Philosophical Problems in Science (2025), **79**, 41–48
(Zagadnienia Filozoficzne w Nauce)
doi:10.59203/zfn.79.724

ARTICLE

Consciousness, normativity, and intrinsic meaning in large language models

Roman Krzanowski

The Pontifical University of John Paul II in Kraków, Poland
Email: rmkrzan@gmail.com

Abstract

Recent advances in large language models (LLMs), together with renewed speculation about artificial consciousness, have intensified debates concerning meaning and semantic intentionality in artificial systems. This paper advances a restricted and conditional thesis: even if LLMs were to possess some form of consciousness, this would not suffice to establish intrinsic semantic intentionality. To clarify this claim, I introduce a hieroglyph-copying thought experiment showing that consciousness, even when accompanied by goal-directed or procedural intentionality, does not guarantee normatively assessable semantic content. Drawing on this result, I argue that contemporary LLM architectures—despite their ability to generate contextually appropriate linguistic outputs—operate through probabilistic optimization processes that do not by themselves ground epistemic commitment or truth-directed representation. The argument does not deny the sophistication of LLM representations nor attempt to resolve competing theories of meaning. Rather, it isolates and defends an insufficiency thesis: consciousness alone does not ground intrinsic semantics. The apparent meaningfulness of LLM outputs is therefore best understood as dependent on human interpretive projection rather than as an internally grounded semantic achievement.

Keywords: large language models, intrinsic semantics, semantic intentionality, consciousness, normativity, philosophy of artificial intelligence

1. Introduction: The Question of Meaning in LLM Systems

Large language models (LLMs) (Vaswani, Shazeer, and Parmar 2017; Zhao et al. 2023; Chang et al. 2024; Fan et al. 2024) exhibit increasingly human-like linguistic capabilities. This development has reignited long-standing philosophical questions concerning the nature of understanding and meaning in artificial systems. Do these systems genuinely understand what they say, or do they merely execute highly sophisticated forms of pattern recognition without semantic comprehension?

Discussions of meaning in AI frequently invoke consciousness. If a system were conscious, the reasoning goes, it might also be capable of understanding or genuinely meaning its linguistic outputs. Some authors (Chalmers 2023; Hinton 2025) have suggested that advanced

AI systems may be exhibiting forms of proto-consciousness, raising the question of whether this could ground semantic understanding.

Throughout this paper, a distinction is drawn between procedural or goal-directed intentionality and semantic intentionality understood as contentful, truth-directed aboutness; the central claim concerns the insufficiency of the former for grounding the latter. By goal-oriented or procedural intentionality, I mean the directedness of actions toward outcomes or goals, which can be captured instrumentally or teleologically without presupposing intrinsic semantic content (Brentano 1874; Millikan 1984; Dennett 1987).

This paper argues that even if LLM systems were granted some form of consciousness, this alone would not be sufficient to secure intrinsic semantic meaning or genuine understanding. Intrinsic semantic meaning is meaning that belongs to a system itself, grounded in its own epistemic and normative relations, rather than imposed by an external interpreter. The paper does not attempt to settle the question of whether AI systems can be conscious. Rather, it examines whether consciousness—*if granted*—would be sufficient for intrinsic semantics or intentionality.

Although the paper refers to the Chinese Room argument, it does not seek to revive it, but rather to clarify a distinct confusion concerning the relation between consciousness and semantic competence.

To address this question, I first introduce the conditions under which a system can be said to possess meaning and understanding, emphasizing intrinsic semantic content, epistemic commitment, and normativity. I then introduce the Hieroglyph-Copying Scenario, which illustrates that consciousness and certain forms of intentionality may be present without semantic understanding. Next I examine large language models from an operational perspective and ask whether, given these conditions, they can be said to possess meaning, understanding, intentionality, or consciousness. For the sake of argument, I provisionally bracket the possibility that LLMs or future AI systems might exhibit some form of consciousness or intentionality. Finally, I argue that even if LLM systems—or future artificial systems—were to possess consciousness and intentionality, this would not by itself be sufficient to grant them intrinsic semantic meaning or genuine understanding.

The present argument is deliberately limited in scope. It does not attempt to refute all functionalist or informational theories of content, nor to settle the general metaphysics of intentionality. Its claim is conditional and restricted: even if large language models were granted forms of consciousness or goal-directed intentionality, this would not by itself suffice to ground intrinsic semantic content.

2. Meaning, Understanding, and the Conditions of Semantic Content

Discussions of meaning in artificial intelligence often proceed too quickly from observable linguistic performance to claims about understanding or semantic competence. In order to assess whether artificial systems can genuinely possess meaning, it is therefore necessary to clarify what is meant by *meaning* in the relevant philosophical sense and to distinguish it from merely functional or behavioral criteria.

The distinction drawn here parallels established contrasts in the literature between instrumental or derived intentionality (Dennett 1987), teleofunctional accounts (Millikan 1984), and truth-conditional or contentful representation (Dretske 1981; Fodor 1990). The present

use of "procedural intentionality" corresponds to forms of directedness that can be captured without invoking intrinsic truth-directed content, while "semantic intentionality" refers to states that are normatively assessable as correct or incorrect for the system itself.

In philosophy, meaning is not exhausted by the production of correct or useful outputs. To say that an expression is meaningful is to say that it is *about* something – that it represents, refers to, or purports to describe some object, concept, or state of affairs. This notion of aboutness is inseparable from normativity: meaningful representations can be correct or incorrect, true or false, appropriate or inappropriate. Meaning, in this sense, is not merely descriptive but evaluative, since it involves standards against which representations can succeed or fail.

Understanding is closely related to meaning but should not be identified with it. To understand is not simply to produce or manipulate meaningful symbols, but to stand in an epistemic relation to what those symbols represent. Understanding involves commitment to content, sensitivity to error, and responsiveness to reasons. A system that understands does not merely generate representations; it is answerable to how things are and can, in principle, recognize its own mistakes.

Intentionality is often invoked as the bridge between meaning and understanding, but here a crucial distinction must be drawn. On the one hand, there is procedural or goal-directed intentionality, understood as the directedness of actions toward outcomes or the systematic following of rules. On the other hand, there is semantic intentionality, which involves contentful mental states such as beliefs and judgments that are subject to standards of truth and justification. While semantic intentionality presupposes meaning, procedural intentionality does not. A system may act intentionally in the procedural sense—by following rules or pursuing goals—without thereby possessing semantic content or understanding.

Consciousness is frequently introduced into these debates as a potential grounding for meaning. It is often suggested that if a system were conscious, it might thereby be capable of genuine understanding. However, consciousness alone does not determine whether a system possesses intrinsic semantic content. Even if a system were granted some form of conscious experience, it would still be an open question whether its representations are normatively constrained – whether they function as beliefs or judgments that can be correct or incorrect for the system itself. Consciousness may be relevant to meaning, but it is not sufficient for it.

In human agents, semantic meaning is typically embedded within broader epistemic practices: interacting with the world, forming beliefs, correcting errors, and taking responsibility for representations. These practices illustrate how meaning is ordinarily grounded, but they do not by themselves establish necessary conditions that any system must satisfy in order to possess meaning. What matters is not the particular biological or social route by which meaning is achieved, but the presence of epistemic commitment and normativity.

The upshot is that neither linguistic performance, nor procedural intentionality, nor even consciousness guarantees intrinsic semantic meaning. A system may produce symbols that are meaningful to others, act intentionally with respect to its tasks, and even enjoy conscious states, while still lacking understanding. The following section introduces a thought experiment—the Hieroglyph-Copying Scenario—that makes this dissociation explicit and clarifies why meaning cannot be reduced to consciousness or intentional action alone.

3. The Hieroglyph-Copying Scenario

The Hieroglyph-Copying Scenario unfolds as follows. A person consciously copies Egyptian hieroglyphs without knowing their meaning. Although these signs may carry meaning for others (for example, Egyptologists), they remain meaningless to the person who copies them. The copier is conscious and intentionally draws the hieroglyphs, yet they do not convey any meaning *for him or her*. The person is not communicating a meaningful message, nor does he or she understand what the symbols stand for.

Of course, for someone versed in Egyptian hieroglyphs, the copied inscriptions may have semantic content. However, this meaning is not possessed by the copier. Rather, it is imposed by an external interpreter. The meaning of the hieroglyphs, in this case, belongs to the observer, not to the agent who produces them.

What do we learn from this scenario? We learn that even conscious agents can generate symbols that are meaningful *for others* without understanding them themselves. Moreover, systems that are conscious and intentional with respect to their actions do not thereby possess intrinsic semantic content. Consciousness and goal-directed intentionality are therefore not sufficient for intrinsic semantics.

The force of the scenario does not depend on introspective access as a necessary condition of meaning. The point is not that lack of meta-awareness entails absence of first-order content, but that consciousness alone—even when present—does not secure truth-directed, normatively assessable representation. The example isolates the insufficiency of consciousness, not the necessity of phenomenology for semantics.

4. Large Language Models: An Operational Overview

Large Language Models like GPT-4 or Claude (or software systems based on the similar architecture or any other systems with LLM architecture; see for a list of such systems: (Hadi et al. 2023; Ariffud 2026) are trained on vast corpora of human-generated text. They learn statistical associations between words, phrases, and contexts, allowing them to generate fluent and contextually appropriate responses (see the review of LLM systems by: Raiaan et al. 2024; Vaswani, Shazeer, and Parmar 2017).

LLMs are composed of large numbers of correlated parameters structured as multi-dimensional data arrays. Through training, these systems learn statistical associations between tokens (word fragments, words, or sequences of words), phrases, and sentences, enabling them to generate fluent and contextually appropriate responses over long sequences (Raiaan et al. 2024; Vaswani, Shazeer, and Parmar 2017). The process by which LLMs "learn" is sometimes likened to statistical pattern recognition (see e.g. Chang et al. 2024; Fan et al. 2024; Zhao et al. 2023; Hadi et al. 2023; Raiaan et al. 2024).

In transformer architectures (Vaswani, Shazeer, and Parmar 2017), tokens are mapped to high-dimensional embedding vectors, whose geometric relations encode statistical co-occurrence patterns. The attention mechanism dynamically weights contextual relations between tokens, enabling context-sensitive prediction of subsequent tokens. These operations allow for remarkable linguistic performance but remain governed by probabilistic optimization objectives rather than truth-evaluative commitments.

5. LLM and Understanding

LLM algorithms optimize predictive accuracy without any reference to the semantic content of the data. This description concerns their operational training objective, not a metaphysical denial of representational structure. Their "knowledge" is a byproduct of massive-scale correlation, not comprehension. The presence of human-in-the-loop reinforcement, such as RLHF (Reinforcement Learning from Human Feedback), may refine outputs, but it does not endow the system with intentionality. Consequently, while their outputs may appear coherent, even insightful, they are not meaningful in the sense defined by philosophical theories of language use.

The apparent meaning in LLM output arises from the interpretation imposed by human users who ascribe meaning to language based on context, experience, and shared conceptual frameworks. The LLM does not thereby know, believe, or mean anything in the intrinsic semantic sense defined above (see (Marcus and Davis 2020; Wolfram 2023). LLMs lack necessary properties requires for a system (any system) to possess meaning.

The LLM does not evaluate whether a response *makes sense* in a semantic or epistemic manner at any stage in this process. This is because it lacks any representation of meaning. Rather, the system optimizes numerical objectives defined over probability distributions within a multilayer computational procedure. The guiding criterion for output generation is, therefore, not semantic understanding. It is statistical optimization – typically the maximization of likelihood or related objective functions.

This point can be illustrated by a simple observation. Consider perturbations in the model's probability space (understood here as a high-dimensional vector space rather than a physical or mathematical "complex" space). These perturbations can yield outputs that are grammatically well-formed yet semantically incoherent. The truth of such an output is also evaluated with respect to the statistical coherence of syntax rather than meaning. This demonstrates that grammaticality and surface coherence are consequences of learned statistical regularities, not indicators of understanding.

6. Philosophical Thought Experiments: Ants, Chinese Rooms, and Hieroglyphs

These considerations do not introduce an independent objection to AI meaning, but reinforce the insufficiency result established by the Hieroglyph-Copying Scenario.

Two thought experiments illustrate the disconnect between symbol manipulation and genuine understanding:

1. Putnam's Ant (Putnam 1981): An ant wandering randomly on the sand may trace a shape resembling Winston Churchill. Despite the resemblance, the ant did not draw Churchill. The interpretation is imposed by an observer, not intended by the ant.
2. Searle's Chinese Room (Searle 1980; Cole 2024): A person inside a room manipulates Chinese symbols using a rulebook without understanding the language. From the outside, the person appears to understand Chinese, but inside, no understanding occurs. This demonstrates that syntactic manipulation does not entail semantic comprehension. Note that the Chinese Room argument does not equate to the claim that linguistic meaning, in a general sense, is solely internal to the mind, the brain, or a system. Searle has repeatedly emphasized the social and institutional dimensions of language, stressing its public, conventional, and culturally embedded character (e.g., Searle 1995). His broader

philosophy of language, therefore, supports a mixed internal–external account of meaning. The Chinese Room argument itself is more narrowly focused. It concerns systems that process linguistic inputs exclusively by means of internally sufficient formal procedures. These are procedures that are complete with respect to input–output mapping but lack any semantic or epistemic perspective. Any system whose operation is exhausted by such internally complete procedures (as described in the Chinese Room) cannot thereby generate intrinsic meaning or understanding.

These examples underscore a critical point: neither syntax, statistical semantics (what LLM systems create from input data), nor even consciousness alone ensures meaning. The Hieroglyph-Copying Scenario is particularly relevant as in this scenario a conscious person is responsible for generating "meaningful signs without understanding them" demonstrating that even conscious system may generate meaningful, for someone else, symbols without understanding their meaning. What is missing is the agent's semantic intentionality.

7. LLMs and understanding—Objections

A common objection to denying that LLMs possess meaning is that LLMs generate outputs by training on meaningful human language and thus inherit meaning through this process. However, this objection fails to grasp the distinction between statistical structure and semantic content. An LLM trained on syntactically correct but semantically meaningless data—such as a corpus of meaningless but grammatically plausible text—would still generate fluent responses. Even if such a corpus is hypothetical, the point illustrates the distinction between statistical regularity and semantic grounding. Yet these responses would not be meaningful, as the model cannot distinguish between meaningful and meaningless input without intentional or social context. Its performance is driven by probabilistic patterns, not by understanding or intention.

Another objection suggests that LLMs generate a kind of meaning through what is sometimes called *statistical semantics*. However, this term is misleading, as it bears little resemblance to *semantics* in the traditional philosophical or linguistic sense. Technically, statistical semantics refers to patterns of co-occurrence—statistical relationships between tokens, words, or phrases—not to meaning as understood through reference, intentionality, or comprehension. The atomic elements of LLMs—tokens and word fragments—are assembled based on probabilistic associations, not semantic understanding. Since semantics plays no role during the training or compilation of an LLM's dataset, it is a dubious claim to suggest that meaning somehow "emerges" from inherently non-semantic processes. Yet this is precisely what some proponents of LLMs—including certain developers and users—appear to argue, presenting it as an emerging capacity or ability of such systems (Schaeffer, Miranda, and Koyejo 2023; Wei et al. 2022).

One may argue (a functionalist) that normativity itself could emerge from sufficiently complex functional organization. The present argument does not deny this as a logical possibility. Rather, it maintains that contemporary LLM architectures, as currently understood, do not instantiate mechanisms that would ground epistemic commitment or truth-directed representation at the system level.

One might also object that intrinsic semantic intentionality could emerge from sufficiently complex computational systems. However, the argument presented here concerns a concep-

tual insufficiency rather than a limitation of scale or architecture: increases in complexity or statistical sophistication do not, by themselves, generate normatively binding epistemic commitments. Without an account of how such normativity could arise from purely formal operations, appeals to emergence remain explanatorily incomplete.

8. Conclusion: Consciousness Without Meaning

This paper set out to examine whether recent claims concerning consciousness and intentionality in large language models can ground genuine meaning or understanding in artificial systems. Rather than attempting to settle the question of whether LLMs are, or could become, conscious, the argument adopted a conditional strategy: even if consciousness and certain forms of intentionality were granted to artificial systems, intrinsic semantic meaning would not thereby follow.

The Hieroglyph-Copying Scenario was introduced to clarify this point. It demonstrated that a conscious agent may intentionally and systematically manipulate symbols that are meaningful to others without understanding them or possessing semantic content for itself. This dissociation shows that consciousness, even when combined with procedural or goal-directed intentionality, is insufficient for intrinsic semantics. Meaning requires more than the capacity to experience or to act intentionally; it requires epistemic commitment, normativity, and the ability to stand in a truth-directed relation to what is represented.

When applied to large language models, this result suggests a principled limitation. LLMs can generate linguistically appropriate and contextually sensitive outputs, but such performance does not by itself amount to understanding. Even if future systems were to exhibit forms of consciousness or intentional states, this would not automatically provide them with intrinsic semantic content. Without epistemic commitment—without beliefs that can be correct or incorrect for the system itself—semantic meaning remains externally imposed rather than internally grounded.

Takeaway message is that "Even if consciousness and certain forms of intentionality are granted, intrinsic semantic meaning does not follow."

The broader implication is that debates about meaning in AI should not focus exclusively on architectural complexity, scale, or even consciousness, but on the conditions under which semantic content can be intrinsic to a system. By separating consciousness from semantic competence, the paper clarifies a persistent conceptual confusion in discussions of artificial intelligence. Meaning without epistemic agency is only apparent meaning: a projection of human interpretation onto systems whose operations, however sophisticated, remain formally and normatively uncommitted.

Statement on the use of AI tools

Parts of the manuscript were linguistically revised with the assistance of AI-based language tools. All philosophical ideas, arguments, and conclusions are the author's own, and the author takes full responsibility for the content of the paper.

References

Ariffud, Muhammad. 2026. *What are large language models (LLMs), how do they work, and popular use cases.* https://www.hostinger.com/tutorials/large-language-models accessed February 22, 2026.

Brentano, Franz. 1874. *Psychology from an Empirical Standpoint.* Translated by A.C. Rancurello, D.B. Terrell, and L.L. McAlister. London: Routledge & Kegan Paul.

Chalmers, David. 2023. *Could a large language model be conscious?* https://arxiv.org/abs/2303.07103 accessed June 22, 2026.

Chang, Yupeng, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, et al. 2024. A Survey on Evaluation of Large Language Models. Version Number: 9, *ACM transactions on intelligent systems and technology* 15 (3): 1–45.

Cole, David. 2024. *The Chinese Room Argument.* https://plato.stanford.edu/entries/chinese-room/ accessed February 22, 2026.

Dennett, Daniel. 1987. *The Intentional Stance.* Cambridge: MIT Press.

Dretske, Fred I. 1981. *Knowledge and the Flow of Information.* Cambridge: MIT Press.

Fan, Lizhou, Lingyao Li, Zihui Ma, Sanggyu Lee, Huizi Yu, and Libby Hemphill. 2024. A Bibliometric Review of Large Language Models Research from 2017 to 2023. *ACM Transactions on Intelligent Systems and Technology* 15 (5): 1–25. https://doi.org/10.1145/3664930.

Fodor, Jerry A. 1990. *A Theory of Content and Other Essays.* Cambridge: MIT Press.

Hadi, Muhammad Usman, Qasem Al Tashi, Rizwan Qureshi, Abbas Shah, Amgad Muneer, Muhammad Irfan, Anas Zafar, et al. 2023. *Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects.* https://doi.org/10.36227/techrxiv.23589741.v4.

Hinton, Geoffrey. 2025. *'Godfather of AI' predicts it will take over the world.* https://www.youtube.com/watch?v=vxkBE23zDmQ accessed February 22, 2026.

Marcus, Gary, and Ernest Davis. 2020. *GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about.* https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/ accessed February 22, 2026.

Millikan, Ruth Garrett. 1984. *Language, Thought, and Other Biological Categories: New Foundations for Realism.* Cambridge: MIT Press.

Putnam, Hilary. 1981. *Reason, truth and history.* Cambridge: Cambridge University Press.

Raiaan, Mohaimenul Azam Khan, Md. Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges. *IEEE Access* 12:26839–26874. https://doi.org/10.1109/ACCESS.2024.3365742.

Schaeffer, Rylan, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems* 36:55565–55581.

Searle, John. 1980. The intentionality of intention and action. *Cognitive Science* 4:47–70.

Searle, John. 1995. *The construction of social reality.* London: Penguin Books.

Vaswani, Ashish, Noam Shazeer, and Niki Parmar. 2017. Attention is all you need. In *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems,* edited by Ulrike Luxburg, Isabelle Gyuon, and Samy Bengio, 6000–6010. New York: Curran Associates Inc.

Wei, Jason, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, et al. 2022. *Emergent Abilities of Large Language Models.* Version Number: 2. https://arxiv.org/abs/2206.07682 accessed February 23, 2026.

Wolfram, Stephen. 2023. *All-In Summit: Stephen Wolfram on computation, AI, and the nature of the universe.* https://www.youtube.com/watch?v=%7B2%7DcQmQIYNI5M accessed February 22, 2026.

Zhao, Wayne Xin, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, et al. 2023. *A Survey of Large Language Models.* Version Number: 18. https://arxiv.org/abs/2303.18223 accessed February 23, 2026.