Forms of Understanding in Artificial Intelligence

Language, Representation, and the Question of Meaning

"PsyAI: Interactions and Resonances of the Human Unconscious and Artificial Intelligence", Kyiv, 22 March 2026

Daniel Defays

Introduction

Recent developments in large language models (LLMs) have revived a longstanding philosophical question: *what does it mean to understand?* Trained on massive corpora of text, these systems exhibit performances that often appear indistinguishable from human linguistic competence. They complete sentences, answer complex questions, solve mathematical problems, generate code, and even produce poetry. For some researchers, the breadth and flexibility of these abilities suggest that LLMs may instantiate a genuine—albeit novel—form of understanding (Chalmers, 2023; Hofstadter, 2023; Giraud, 2025). For others, however, such systems remain sophisticated pattern-matching devices, devoid of any access to meaning (Bender, 2021; Dentella et al., 2024).

This disagreement points to a deeper issue: the absence of a unified and operational definition of understanding. Recent philosophical work (e.g., Chalmers, 2023; Lyre, 2024) increasingly supports the view that understanding is not a unitary phenomenon but rather a multidimensional construct. Human understanding appears to involve several distinguishable forms or levels, which cannot be adequately captured by a minimal definition such as "access to meaning." Such a definition, while intuitive, proves insufficient to structure the current debate.

The aim of this paper is to use the evaluation of LLMs' capacities as a lens through which to revisit the concept of understanding itself. LLMs challenge traditional assumptions by exhibiting forms of competence that have no clear precedent in either biological or artificial systems. This raises a fundamental question: is understanding primarily grounded in embodiment, or can it emerge from linguistic structure alone?

The paper is organized as follows. After a brief review of classical approaches to evaluating understanding, we propose a fourfold classification, drawing in part on typologies introduced by Chalmers (2023) and Lyre (2024). We then examine how LLMs fit within this framework. Finally, we discuss how this analysis calls into question the adequacy of existing theoretical

models and opens broader issues concerning the roles of language, reference, and embodiment in the constitution of understanding and meaning.

Approaching the Question

There are, broadly speaking, three main ways of assessing whether a system understands (see, for instance, Lyre).

The first consists in examining its behaviour—what may be called an extrospective approach. From this perspective, understanding is inferred from observable performance: does the system answer correctly, reason coherently, explain the terms it uses, or deploy counterfactuals appropriately? This is the standpoint of an external observer evaluating the system from the outside, and it underlies the logic of the Turing Test. Contemporary research has refined this approach through increasingly systematic evaluation protocols. Numerous benchmarks—such as SQuAD, GLUE, Winograd Schema Challenge, BIG-bench, MMLU, and ARC—aim to measure a wide range of "cognitive" capacities. These include the ability to draw inferences, recognize patterns, form analogies, use common sense, and, more broadly, construct workable models of the world. In this framework, such performances are often treated as proxies for understanding. A well-known objection to this approach was formulated by John Searle in his Chinese Room argument: meaning cannot be reduced to formal symbol manipulation, and behavior alone remains fundamentally ambiguous. The same observable output may result from radically different internal processes.

A second strategy, more natural in the case of humans, is the introspective approach. Here, one directly asks the subject whether they understand what they are saying. Applied to artificial systems, this method yields problematic results. In February 2026, when asked whether it understands the text it generates, ChatGPT replied: "No. I predict patterns based on training data." When further asked whether it builds a model of the world, it answered: "Not in the human sense. I associate linguistic contexts statistically." Such responses seem to deny understanding. However, introspection is known to be unreliable even in humans, and even more so in artificial systems explicitly trained to avoid attributing to themselves consciousness or agency. Such answers should therefore be interpreted with caution. While the model denies constructing representations of the world in the usual sense, recent studies suggest that it may in fact develop internal structures that function in precisely this way—a point to which I will return later.

A third approach consists in examining the internal mechanisms of the system—a mechanistic approach. In the case of LLMs, we have partial access to how "meaning" is represented and processed. Words, or subword units, are encoded as numerical vectors (embeddings), and semantic relationships—such as similarity, opposition, or functional association—are captured through geometric relations within high-dimensional spaces. These representations are learned through a simple objective: next-token prediction. By training the model to complete sentences, structured representations of linguistic and, to some extent, world knowledge emerge. This makes it possible, at least in principle, to probe these internal spaces and search for representations of meaningful features. However, this approach also faces significant limitations. First, the scale of these models is immense: billions of parameters operating in spaces of thousands of dimensions make comprehensive interpretation extremely difficult. Second, the structures we are able to identify may not be the most relevant ones. The search for interpretable patterns risks projecting human-readable organization onto systems

whose operative principles may remain opaque. As cognitive science has long emphasized, the causes of behavior are not always transparent, even to the agents themselves.

In sum, each of these three approaches captures an important aspect of the problem, yet none is sufficient on its own. A more structured framework is therefore required. In what follows, I propose such a framework—partly inspired by the work of Lyre and Chalmers—and use it to assess the forms of understanding exhibited by LLMs.

What Do We Mean by Understanding?

Various frameworks have been proposed to characterize understanding. Holger Lyre, for instance, distinguishes between different forms of semantic grounding. For Lyre, a central feature of meaning is the anchoring of words in the objects to which they refer. This anchoring can occur through causal chains that ultimately fix reference. However, meaning is not exhausted by such causal relations. It also depends on the functional roles that words play within a system, as well as on their use within broader social practices. In this sense, social grounding can be seen as an extension of functional grounding beyond the boundaries of the individual system.

David Chalmers proposes a complementary threefold distinction between explanatory (e-), use-based (u-), and phenomenal (p-) understanding. One e-understands a phenomenon if one can explain it; one u-understands it if one can use it appropriately—for instance, to answer questions, draw inferences, or manipulate it in reasoning; and one p-understands it if one has the subjective experience of understanding. This last dimension explicitly introduces consciousness, a feature not directly addressed in Lyre's framework.

Building on these approaches, I propose a fourfold classification of understanding.

Referential Understanding

Referential understanding consists in linking words to objects, scenes, or entities external to language. To understand the word *tree* is to connect it to something in the world. This corresponds closely to Lyre's notion of causal grounding.

At first glance, LLMs appear incapable of such understanding, as they manipulate symbols without direct perceptual access to the world. However, recent research complicates this view, suggesting that some forms of referential structure may emerge indirectly from purely linguistic training. This point will be examined in the next section.

Inferential Understanding

Not all words refer directly to external entities. Concepts such as *freedom*, *democracy*, or *irony* derive much of their meaning from their relations to other concepts. Inferential understanding concerns precisely these relations.

In this respect, Ludwig Wittgenstein's claim that "the meaning of a word is its use in language" is particularly relevant. This form of understanding aligns with what Chalmers (2023) calls explanatory understanding: the capacity to explain, infer, and reason on the basis of abstract structures.

LLMs clearly exhibit strong inferential capacities. They manipulate complex conceptual networks, generate explanations, and reason about counterfactuals—for instance, regarding physical laws such as gravity or their consequences.

Pragmatic (Intentional) Understanding

Pragmatic understanding concerns the grasp of intentions and the ability to act appropriately in context. It corresponds to what Chalmers calls u-understanding: one understands something when one can use it effectively to achieve goals.

This dimension is also closely related to Lyre's notion of functional grounding. As already written, when such functional integration extends beyond the system itself and involves interaction with other agents, Lyre characterizes it as social grounding.

Pragmatic understanding presupposes, at least in its richer forms, a theory of mind—that is, the capacity to attribute mental states such as beliefs, desires, and intentions to others. Recent large language models appear, at least functionally, to approximate this capacity. Their responses suggest the construction of structured internal models in which agents are represented as acting on the basis of beliefs and goals.

