Philosophy Compass

WILEY

ARTICLE OPEN ACCESS

The Philosophy of Language Models

Raphaël Millière[1] | Cameron Buckner[2]

[1]Faculty of Philosophy, University of Oxford, Oxford, UK | [2]Department of Philosophy, University of Florida, Gainesville, Florida, USA

Correspondence: Raphaël Millière (raphael.milliere@philosophy.ox.ac.uk)

Received: 12 March 2026 | **Revised:** 12 March 2026 | **Accepted:** 18 May 2026

ABSTRACT

The success of large language models (LLMs) across many domains of AI research has generated intense debate. Some attribute their impressive performance on complex tasks to human-like linguistic and cognitive capacities, whereas others ascribe it to shallow pattern matching. These disputes stem from deep-seated philosophical disagreements about the nature of language and cognition. We provide an opinionated survey of these disagreements across core topics in the philosophy of mind and language, including syntactic competence, compositionality, linguistic meaning, representation, attitudes, reasoning, agency, and consciousness. We contend that progress on these issues requires not only clarity about background philosophical commitments but also, in many cases, close engagement with emerging empirical evidence.

1 | Introduction

In his *Discourse on the Method*, Descartes famously claims that while we might conceive of a machine designed to imitate human behavior—including speech—"it is not conceivable that such a machine should produce different arrangements of words so as to give an appropriately meaningful answer to whatever is said in its presence, as the dullest of men can do" (Descartes 1985, 140). Almost four centuries later, large language models (LLMs) appear to challenge Descartes' claim. For philosophers, these systems can be seen as thought experiments brought to life, reviving long-standing philosophical issues and introducing new ones.

LLMs are undeniably impressive engineering artifacts. Since the first widely recognized LLM (GPT-3) was unveiled in 2020, these models' capabilities have advanced at breakneck speed. They can fluidly converse with humans on virtually any topic and reliably pass short controlled Turing tests (Jones et al. 2025). Leading LLMs have matched or exceeded human performance on benchmarks spanning graduate-level scientific knowledge, advanced mathematics, computer programming, and various general and professional exams.[1] Some AI researchers suggest that LLMs display initial "sparks of artificial general intelligence" (Bubeck et al. 2023) and might even equal or surpass general human intelligence within a few years (Kokotajlo et al. 2025).

Nevertheless, many remain skeptical that LLMs' seemingly impressive performance warrants ascriptions of human-like linguistic and cognitive competence. Critics point to LLMs' architecture, learning environment, and spectacular failures as evidence for a deflationary account that attributes their capabilities to memorization and shallow pattern matching rather than deeper cognitive sophistication (Bender et al. 2021; Kambhampati, Stechly, and Valmeekam 2025; Mitchell and Krakauer 2023). For every claim that LLMs possess some human-like competence—"understanding," "reasoning," "belief"—there are equally forceful skeptical dismissals.

These disagreements turn on deep and often unacknowledged philosophical issues. The present paper aims to elucidate the philosophical concepts, theories, arguments, and evidence that bear on debates about the capacities and limitations of LLMs. We begin by clarifying what LLMs are and outlining methodological issues (Section 2), then discuss core debates about syntactic competence (Section 3), compositionality (Section 4),

meaning (Section 5), representation (Section 6), propositional attitudes (Section 7), reasoning (Section 8), and finally agency and consciousness (Section 9).

2 | Conceptual Preliminaries

2.1 | What Are Language Models?

Language models are artificial neural networks: computational systems loosely inspired by biological brains that learn to perform tasks from data rather than by following explicitly programmed rules.[2] A neural network consists of layers of interconnected nodes, or "neurons," where each connection has an associated numerical weight. When given an input (e.g., a sequence of words), the network propagates information forward: each neuron computes a weighted sum of its inputs, applies an activation function, and passes the result to the next layer. The entire network thus implements a complex transformation from input to output, determined by its architecture (how the layers are wired together) and the values of its weights. The learning process, or "training," involves gradually adjusting the network's weights to optimize a *loss function*, which measures performance on a given task (the *learning objective*). An optimization algorithm iteratively updates the weights on each training example to reduce this loss. Repeating this process across millions or billions of examples gradually tunes the network into a configuration that performs the task well.

The learning objective that gives language models their name is, as one might expect, language modeling: learning a probability distribution over sequences of text. Rather than processing words directly, they operate over a fixed vocabulary of *tokens*, which may represent whole words or subword units. Language models are trained through next-token prediction by computing the probability of each possible next token given a preceding sequence: $P(\text{token}_i|\text{token}_1, \text{token}_2, ..., \text{token}_{i-1})$. For example, given "To be or not to __," the model might assign the highest probability to "be". After training, it can generate text to complete an input sequence by iteratively predicting and appending the most probable next tokens, feeding the growing sequence back into itself at each step.

The dominant architecture for language models is called the transformer.[3] Its central innovation is the attention mechanism, which allows the model to encode contextual relationships between tokens. Each input token is first mapped to a numerical vector (an embedding), initially generic but progressively refined across layers based on its context. At each layer, attention updates these vectors by computing, for every token, a weighted combination of information from all others, where weights reflect contextual relevance. This enables the model to capture dependencies between tokens (e.g., inferring that "it" in "The student chose the class because it was easy" refers to "class"). Multiple attention heads perform this operation in parallel at each layer of the network. Each head can specialize in tracking different types of relationships between tokens, such as subject–verb agreement or anaphoric reference. As the input sequence flows across layers, the model constructs increasingly rich context-sensitive representations of the sequence's tokens, which guide its prediction for the next token.

Next-token prediction is only part of the story. Creating a conversational chatbot such as ChatGPT involves two stages: pretraining and posttraining.[4] During pretraining, a transformer neural network is trained on the next-token prediction objective using an internet-scale corpus of text and code, often containing trillions of words from websites, books, articles, and other sources. The result is a base model: a highly capable but hardly useful completion engine, which is good at continuing text in plausible ways, but cannot reliably follow instructions or engage in conversation. Posttraining adapts this base model for practical use. It typically involves fine-tuning (i.e., further training) it on a curated set of prompt-response examples that teach it the format and norms of conversation. A further alignment step then fine-tunes the model's behavior using human-ranked responses to favor outputs reflecting behavioral norms such as helpfulness, honesty, and harmlessness.

2.2 | Background Methodological Issues

Disputes about the capacities and limitations of LLMs are starkly polarized. The same system that one researcher views as genuinely intelligent strikes another as a pattern-matching device whose surface linguistic fluency invites reckless anthropomorphism. Most troubling, we lack consensus on how to arbitrate these disagreements. This is where philosophical analysis earns its keep. Rich psychological terms such as "reasoning" or "understanding" are used in different ways by different authors across disciplines (Shevlin and Halina 2019). Without clarifying what we mean by these terms, we risk talking past one another. More fundamentally, researchers bring different background assumptions about the nature of meaning, reference, representation, or psychological attitudes to these debates. These assumptions—often left implicit—determine which evidence counts as relevant and how that evidence should be interpreted. Mapping these philosophical choice points explains why knowledgeable researchers assessing the same system can reach radically different conclusions.

Evidence about LLMs falls into two categories: properties of untrained models (architecture, learning objective, training data) and properties of trained models (behavioral performance, internal processing mechanisms). Given particular background commitments, some questions about LLMs can be settled before we even train them. For example, the training data and learning objective of LLMs plausibly constrain what they can learn and represent. If a model only ever learns from predicting the next token in text sequences, can it form representations of the world that text describes? It is reasonable to think that properties of untrained models can help us address these kinds of questions. There are also technical arguments that infer limitations from properties of untrained models. For example, complexity theory can provide provable bounds on what transformer architectures can compute under various constraints (Strobl et al. 2024). This implies that certain computational problems are, in principle, beyond their reach without specific architectural modifications or inference strategies.

However, we must resist what we might call the redescription fallacy: inferring that because LLMs are pretrained to minimize next-token prediction error, all they do is predict tokens, and

therefore cannot possess sophisticated capabilities such as reasoning. The fallacy rests on two mistakes: it ignores that base language models, as we have seen, undergo subsequent fine-tuning that optimizes them for new functions beyond next-token prediction, and it confuses the learning objective that shaped a system with a complete description of what that system does.[5] The latter concern relates to the claim that LLMs exhibit "emergent capabilities" (Wei et al. 2022). One version of this claim suggests that certain capacities can arise as byproducts of training on next-token prediction without needing to be deliberately instilled through task-specific fine-tuning. For example, a language model trained only to predict text may develop the ability to solve mathematical problems or translate between languages. In other words, we cannot reliably predict all capabilities a model will develop until it has been trained. Consequently, to arbitrate substantive philosophical disputes about the functional capacities of LLMs, analyzing pre-training conditions is often insufficient: we must turn to empirical study of the trained artifacts themselves.

Whether this observation justifies an appeal to emergence is debated because the capabilities of trained models can often be traced back a posteriori to features of the training data (Rogers 2024). For example, it is not so mysterious or impressive that an LLM can generate plausible-looking chess moves if its training data contain thousands of chess game transcripts. Indeed, there is a common skeptical worry that LLMs merely memorize and regurgitate training sequences. Because they are trained on a significant subset of all human outputs, parroting this data could produce intelligent-seeming behavior.[6] This recalls Block (1981)'s "Blockhead"—a thought experiment proposed as a counterexample to the Turing test, in which a machine can hold intelligent conversations by simply searching for user inputs and responding with matching strings using a giant database of all possible exchanges. However, LLMs are not Blockheads in a literal sense. Although they do memorize some training sequences verbatim (Prashanth et al. 2025), they demonstrably produce genuinely novel outputs absent from their training corpus (Padmakumar et al. 2025), and their capacity for memorization is bounded (Morris et al. 2025).

A more sophisticated skeptical position holds that LLMs remain limited to "pattern matching" through interpolation between memorized examples. In statistical learning, interpolation refers to making predictions for inputs (or tasks) that lie in regions of the input/task space that are well supported by the training distribution—that is, cases that are sufficiently similar to training examples under an appropriate notion of similarity. In contrast, extrapolation involves making predictions in regions with little or no training support (often described as out-of-distribution generalization). The sophisticated skeptic claims LLMs largely succeed in the former regime and fail to generalize reliably in the latter. This concern is substantiated by evidence that their success is often brittle, such that superficial changes to a task's phrasing or structure can cause performance to collapse. Skeptics argue that this is evidence that LLMs' seemingly sophisticated capabilities are contingent on task familiarity rather than an understanding of the task's underlying structure.

This raises an important methodological issue: we cannot always take LLMs' success on flashy benchmarks at face value because they can be right for the wrong reasons. This relates to the familiar distinction from cognitive science between performance and competence: a system's performance on a task provides at best defeasible evidence for its underlying competence in the task domain.[7] When LLMs succeed on benchmarks but fail on trivial variations, performance does not reflect genuine competence—which can be due to training data contamination (benchmark items or similar content appearing in the training corpus) or exploitation of shallow regularities (reliance on superficial statistical patterns that happen to correlate with correct answers rather than the underlying task structure). Conversely, performance failures may not always indicate a lack of competence, as performance can be hindered by auxiliary factors unrelated to the target capacity.[8] In either case, we must interpret behavioral evidence with caution. To address these worries, we should design behavioral experiments applying best practices from cognitive science. Such experiments should be clear about the target construct and task demands, and use novel stimuli and carefully controlled conditions to rule out confounding explanations based on superficial cues or training data associations (Frank 2023). Furthermore, we can go beyond behavioral evidence by using interpretability methods to understand the internal mechanisms that cause LLMs' performance on the task (Grzankowski 2024; Millière and Buckner 2026). As we shall see, these two strategies—well-designed behavioral experiments and causal interpretability—can help us arbitrate disputes about whether LLMs' performance patterns are indicative of underlying competence in various domains where properties of trained models are particularly relevant.

3 | Linguistic Competence

It is uncontroversial that language models learn about the statistical distribution of words in their training corpora. But should we take them seriously as "models of language" in a deeper sense? A central debate in modern linguistics, originating with Chomsky's arguments against the statistical models of his day, concerns whether the statistical properties of language data are sufficient for learning its underlying structure. Chomsky famously argued that grammaticality is distinct from statistical frequency; for example, the nonsensical but grammatical sentence "Colorless green ideas sleep furiously" could be judged by a statistical model as equally remote from English as its ungrammatical counterpart, "Furiously sleep ideas green colorless" (Chomsky 1957). For decades, this view held that linguistic competence could only be characterized by formal, rule-based systems, and that learning such a system from data alone was impossible (Everaert et al. 2015).

Contemporary language models challenge this long-held skepticism.[9] Drawing on methods from psycholinguistics, researchers have conducted controlled behavioral experiments to evaluate models' implicit linguistic knowledge. A common strategy involves testing models on minimal pairs of stimuli, such as two sentences that are identical except for a single feature that makes one grammatical and the other ungrammatical. Instead of asking for explicit grammaticality judgments, researchers measure the probability a model assigns to a correct continuation versus an ungrammatical one. For instance, to test for sensitivity to subject–verb agreement, a model given "The keys to the cabinet __" is evaluated on the relative probabilities it assigns

to "are" versus the incorrect singular "is". The intervening noun "cabinet" serves as an attractor, designed to test whether the model is sensitive to the sentence's hierarchical structure (by correctly identifying "keys" as the subject) or whether it is merely relying on a shallow heuristic like agreement with the linearly closest noun (Linzen et al. 2016). A large body of evidence from such behavioral studies, covering a wide range of syntactic phenomena, suggests that modern language models' predictions are indeed sensitive to abstract syntactic structure—including nonlocal dependencies that were previously thought to be beyond the reach of statistical learners.

We encounter here the familiar concern from Section 2.2: models might achieve good performance on these tests by relying on shallow heuristics instead of representing the underlying syntactic structure. One way to rule this out involves controlling a model's training data to assess whether it can genuinely generalize out-of-distribution—that is, correctly apply learned rules to novel examples that differ in targeted ways from the training data. For example, models trained exclusively on text where subjects and verbs are adjacent can still correctly generalize the rules of agreement to novel sentences where subjects and verbs are separated by many words (Ahuja et al. 2025). Another approach involves probing the models' internal activations to see if linguistic information can be decoded from them. On its own, this kind of decoding method is merely correlational: it can show that information is present in model activations, but not that the model *causally relies* on it for its predictions.[10] To establish a causal link, researchers can actively manipulate a model's internal states to observe the impact on its behavior. For example, studies have demonstrated that erasing number-related information from a model's activations directly impairs its ability to perform subject–verb agreement, providing strong evidence that the model causally uses this encoded information (Lasri et al. 2022).

This causal approach is central to the nascent field of mechanistic interpretability. In the philosophy of science, a mechanism is an organized system of components whose activities and interactions produce a phenomenon (Machamer et al. 2000). In the case of neural networks, the relevant phenomenon is the network's behavior on a task of interest, and the relevant mechanism is typically a subset of the network—also known as a "circuit"—that is causally responsible for that behavior. The aim of mechanistic interpretability is to identify the circuits responsible for a network's behavior on a particular task, and understand the computations these circuits implement.[11] Recent work in this area has identified interpretable circuits that implement syntactic computations such as subject–verb agreement (Marks et al. 2024) and filler-gap dependencies (Boguraev et al. 2025) in language models. The upshot of this research is that modern language models can acquire sensitivity to hierarchical syntactic structure and nonlocal dependencies. This challenges the categorical skepticism underlying Chomsky's original arguments against statistical approaches to language.[12]

Whether these findings bear on debates about language acquisition is a further question.[13] Experiments with language models put pressure on the strong claim that syntax is not learnable from data even in principle. However, this claim should be distinguished from the so-called "poverty of the stimulus" argument (POS).[14] POS holds that the primary linguistic data available to children is compatible with many possible grammars, and that the correct grammar is not in any pretheoretic sense simpler or more natural than the alternatives—giving a general-purpose learner no reliable basis for selecting it. However, children reliably acquire the correct grammar—including structure-dependent operations—without making errors consistent with competing alternatives, and without receiving explicit correction. From this, POS concludes that children must possess innate, domain-specific constraints that guide them toward the correct grammar. LLMs do not directly challenge this argument because their training data differs radically from the data children learn language from, both in nature and scale. They learn from tokenized text and code rather than speech, and are exposed to orders of magnitude more words than any child encounters. However, experiments with small language models trained in more developmentally plausible scenarios can in principle shed light on aspects of language acquisition (Warstadt and Bowman 2022; Wilcox et al. 2025). By manipulating the training data of these small models, we can test specific learning hypotheses—for example, whether certain constructions are learnable from indirect evidence alone (Leong and Linzen 2024).

4 | Compositionality

A related debate concerns whether language models can account for the productivity and systematicity of language and thought. In a linguistic context, productivity refers to the ability of speakers to produce and understand a virtually infinite number of novel sentences from a finite set of words and grammatical rules. Systematicity describes the lawful relationships between linguistic abilities; for example, any speaker who understands the sentence "Jerry loves Paul" can also understand "Paul loves Jerry". The classic explanation for both of these phenomena is the principle of compositionality—the claim that the meaning of a complex expression is determined by the meanings of its constituent parts and the way they are syntactically combined (Partee 1984). Because productivity and systematicity are commonly taken to be features of thought beyond natural language, the principle of compositionality has been applied more broadly to mental representations. In particular, proponents of the "language of thought" hypothesis argue that much of cognition involves language-like representations that have a compositional syntax and semantics, allowing for coherent, rule-governed mental processes (Fodor 1975; Quilty-Dunn et al. 2023).

Compositionality has long been considered a major challenge for artificial neural networks, by contrast with classical symbolic architectures. The latter operate on discrete symbols with an explicit compositional structure, and their processes are sensitive to this structure. In contrast, neural networks use distributed representations (patterns of activation across many units) and learn statistical regularities from data by adjusting connection weights; they lack built-in mechanisms for representing and manipulating discrete, semantically meaningful constituents in a causally direct way. This motivated Fodor and Pylyshyn (1988)'s infamous dilemma: either neural network models fail to capture the systematicity and productivity of language and cognition, or, where they do succeed, they merely implement a classical symbolic architecture.

This long-standing debate has been reignited by LLMs' apparent proficiency in parsing and generating novel sentences with complex compositional structure. It has prompted researchers to test their capacity for compositional generalization: the ability to parse and produce novel combinations of previously seen elements. A model that generalizes compositionally should be able to systematically recombine familiar words and structures to handle inputs that are structurally different from those in its training data. However, because LLMs are trained on vast uncontrolled corpora that span virtually any syntactic construction, it is difficult to determine whether their success stems from genuine compositional generalization or from having encountered and memorized similar constructions during training. To address this challenge, researchers turn to synthetic datasets designed to exhibit a controlled distribution shift between the training and test data (Keysers et al. 2019; N. Kim and Linzen 2020; Lake and Baroni 2018). In these datasets, the training data is intentionally restricted—for example, it might contain a word only in a subject position, or sentences with a recursion depth of only one or two. The test data then contains novel combinations of these familiar elements, such as the same word used in an object position, or sentences with deeper recursive structures. Success on these tests requires a model to generalize beyond the specific patterns in its training data, providing stronger evidence for a grasp of abstract compositional rules.

Findings from this line of research are nuanced, and their interpretation is still debated. We can draw a distinction between two aspects of compositional generalization: lexical generalization (recombining known words in familiar grammatical structures) and structural generalization (applying rules to create grammatical structures not seen in training) (Donatelli and Koller 2023). Standard transformer models achieve near-perfect accuracy on tests of lexical generalization in synthetic benchmarks. However, they fail almost completely on tests of structural generalization, such as generating sentences with deeper prepositional phrase recursion or using a prepositional phrase to modify a subject noun when only object modifications were seen in training (Yao and Koller 2022). However, recent work suggests these failures may be misleading, arguing that they stem not from a fundamental inability to generalize but from models latching onto superficial artifacts in the benchmark's logical form (Z. Wu, Qiu, et al. 2024). When these artifacts—such as spurious correlations between a structure and its linear position—are removed through meaning-preserving adjustments to the training data, standard transformers show significant improvement on these tests of structural generalization.

An influential paper by Lake and Baroni (2023) proposes "meta-learning for compositionality" (MLC) to improve compositional generalization. Instead of using a single massive dataset, MLC trains a transformer on a curriculum of many distinct, randomly generated compositional tasks. This process forces the model to "learn how to learn" by inferring the compositional rules of a new mini-language from few examples. In direct comparisons on novel instruction-following tasks, their MLC-trained model not only achieves human-level accuracy but also reproduces characteristic human inductive biases and error patterns, which sometimes deviate from perfect systematicity. However, the claim that this work conclusively resolves the first horn of Fodor and Pylyshyn's dilemma is contested. Critics argue the model's success is brittle and not truly systematic, demonstrating that it fails on trivial variations of learned structures, is sensitive to arbitrary features of the input (such as the specific mapping between labels and meanings), and cannot generalize beyond the structural constraints of its meta-training distribution (e.g., to a greater number of repetitions) (Goodale and Mascarenhas 2023; Woydt et al. 2025). These failures, they contend, show that the model has not acquired genuine structure-sensitive rules or compositional competence in the classical sense, but rather a highly sophisticated form of pattern matching. The debate thus turns on whether the goal is to model the guaranteed competence of an idealized symbolic system or the more nuanced, and sometimes unsystematic, performance of human subjects.

Once again, we face the concern that mere behavioral success on compositional generalization tasks might not reflect underlying compositional competence. Interpretability methods can address this concern by assessing if models actually learn systematic computations over structured representations (Vegner et al. 2025). Researchers have begun to uncover evidence for such mechanisms. For instance, transformers trained on a synthetic dataset can learn systematic variable binding—a cornerstone of symbolic computation—by developing specialized circuits that use the model's vector space as a form of addressable memory (Y. Wu et al. 2025). On specific binding tasks, actual language models have been found to implement a mechanism that assigns abstract "binding ID" vectors to entities and their attributes, where the association is encoded through vector addition in a dedicated subspace of the model's activations (Feng and Steinhardt 2023). Further work has reverse-engineered the algorithm that a transformer implements to achieve compositional generalization on a synthetic task, identifying specific attention heads responsible for distinct computational roles such as broadcasting positional indices and retrieving function arguments (Tang et al. 2025). Collectively, these findings suggest that Transformers can induce systematic computations using vector-based representations with latent compositional structure; whether this suggests that they actually implement a classical architecture with the hallmarks of a "language of thought" remains a matter of debate.[15]

5 | Meaning, Reference, and Communication

LLMs produce sentences that seem meaningful to us. However, are we merely projecting meaning onto fundamentally meaningless strings because they are similar to sentences uttered by humans? This question has captured the interest of linguists and philosophers alike. As systems trained only on text without direct interaction with the world, LLMs face a version of the classic "grounding problem" from Harnad (1990): a system cannot learn the meaning of words by simply consulting a dictionary, as this would take it from one set of symbols to another in a circular loop without connection to those symbols' worldly referents. A language model trained exclusively on a text corpus appears to be in an analogous predicament: it can master the complex statistical relationships between tokens—learning which tokens are likely to follow others—but it seemingly never connects any of those symbols to the nonlinguistic, real-world referents they signify (Mollo and Millière 2025). This problem motivates the skeptical position that LLMs' outputs are intrinsically meaningless. For example,

Bender and Koller (2020) argue that LLMs are constitutively unable to learn linguistic meaning and produce genuinely meaningful outputs, because the necessary connection between language and the world is absent from their learning environment. Likewise, Titus (2024) argues that the behavior of LLMs is best explained by sensitivity to statistical properties rather than genuine semantic properties.

This debate—and the kind of evidence that can be brought to bear on it—depends crucially on background metasemantic commitments.[16] The first major line of response to the skeptical view comes from philosophers drawing on externalist theories of reference developed by Kripke (1980) and Putnam (1975). According to these theories, linguistic expressions can refer to objects in the world not through speakers' descriptive knowledge or direct experience, but through causal-historical chains connecting current usage to initial events of "baptism". This opens a potential route for LLMs' outputs to refer: if the human-generated training data contains referential terms embedded in appropriate causal chains, LLMs may inherit these referential connections. Mandelkern and Linzen (2024) explicitly develop this line of argument. Drawing on Kripke's arguments from ignorance and error—which show that speakers can refer successfully even with minimal or incorrect beliefs about referents—they argue that words generated by LLMs could refer to extra-linguistic entities and properties as long as LLMs can be considered part of a linguistic community with the right causal-historical connections between words and referents.[17]

This externalist approach faces a challenge regarding the role of intentions in reference transmission (Ostertag 2025). Kripke's original account stipulated that a speaker must intend to use a name with the same reference as the person from whom they learned it; however, it remains unclear whether LLMs can form such intentions. A few different strategies have been proposed to address this concern. Koch (2025) argues that LLMs' architectural design can functionally replace human referential intentions. Because LLMs are built to reproduce patterns from training data, their design ensures the continuity between past and future usage that Kripke's intention requirement was meant to secure. This allows LLMs to achieve successful reference for proper names and natural kind terms without possessing mental states.[18] Pepp (2025) goes further, defending an "austere" account of reference borrowing that eliminates the intention requirement entirely. Drawing on cases of involuntary reference and children's linguistic competence, she argues that reference transmission can occur through mechanical repetition, making LLMs' lack of intentions unproblematic for basic reference borrowing. Lederman and Mahowald (2024) propose a similarly minimalist view called bibliotechnism, which posits that LLMs are cultural technologies whose outputs are derivatively meaningful by virtue of causally inheriting meaning from the human-produced text in their training data. Likewise, Borg (2025) contends that such inherited causal–historical links are sufficient for the outputs of LLMs to possess type-level semantic content through semantic deference to the linguistic community, even though LLMs still lack the "original intentionality" required for agency or conscious understanding.

A further challenge for externalist views is what Lederman and Mahowald (2024) call the "novel reference problem". LLMs appear capable of introducing new names for entities, including entities they create (such as components of ASCII art). This is puzzling for views that ground the meaning of LLMs' outputs purely in mechanical transmission from their training data because novel reference seems to require some form of referential intention. In response, Lederman and Mahowald (2024) propose to extend bibliotechnism by embracing an interpretationist approach: if attributing attitudes such as beliefs and intentions to LLMs provides the most tractable explanation of their behavior, then they can be said to have such attitudes.[19] In particular, LLMs could be said to have the referential intentions necessary for introducing new names, providing a solution to the novel reference problem. Pepp (2025) suggests instead that novel reference might be grounded not in full-blown propositional attitudes but in a more primitive capacity for "basic reference": a pre-linguistic ability to "think of" or "have in mind" a particular object, which LLMs might achieve through their processing of inputs and generated content.

An alternative strategy sidesteps reference and externalism, drawing instead on conceptual role semantics. On this view, meaning arises from the network of inferential roles a concept plays within a wider system (Block 1986). Piantadosi and Hill (2022) argue that LLMs' performance suggests they are learning these meaning-defining conceptual roles. They challenge the primacy of reference by noting that many perfectly meaningful concepts such as "justice" or "perpetual motion machine" lack clear real-world referents. Their meaning is instead derived from their position within a web of related concepts. Because text is a rich source of evidence for these conceptual relationships (as it is generated by humans using them), LLMs can learn these roles without direct interaction with the world. However, this approach faces its own philosophical challenges. Block himself questioned whether conceptual role constitutes genuine content rather than merely a determinant of content, and the view struggles with well-known problems including radical meaning holism (where any revision to one part of the conceptual network threatens to alter all meanings within the system) and the difficulty of deriving truth conditions from purely syntactic roles. Most critically, it remains unclear whether the statistical correlations learned through next-token prediction genuinely instantiate the inferential and causal roles that Block took to be necessary for meaning, or whether they merely simulate those roles while lacking the requisite normative force. This concern is particularly acute when we note that Block's original formulation required mediation between sensory inputs and behavioral outputs, neither of which LLMs possess in any straightforward sense.

6 | Representation and World Models

Debates about semantic content in LLMs extend beyond their outputs to questions about their internal representations. Machine learning practitioners often use the term "representation" loosely for any activation pattern that correlates with an input feature or mediates between layers.[20] However, philosophers and cognitive scientists typically demand a more substantial notion that captures its unique explanatory role. Theories of representational content, especially those rooted in naturalistic traditions, have converged on two main conditions for content determination:[21]

1. **Information condition:** A representational vehicle (an internal state or activation pattern in a neural network) must reliably carry information about the feature of the environment it purports to represent. What a representation is about depends on what is causally upstream from it: a vehicle *R* represents *X* because instantiations of *X* reliably cause *R*.
2. **Use condition:** The information carried by the vehicle must play a role in guiding the system's downstream processing and behavior. What a representation is about depends on how it is exploited by the system: a vehicle *R* represents *X* because it drives computations and outputs that are appropriate for dealing with *X*. This implies that the vehicle must play a mechanistic and causal role; if *R* were altered, the system's behavior should change in an intelligible way related to its putative content.

A theory of representation based solely on simple correlation (information condition) or behavioral output (use condition) faces the challenge of explaining the possibility of error. A genuine representation must be capable of being incorrect. For example, if a frog's neural state is caused by flies but also, on occasion, by bees, a simple causal theory would imply the state's content is not FLY but the disjunction FLY-OR-BEE. In that case, the representation is never incorrect; it is simply a correct representation of a disjunctive property. The capacity for error, or normativity, is what distinguishes a representation from a mere causal effect or a reflex. Thus, any adequate theory of content must also provide the resources to ground correctness conditions. To address the problem of misrepresentation, naturalistic theories typically refine these conditions by appealing to function. Representational content is determined not by everything that can cause a state or how it *can* be used, but by what it is supposed to be caused by and used for. This function can be established through a history of selection—such as biological evolution or, crucially for LLMs, a learning process—which has selected the vehicle specifically for its role in processing information about a certain feature to successfully perform a task.[22] Misrepresentation then occurs when a state is caused by something other than what it has the function to detect (a malfunction of information-gathering) or is used in a way that is inappropriate given its content (a malfunction of use). For some artificial systems, this function might also be assigned by the intentions of their designers.

These criteria can be operationalized to make representational claims about LLMs empirically tractable (Harding 2023). For an activation pattern in a given layer to count as a representation of some feature, it must satisfy the conditions outlined above. To illustrate the methodology concretely, consider the hypothesis that a language model represents the grammatical number (singular or plural) of the subject noun phrase when predicting subject–verb agreement. To test the information condition, we must show that the activation pattern contains decodable information about the target feature. This is typically done by training a simple classifier, known as a probe, to predict the feature from the model's activations. In our example, we would train a probe to predict whether the subject is singular or plural from the activation pattern at a specific layer, using sentences where the ground-truth number is known. High prediction accuracy on a held-out test set suggests the activation pattern carries generalizable information about grammatical number, making it a candidate representation of that feature.

To test the use condition, we must show that the model causally relies on this information for downstream processing. Satisfying the information condition alone is insufficient: a model might encode information that is merely epiphenomenal, playing no role in generating outputs. Causal reliance can be tested through targeted interventions on the activation pattern—for instance, by ablating or corrupting the information about grammatical number while leaving other information intact. In our example, if selectively removing number information causes the model to produce verbs that no longer agree with the subject—whereas it performed correctly before the intervention—we have evidence that the model was causally exploiting the number information to generate appropriate verb forms.

Together, these tests establish that an activation pattern (i) carries decodable information about a feature and (ii) is causally implicated in the model's behavior. However, they do not by themselves establish that the activation pattern possesses the normative properties characteristic of genuine representation—namely, the capacity to be correct or incorrect about that feature. As noted above, this normative dimension must be grounded in a theory that specifies what the representation has the *function* of being about, typically by appeal to a history of selection. Without such a theory, we cannot distinguish a genuine representation (which can misrepresent) from a mere causal mediator (which cannot). Once a metasemantic theory supplies a correctness standard and a candidate content-ascription, however, a further diagnostic becomes available.[23] We can assess whether the model's errors are systematically associated with the representational vehicle taking a value that corresponds to an incorrect feature attribution. Returning to our example: suppose that, on a sentence with a singular subject, the model erroneously predicts a plural verb. If we hypothesize that this failure is due to misrepresentation of grammatical number, we can test this by examining whether the probe classifies the activation pattern as encoding PLURAL rather than SINGULAR. If so, we can further intervene on the activation pattern to shift it toward values the probe would classify as SINGULAR, while leaving other encoded information intact. If this targeted correction causes the model to now predict a singular verb, we have defeasible evidence that the original error was indeed a case of misrepresentation: the model's internal state carried incorrect information about grammatical number, and this incorrect information caused the erroneous output.

As discussed in Section 3, there is converging evidence from mechanistic interpretability research using probes and causal interventions as described above that LLMs represent syntactic features (Millière 2026). This is consistent with the hypothesis that next-token prediction, as a learning objective, creates a strong selection pressure on the internal states of LLMs to track the syntactic roles of tokens in order to make correct predictions. However, this is a purely intra-linguistic function; there is a further substantive question about whether LLMs can acquire representations about the world, that is, representations about extra-linguistic entities, properties, and events. To achieve this, LLMs would need not only to stand in appropriate causal-informational relations to the world but also to have a history of selection that has endowed them with the function of carrying this information. As we have seen in Section 5, the

causal-informational link, though indirect, is arguably established through the training corpus, which is itself causally shaped by human interaction with the world. However, on the face of it, next-token prediction is not a suitable learning objective to endow the internal states of LLMs with the requisite world-involving function (Butlin 2021). Nonetheless, Mollo and Millière (2025) argue that such function can be selected for through the *post-training* history of LLMs, and particularly during fine-tuning on human preferences, where internal states are selected based on their contribution to outputs that satisfy world-involving norms such as factual accuracy.[24] This entails, perhaps counter-intuitively, that multimodality and embodiment are neither necessary nor sufficient for an LLM to acquire representations about extra-linguistic reality; the crucial factor is a learning history that selects for world-representing representational functions.

A further substantive—if somewhat nebulous—question is whether LLMs can acquire "world models," a term used with considerable variation in AI and cognitive science. Building on Yildirim and Paul (2023), we can define a rather minimalist notion of a world model as a structure-preserving, behaviorally efficacious representation of the entities, relations, and/or processes in a system's environment. The "structure-preserving" condition is best understood through the lens of structural correspondence theories of representation (Shea 2018). On this view, a system of internal states represents a domain by instantiating a set of relations (e.g., geometric relations in an activation space) that mirrors the relations among entities in that domain. However, a mere structural correspondence is insufficient to ground content, as such correspondences are ubiquitous and can be found trivially. The correspondence must be exploited by the system, which is what the "behaviorally efficacious" condition captures: for the correspondence to be content-grounding, the system's success at its task must depend on it.

Perhaps the most compelling evidence that LLMs could in principle meet this minimal definition comes from models trained in formally constrained domains. In a now-classic study, a transformer model (Othello-GPT) was trained exclusively on transcripts of moves from the board game Othello (K. Li et al. 2023). Subsequent interpretability work revealed that the geometry of the model's activation space could be mapped onto the game board's structure, such that linear probes could read the board state from the model's activations, and causal interventions provided evidence of exploitation (Nanda et al. 2023). Intervening on the model's internal activations to "flip" the representation of a board square caused the model's subsequent move prediction to change in a way that was legally consistent with that counterfactual board state. This suggests both that the model's processing is causally sensitive to the internal relational structure and that its success (making legal moves) depends on the integrity of the correspondence between its internal map and the board. Othello-GPT thus appears to be a paradigmatic case of a system that has acquired a "world model" in this minimal sense by learning to exploit a structural correspondence. Whether actual LLMs can also exploit structural correspondences to real-world structures is more controversial (Mollo and Millière 2025; Williams 2025).

However, this definition leaves open at least two further questions about the stringency of the criteria. First, should the representations constituting the world model be globally coherent and systematic? A globally coherent model would not merely contain representations of individual states or relations but would represent the underlying structure of the domain in a way that respects its global constraints and equivalences. Given that the "world" of Othello is a closed, formal system with deterministic rules, Othello-GPT might be thought to be a prime candidate for satisfying this stronger condition. However, further interpretability research suggests that Othello-GPT does not implement a single, succinct, and globally consistent algorithm for representing the board state. Instead, it appears to have learned a "bag of heuristics": a large collection of independent, localized, and sometimes conflicting rules that are aggregated to produce a final, statistically reliable prediction (jylin04 et al. 2024; Vafa et al. 2024).

Second, what kind of behavioral efficacy is required? A stronger condition, often implicit in cognitive science, is that the relevant representations must support model-based reasoning and planning (Wong et al. 2023). On this view, a world model is not merely a static map of the world that guides immediate, reactive responses. Rather, it is a dynamic resource that can be manipulated "offline" to simulate possible future states, infer unobserved causes, and evaluate counterfactuals. It is unclear whether Othello-GPT meets this more demanding standard. Although its internal representations of the board state do guide its next action, this could be interpreted as a sophisticated learned policy rather than genuine model-based planning. There is no evidence that the model uses its internal board representations to "play out" multiple future move sequences to find an optimal path to victory.[25]

7 | Propositional Attitudes

Moving one step further in the comparison with human cognition, we can ask whether LLMs have propositional attitudes like beliefs or desires. Here too we encounter a familiar spectrum of positions, from eliminativism to realism. The default skeptical view, echoing critiques from Bender et al. (2021), is that ascribing such attitudes to LLMs is a category error rooted in anthropomorphism. A slightly weaker form of this skepticism might concede a fictionalist stance: pretending that LLMs have attitudes may be a useful fiction for facilitating fluid conversation but this does not commit us to their real existence.

However, as we have seen in the previous section it can be argued that LLMs do have the capacity to represent extra-linguistic reality under certain assumptions about their learning history. A cautious representationalist approach, advanced by Goldstein and Levinstein (2024), argues that while evidence from mechanistic interpretability does provide strong reasons to think LLMs possess robust internal states that satisfy the main conditions for representation, it is debatable whether these representations constitute full-fledged beliefs within a stable folk-psychological framework of belief-desire-action. In particular, the instability of LLMs' outputs across different prompts challenges the hypothesis that they have stable belief-like states. Chalmers (2025) proposes a middle path, advocating for a program of "propositional interpretability" that seeks to identify "generalized propositional attitudes," abstracting away from specific properties of belief and desire in human cognition. This approach aims to be

explanatorily useful without getting mired in debates over whether LLMs have minds in the way humans do.

On a more realist approach adjacent to interpretationism, if an LLM's coherent, seemingly rational, and goal-directed linguistic behavior is best explained by ascribing psychological attitudes to it, we are warranted in making such ascriptions without turning to mechanistic evidence. Cappelen and Dever (2025) advance a particularly strong version of this view. They argue that our everyday, reliable practice of identifying mental acts like answering questions should be applied to LLMs just as it is to humans. Because our interactions consistently and literally lead us to say LLMs perform these acts, we have a strong initial justification for these ascriptions. This justification then extends, through a commitment to cognitive holism, to the entire network of mental states that such acts presuppose; one cannot rationally separate the act of answering from the underlying beliefs and intentions that make it a coherent action, meaning that our justification for the former is also a justification for the latter.

There is a further debate about whether and how we could reliably detect attitudes such as beliefs in LLMs, if they do have them. Herrmann and Levinstein (2024) propose four conditions that any internal state must meet to be considered a belief-like representation: accuracy (it should track truth), coherence (it should be logically and semantically consistent), uniformity (the same decoding method should work across different domains), and use (the LLM must actually employ the representation to guide its outputs). The authors argue that these criteria, taken together, provide a rigorous standard for the measurement of beliefs, but current methods fail to meet this standard. Levinstein and Herrmann (2024) show empirically that popular probing methods are brittle and fail to generalize. Probes trained to identify a truth-representation for a set of statements perform worse than chance when tested on the negations of those same statements, suggesting they have learned a superficial correlate of truth (e.g., "is a simple, positive sentence found on Wikipedia") rather than a robust, general representation of truth itself. Keeling and Street (2025) extend this epistemic skepticism to the attribution of credences (or degrees of belief). They argue that all three major methods for measuring LLM confidence are unreliable. Prompting for reported confidence is confounded by the stochastic nature of text generation, while consistency-based estimation across multiple trials is distorted by user-controlled parameters such as temperature and sampling methods. Finally, deriving credences from output probabilities faces an intractable "bridging problem" between the probability of syntactic tokens and the credence in semantic propositions.

8 | Reasoning

Leading LLMs are often described as sophisticated reasoners; but can they actually reason? This debate is characteristically muddled by potential verbal disputes about the definition of "reasoning". Some philosophical accounts of reasoning are quite demanding, treating it as a personal-level process of attitude revision in which a subject actively changes her mind—forming, revising, or relinquishing beliefs or intentions—on the basis of other attitudes she holds as reasons. A transition between attitudes counts as reasoning only if the agent appreciates the rational support relation between them. Boghossian (2014) captures this with his influential "Taking Condition," which posits that for a transition to count as an inference, the reasoner must take her premises to justify her conclusion.[26] Some accounts tie the capacity for reasoning to other sophisticated mental phenomena such as propositional attitudes, personal-level aims, and explicit metacognition. As we have seen in Section 7, it is controversial that LLMs have propositional attitudes at all, let alone human-like beliefs and intentions; if they do not, it straightforwardly follows from this kind of account that LLMs cannot reason.[27]

However, many philosophers and cognitive scientists reject such accounts as over-intellectualized, insofar as they exclude the vast swaths of reasoning that may occur unconsciously or in nonhuman animals. This motivates naturalistic approaches that characterize reasoning in functional or computational terms. The challenge is to specify what makes a transition between representations inferential rather than merely causal or associative, by drawing on further distinctions between the kinds of computations available to cognitive systems.[28] Shea (2023) offers a general distinction between content-specific computations, whose validity depends on the particular contents of their inputs and outputs, and noncontent-specific computations that use formal procedures or algorithms invariant across the particular contents of their inputs and outputs. Content-specificity is characteristic of purely associative transitions, where a specific input pattern is mapped to a specific output based on learned statistical correlations. A classic example is a neural network trained to map a specific distribution of pixel values to the label DOG, or an LLM that learns a specific disposition to generate the token "cat" after the sequence "Curiosity killed the". In contrast, noncontent specific computations are the hallmark of classical symbolic programs, and provide the basis for the flexibility, generality, and systematicity often associated with reasoning. The rule of modus ponens, for instance, is applied in the same way regardless of what its premises are about.

Quilty-Dunn and Mandelbaum (2018) draw a related distinction between merely associative transitions and bare inferential transitions (BITs). A BIT is a noncontent-specific computation that meets two further constraints: first, its inputs must be "discursive" representations with a language-like constituent structure, and second, the rule governing the transition must be a "logical" rule geared toward truth-preservation. A BIT is thus the sub-personal realization of a formal inference. Although less demanding than the intellectualist account, this standard is still stringent, requiring not just content-independent processing but a specific, quasi-sentential representational format and sensitivity to logical relations. At the most inclusive end, some theories establish a "lower bound" for inference designed to accommodate flexible cognition in nonlinguistic animals (Buckner 2019b). On such views, reasoning does not require representations to have a language-like, propositional format, nor must the rules be strictly logical. Instead, practical inference can be realized through similarity-based categorization over rich, configural representations (like mental maps or models), guided by ecologically rational heuristics rather than formal logic. For an LLM to reason in this minimal sense, its internal state transitions would need to be guided by generalizable structure-sensitive procedures operating over its vector

embeddings, rather than just activating memorized input-output patterns.

Evidence on the validity and robustness of reasoning-like behavior in LLMs is somewhat mixed. On the one hand, a wealth of positive findings shows that the best LLMs can match or surpass human performance on complex reasoning problems. For example, LLMs are increasingly proficient at solving advanced competition-level mathematical problems that require valid step-by-step proofs (Luong and Lockhart 2025). They have also been found to achieve near-perfect performance and/or match humans at various logical reasoning problems (Liu et al. 2023), commonsense reasoning problems (Talmor et al. 2021), formal and semantically laden analogical reasoning tasks (Musker et al. 2025; Webb et al. 2023), and Theory of Mind tasks (Kosinski 2024; Strachan et al. 2024). On the other hand, LLMs still exhibit striking failure modes on deceptively simple puzzles. For example, the same LLMs that solve advanced math problems can still fail at variations on trick questions such as "What is heavier: a kilogram of metal or a kilogram of feathers?" (Zečević et al. 2023). This motivates an adversarial strategy using "counterfactual tasks": trivial variations on problems that are easy for humans but cause LLM performance to collapse (Lewis and Mitchell 2024; McCoy et al. 2024; Philip and Hemang 2024; Ullman 2023; Z. Wu, Manning, et al. 2024). These failures are often interpreted as symptomatic of a deeper gap between performance and competence; namely, LLMs are overly sensitive to irrelevant task features that reveal their tendency to memorize common content-specific patterns in their training data. If all LLMs do to solve reasoning problems is a form of shallow interpolation between nearest memorized training data points, instead of rule-governed inferential transitions, calling it "reasoning" seems less appropriate.

A common line of response to instances of reasoning errors in LLMs is that humans are susceptible to similar failure modes, and yet are deemed capable of reasoning. In fact, a well-documented phenomenon in psychology is that human reasoning is also subject to so-called "content effects," where performance on logical tasks is significantly influenced by the semantic content of the problem rather than by its formal structure alone (Evans et al. 1983). For instance, humans are more likely to endorse a logically invalid syllogism if its conclusion is believable, and they find it easier to solve the Wason selection task when it is framed in familiar, realistic terms (e.g., checking for underage drinkers) than in abstract ones (e.g., letters and numbers). Human cognition appears to use concepts that can support both content-specific pattern recognition and rule-based, noncontent-specific inference (Shea 2024). Classic examples of "content effects" on reasoning can be seen as cases where content-specific associations are either facilitating or interfering with noncontent-specific transitions. Lampinen et al. (2024) find that LLMs exhibit precisely these kinds of human-like content effects. Across tasks like natural language inference, syllogistic reasoning, and the Wason selection task, they show that LLMs, like humans, are more accurate when semantic content aligns with logical validity and are similarly biased when the two conflict. Furthermore, LLM confidence on these tasks correlates with human response times, suggesting that the problems humans find more difficult are also those that elicit lower confidence from models.

To arbitrate disputes about whether LLMs can engage in rule-governed, noncontent-specific inferential transitions, we ought to look beyond behavior. Recent work in mechanistic interpretability provides compelling evidence that transformers, when trained on appropriate formal tasks, can indeed learn to implement noncontent-specific computations that approximate, if not implement, BITs. For example, Y. Wu et al. (2025) show that a transformer trained to dereference variables in symbolic programs develops a systematic, multi-step mechanism for tracking assignment chains. Using causal interventions, they show that the model learns to use its vector space as a kind of addressable memory, with specialized attention heads routing information between token positions in a way that is sensitive to the program's structure but independent of the specific variables or values involved. This learned algorithm for variable dereferencing is a clear case of noncontent-specific computation. Furthermore, the developmental trajectory they uncover—from shallow, content-bound heuristics to a general, content-independent mechanism—suggests a process of learning that progressively abstracts away from content-specific association. Whether this learned mechanism constitutes a genuine BIT remains an open question, contingent on whether the model's internal representations possess the requisite discursive format and its operations the counterfactual robustness of a built-in logical rule. Nonetheless, these findings strongly suggest that the architectural biases of transformers are sufficient to induce the emergence of structure-sensitive, noncontent-specific computational processes from statistical learning alone. Whether this finding translates to actual LLMs trained on text is complicated. For example, evidence suggests that LLMs can form a general, content-independent circuit to solve reasoning tasks such as syllogisms, but this circuit can be overridden by conflicting world knowledge encoded in other parts of the model, particularly in cases of belief-inconsistent reasoning (G. Kim et al. 2025).

Interestingly, the most advanced LLMs that exist today undergo a specific posttraining stage that teaches them to produce a "reasoning trace," also known as a "chain of thought" (CoT), before answering user queries (DeepSeek-AI et al. 2025; OpenAI et al. 2024). These so-called "large reasoning models" (LRMs) consistently outperform regular LLMs on reasoning tasks, and are often described as implementing a kind of slow, deliberate, noncontent-specific reasoning process as opposed to fast associative transitions (Z. Z. Li et al. 2025). Generating a CoT enables these models to solve a larger class of problems, including inherently serial multi-step problems that cannot be solved in a single forward pass (Z. Li et al. 2023; Merrill and Sabharwal 2024). However, generating "reasoning traces" may not be reliable evidence for bona fide reasoning (Kambhampati, Stechly, and Valmeekam 2025). When human reasoning is mediated by external reasoning traces, for example, by writing down each step of a proof on a scratchpad, the output at one step is a key part of the input at the next step. This is not necessarily the case with LRMs; in fact, there is evidence that the semantic content and logical validity of the generated trace can be surprisingly uncorrelated with the final answer's correctness, and that models can even achieve superior performance when trained on traces that are algorithmically invalid or nonsensical (Kambhampati, Stechly, Valmeekam, et al. 2025; Stechly et al. 2025).

Ultimately, then, whether current LLMs can reason depends on both background assumptions and controversial evidence. On normatively rich views that require personal-level attitudes and endorsement, the case for LLM reasoning remains somewhat tenuous. On more naturalistic accounts, which ground reasoning in regimented inferential transitions between structured representations, there is increasingly suggestive behavioral and mechanistic evidence that transformer-based LLMs can meet this requirement in principle, if not in practice. As with other philosophical questions about LLMs, what looks like a substantive disagreement may mask a deeper ambiguity in our explanatory targets, and resolving it requires both conceptual clarity and empirical work.

9 | Agency and Consciousness

A final set of questions about the psychological capacities of LLM-based systems concerns whether they can be meaningfully ascribed some form of agency and even consciousness. These questions have gained increasing attention following the development of so-called "agents"—modular systems that integrate LLMs with external tools, databases, and symbolic scaffolding.[29] Butlin (2025) argues that a system is an agent if it is sensitive to the "instrumental value" of its outputs, meaning it can learn to select actions that help achieve a goal over multiple steps. Many LLM-based "agents," particularly those fine-tuned based on human preferences or equipped with additional modules such as memory and reflection mechanisms, arguably meet this standard: they systematically modify their behavior based on feedback to better achieve goals. Butlin further distinguishes agency from having desires, which require not only pursuing goals but also using them in practical reasoning and acquiring a coherent, integrated set of goals through learning. Most LLM-based "agents" have their goals specified by designers and do not develop them through a process that would create a unified set of personal values, thus falling short of having desires in a more substantive, human-like sense. Similarly, Dung (2024) proposes that agency is not a binary property but a multidimensional construct. Although LLMs score low on dimensions like autonomy (requiring external prompts) and efficacy (lacking direct physical interaction), they exhibit goal-directedness in pursuing objectives such as accurate text prediction or reward maximization. Overall, their agency remains significantly limited and distinct from that of humans—they are more like passive tools than self-initiating actors that can directly affect the world.

Even if LLM-based systems possess some limited form of agency, this does not entail that they are conscious. A survey of US adults found that 67% already attributed some degree of phenomenal consciousness to ChatGPT, with these attributions being positively correlated with usage frequency (Colombatto and Fleming 2024). However, among philosophers and cognitive scientists, there is an emerging consensus that current LLMs are unlikely to be conscious. A first, in-principle objection to LLM consciousness is motivated by the view that consciousness is not a substrate-independent computational phenomenon amenable to multiple realization (Cao 2022; Godfrey-Smith 2016; Seth 2025). This position, sometimes called biological naturalism, holds that consciousness depends on the specific, fine-grained functional properties of living, metabolic systems. Proponents argue that evolution has generatively entrenched cognitive functions with their physiological basis, multiplexing informational, metabolic, and self-maintaining processes in ways that are nomologically difficult, if not impossible, to replicate in a nonbiological substrate like a silicon chip.

Even among philosophers who believe consciousness could theoretically exist in artificial systems, there is broad consensus that LLMs based on the transformer architecture lack the critical features required by leading theories of consciousness (Butlin et al. 2023; