not applicable.

Declarations

Conflict of interest The author declares no conflict of interest.

References

Allen, C., & Wallach, W. (2009). *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Arkoudas, K. (2023). GPT-4 can’t reason. *arXiv*. https://api.semanticscholar.org/CorpusID:260704128

Bachmann, G., & Nagarajan, V. (2024). The pitfalls of next-token prediction. *arXiv*. https://api.semanticscholar.org/CorpusID:268364153

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198).
Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). ACM.
Boghossian, P. A. (1993). Does an inferential role semantics rest upon a mistake? In E. Villanueva (Ed.), *Philosophical Issues* (Vol. 3, pp. 73–88). Ridgeview Press.
Boghossian, P. A. (1994). Inferential role semantics and the analytic/synthetic distinction. *Philosophical Studies, 73*, 109–122.
Boghossian, P. A. (2003). Blind reasoning. *The Aristotelian Society Supplementary, 77*, 225–248.
Boghossian, P. A. (2012). Inferentialism and the epistemology of logic: Reflections on Casalegno and Williamson. *Dialectica, 66*, 221–236.
Bowman, S. R. (2023). Eight things to know about large language models. *arXiv*. https://doi.org/10.48550/arXiv.2304.00612
Brandom, R. (1983). Asserting. *Noûs, 17*(4), 637–650.
Brandom, R. (1994). *Making it explicit: Reasoning, representing, and discursive commitment*. Harvard University Press.
Brandom, R. B. (2000). *Articulating reasons: An introduction to inferentialism*. Harvard University Press.
Brandom, R. (2007). Inferentialism and some of its challenges. *Philosophy and Phenomenological Research, 74*(3), 651–676.
Brandom, R. (2014). Intentionality and language. In N. J. Enfield, P. Kockelman, & J. Sidnell (Eds.), *The Cambridge handbook of linguistic anthropology* (pp. 347–363). Cambridge University Press.
Brandom, R. B. (2019). *A spirit of trust: A reading of Hegel's phenomenology*. Harvard University Press.
Brentano, F. (1970). Psychology from the empirical standpoint. In H. Morick (Ed.), *Introduction to the philosophy of mind: Readings from Descartes to Strawson* (pp. [page numbers if known]). Scott, Foresman.
Browning, J., & Lecun, Y. (2022). AI and the limits of language. *Noema*. https://www.noemamag.com/ai-and-the-limits-of-language
Butlin, P., & Viebahn, E. (2023). AI assertion. *Ergo: An Open Access Journal of Philosophy*.
Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.
Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. Praeger.
Dennett, D. C. (1987). *The intentional stance*. MIT Press.
Descartes, R. (1641/1984). *Meditations on first philosophy* (J. Cottingham, Trans.). Cambridge University Press.
Elkins, K., & Chun, J. (2020). Can GPT-3 pass a writer's turing test? *Journal of Cultural Analytics, 2371*, 4549.
Fodor, J. A. (1975). *The language of thought*. Harvard University Press.
Frápolli, M. J., & Wishin, K. (2018). From conceptual content in big apes and AI, to the classical principle of explosion, an interview with Robert Brandom. *Disputatio, 8*(9). https://disputatio.usal.es/vols/vol-8-no-9/frapolli-wischin-interview/
Freiman, O., & Miller, B. (2020). Can artificial entities assert? In S. Goldberg (Ed.), *Oxford handbook of assertion* (pp. 415–434). Oxford University Press.
Friend, S. (2019). The fictional character of scientific models. In A. Levy & P. Godfrey-Smith (Eds.), *The scientific imagination* (pp. 102–127). Oxford University Press.
Frigg, R. (2021). Scientific modelling and make-believe. In S. Sedivy (Ed.), *Art, representation, and make-believe: Essays on the philosophy of Kendall L. Walton* (pp. 367–383). Routledge.
Green, M., & Michel, J. G. (2022). What might machines mean? *Minds and Machines, 32*(3), 323–338. https://doi.org/10.1007/s11023-022-09589-8
Grice, H. P. (1957). Meaning. *Philosophical Review, 66*(3), 377–388.
Hazra, R., Sygkounas, A., Persson, A., Loutfi, A., & Zuidberg Dos Martires, P. (2025). REvolve: Reward evolution with large language models for autonomous driving. *arXiv*. https://arxiv.org/abs/2406.01309
Hazra, R., Venturato, G., Martires, P. Z., & De Raedt, L. (2024). Can large language models reason? A characterization via 3-SAT. *arXiv*. https://arxiv.org/abs/2408.07215
Hellström, T. (2013). On the moral responsibility of military robots. *Ethics and Information Technology, 15*(2), 99–107.
Jiang, B., Xie, Y., Hao, Z., Wang, X., Mallick, T., Su, W. J., Taylor, C. J., & Roth, D. (2024). A peek into token bias: Large language models are not yet genuine reasoners. In *Proceedings of EMNLP 2024*.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems* (Vol. 35, pp. 22199–22213). https://proceedings.neurips.cc/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf

Kriegel, U. (2011). *The sources of intentionality*. Oxford University Press.

Kyburg, H. E., Jr. (1990). Probabilistic inference and non-monotonic inference. In R. D. Shachter, T. S. Levitt, L. N. Kanal, & J. F. Lemmer (Eds.), *Machine intelligence and pattern recognition* (Vol. 9, pp. 319–326). North-Holland.

Kyburg, H. E., Jr., & Teng, C. M. (2001). *Uncertain inference*. Cambridge University Press.

Kyburg, H. E., Jr., & Teng, C. M. (2006). Nonmonotonic logic and statistical inference. *Computational Intelligence, 22*(1), 26–51.

Li, K. (2023). Do large language models learn world models or just surface statistics? *The Gradient*. https://thegradient.pub/othello/

Levy, A. (2015). Modeling without models. *Philosophical Studies, 172*(3), 781–798.

Lobo, E., Agarwal, C., & Lakkaraju, H. (2024). On the impact of fine-tuning on chain-of-thought reasoning. *arXiv*. https://arxiv.org/abs/2411.15382

Loeffler, R. (2017). *Brandom*. Polity Press.

Mallory, F. (2023). Fictionalism about chatbots. *Ergo: An Open Access Journal of Philosophy, 10*, 38. https://doi.org/10.3998/ergo.4668

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*. https://arxiv.org/abs/2410.05229

Nickel, P. J. (2013). Artificial speech and its authors. *Minds and Machines, 23*(4), 489–502. https://doi.org/10.1007/s11023-013-9303-9

Peng, B., Narayanan, S., & Papadimitriou, C. (2024). On limitations of the transformer architecture. *arXiv*. https://api.semanticscholar.org/CorpusID:267636545

Peregrin, J. (2014). *Inferentialism: Why rules matter*. Palgrave Macmillan.

Peregrin, J. (2018). Is inferentialism circular? *Analysis, 78*, 450–454.

Peregrin, J. (2021). Do computers 'have syntax, but no semantics'? *Minds and Machines, 31*, 305–321.

Piantadosi, S. T., & Hill, F. (2022). Meaning without reference in large language models. *arXiv*. https://arxiv.org/abs/2208.02957

Ryle, G. (1949). *The concept of mind*. Hutchinson.

Saparov, A., & He, H. (2023). Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=qFVVBzXxR2V

Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417–457. https://doi.org/10.1017/S0140525X00005756

Searle, J. (1983). *Intentionality*. Cambridge University Press.

Sellars, W. (1953). Inference and meaning. *Mind, 62*(247), 313–338.

Sellars, W. (1958). Counterfactuals, dispositions, and causal modalities. In H. Feigl, M. Scriven, & G. Maxwell (Eds.), *Minnesota studies in the philosophy of science: Volume II: Concepts, theories, and the mind-body problem* (pp. 225–308). University of Minnesota Press.

Sellars, W. (1962). Naming and saying. *Philosophy of Science, 29*(1), 7–26.

Skinner, B.F. (1953). *Science and Human Behaviour*. New York: Macmillan.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Yang, Y. (2023). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=uyTL5Bvosj

Stechly, K., Marquez, M., & Kambhampati, S. (2023). GPT-4 doesn't know it's wrong: An analysis of iterative prompting for reasoning problems. In *NeurIPS 2023 foundation models for decision making workshop*. https://openreview.net/forum?id=PMtZjDYB68

Toon, A. (2012). *Models as make-believe: Imagination, fiction, and scientific representation*. Palgrave Macmillan.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, New Series, 59*(236), 433–460.

Valmeekam, K., Olmo, A., Sreedharan, S., & Kambhampati, S. (2022). Large language models still can't plan (a benchmark for LLMs on planning and reasoning about change). In *NeurIPS 2022 Foundation Models for Decision Making Workshop*. https://openreview.net/forum?id=wUU-7XTL5XO

Vaswani, A., et al. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Walton, K. (1990). *Mimesis as Make-Believe: On the Foundations of the Representational Arts*. Cambridge: Harvard University Press.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Fedus, W. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems* (Vol. 35, pp. 24824–24837). https://proceedings.neurips.cc/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf

Williams, I., & Bayne, T. (2024). Chatting with bots: AI, speech acts, and the edge of assertion. *Inquiry*, 1–24. https://doi.org/10.1080/0020174X.2024.2434874

Wittgenstein, L. (1953). *Philosophical investigations*. Blackwell.

Wittgenstein, L. (1958). *The Blue and Brown Book.* Oxford: Blackwell.

Wittgenstein, L. (1969). *On certainty*. Blackwell.

Zhang, H., Li, L. H., Meng, T., Chang, K.-W., & Van den Broeck, G. (2023). On the paradox of learning to reason from data. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence* (pp. 3365–3373). https://doi.org/10.24963/ijcai.2023/375