For example, when presented with a social scenario—such as a waiter accidentally spilling wine—ChatGPT produces nuanced responses that simulate the attribution of mental states. Such performances, however, call for caution. A distinction must be maintained between the simulation of intentionality and genuine intentionality, and the attribution of the latter to artificial systems remains philosophically contested.

Experiential Understanding

Experiential understanding is grounded in embodiment and lived experience. To understand *slippery* is, in part, to have nearly fallen; to understand *fatigue* is to have experienced it. To understand the word *spring* is to anticipate warmth, to sense renewal, perhaps even to associate it with particular smells or sensations.

This form of understanding is intrinsically tied to sensorimotor experience and to the organism's situation in the world. In this sense, current artificial systems lack experiential understanding: they possess neither bodies, nor sensations, nor a lived engagement with their environment.

Weak form of referential understanding

The current absence of perceptual access to the world seems to deny any form of referential understanding to artificial networks. Several researches on the way artificial systems process information show however that they generate representations of the world that act like a reference to the external world. I will give four examples of the construction of such models of the world.

“Draw Me a Sheep”

In 2023, Thibaut Giraud (2025) asked a language model trained exclusively on text to “draw a sheep” using a simple graphical programming language. The model had never processed images, yet it produced a recognizable schematic representation. Through co-occurrence patterns and linguistic associations, the system appeared to construct an internal representation of the concept.

Stacking Objects

When asked how to stably stack a book, a laptop, nine eggs, a bottle, and a nail, ChatGPT proposes placing heavier, flatter objects at the bottom. It is difficult to imagine such an answer without some minimal internal model of physical constraints. Hofstadter (2023) suggests that when words “act as” things in the world, they thereby refer—and possibly mean—those things.

Learning Othello

Li et al. (2022) trained a small language model on sequences of moves from Othello games, represented solely as coordinates. The model was never given explicit information about a board, tokens or rules of the game. However, subsequent analysis revealed that the model had internally encoded representations of board states. This information was deeply embedded in the network’s activations and played a causal role in predicting subsequent moves.

Musical Embeddings

Music lacks referential semantics in the ordinary sense. However, when applying embedding techniques—similar to those used in language models—to sequences of notes considered as words, the resulting geometric representations encode pitch relationships, even though no explicit notion of pitch was provided. The system reconstructs the structural features necessary for successful prediction, suggesting that referential-like organization may emerge internally as a consequence of formal constraints.

LLMs trained purely to optimize next-token prediction appear to develop emergent capacities that extend beyond simple statistical completion. These include the construction of internal world models, potentially grounding a form of referential understanding.

Discussion

The fourfold distinction proposed in this paper, as well as the conclusions drawn from it, may be challenged on several grounds.

First, the classification itself may be regarded as incomplete. For example, Chalmers (2023) distinguishes an additional category—phenomenal understanding—which concerns the subjective, experiential dimension of cognition. Do LLMs possess anything like a feeling of understanding? Are they aware of their own capacities? Such questions shift the discussion toward the problem of machine consciousness. Whether phenomenal experience is constitutive of understanding is itself an open and contested issue. In any case, determining whether LLMs are conscious is a distinct and considerably more demanding problem—one

that lies beyond the scope of the present inquiry. The aim of this paper has been more modest: to examine functional and structural forms of understanding rather than subjective awareness.

Second, one may question the criteria used to attribute inferential understanding. Is the ability to generate coherent and convincing explanations sufficient evidence that a system genuinely understands inferential relations? Does producing an explanation imply grasping it? Here, we enter a longstanding philosophical debate about the relationship between performance and competence, simulation and possession. The fact that a system can behave as if it understands does not settle the ontological status of that understanding.

Third, the proposed classification may be questioned on a more fundamental ground. Contemporary developments in artificial intelligence suggest, as I have shown, that several forms of understanding appear to emerge as a natural consequence of advanced language mastery. If inferential reasoning, referential modeling, and even certain pragmatic capacities arise from large-scale linguistic training alone, does it still make sense to treat them as genuinely distinct forms? The apparent plurality of forms may reflect degrees of structural complexity rather than fundamentally different modes of understanding. One might therefore argue that the current trajectory of artificial learning invites a different type of distinction—less between functional categories and more between two overarching dimensions of cognitive organization. On the one hand, there is a form of understanding directly grounded in language mastery, which could be described as a language-based understanding: a competence arising from the internal organization of linguistic structures and their inferential relations. On the other hand, there is a more organic dimension, rooted in embodied and experiential engagement with the world, yet extending beyond mere sensorimotor coupling. From this perspective, the relevant divide might not lie within understanding itself, but between two fundamentally different sources of cognitive organization: linguistically emergent structure, on the one hand, and biologically embodied life, on the other. If this is correct, the fourfold taxonomy proposed above would require refinement—not because it is mistaken, but because artificial systems reveal a possible reconfiguration of the boundaries through which we traditionally analyze understanding.

Taken together, these criticisms do not invalidate the proposed framework, but they highlight the need for conceptual caution and methodological refinement.

The analysis of internal referential structures developped by the system raises an additional issue. As shown in the preceeding section, in several cases, we can establish correlations between the system's behaviour and representations of the world—for example in domains such as Othello/Reversi or music. But a more fundamental question remains: how certain can we be that these behaviors only rely on representations comparable to ours? Maybe other kinds of causal structures are at work—structures that are not directly intelligible or readable to us?

Conclusions

LLMs have not solved the problem of understanding. But they have destabilized it. The fourfold distinction proposed here allows for a more nuanced response to our initial question. The forms of understanding displayed by LLMs differ fundamentally from ours: they lack experiential grounding, and their referentiality is structurally inferred rather than perceptually anchored.

Language mastery appears to function not merely as one capability among others, but as a foundation from which a broad range of competencies can develop. To put it differently, in order to optimize text prediction, LLMs appear to develop novel capacities—such as problem solving, forms of reasoning, and also what may plausibly be described as a weak form of understanding. This could show that understanding may not require consciousness, not even require embodiment- at least not entirely.

And this leads to a provocative thought: maybe human reference is also not fully grounded in raw reality. Maybe much of our understanding is also linguistic, and inferential. It may not require embodiment—at least not entirely. This would mean understanding is an emergent organizational effect.

This leaves us with an open question: Do we understand because we are embodied or because we are structured by language?

References

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the dangers of stochastic parrots: Can language models be too big?* Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). https://doi.org/10.1145/3442188.3445922

Dentella, V., Guenther, F., Murphy, E., Marcus, G., & Leivada, E. (2024). *Testing AI on language comprehension tasks reveals insensitivity to underlying meaning.* arXiv. https://arxiv.org/abs/2302.12313

Chalmers, D. J. (2023, June 10). *Stochastic parrots or emergent reasoners: Can large language models understand?* Talk presented at the Large Language Models & Understanding session, ISC UQAM Summer School, Montreal, Canada.

Defays, D. (2024). From melodic note sequences to pitches using word2vec. *arXiv*. https://doi.org/10.48550/arXiv.2410.22285

Giraud, T. (2025). *La parole aux machines: Philosophie des grands modèles de langage*. Grasset.

Hofstadter, D. (2023, April). *Is there an "I" in AI?* https://berryvilleiml.com/wp-content/uploads/Is-there-an-%E2%80%9CI%E2%80%9D-in-AI-.pdf

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., & Wattenberg, M. (2024). *Emergent world representations: Exploring a sequence model trained on a synthetic task.* arXiv. https://arxiv.org/abs/2210.13382

Lyre, H. (2024). *"Understanding AI": Semantic grounding in large language models.* https://arxiv.org/pdf/2402.10992

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., & Fedus, W. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=V3U58fO1Nf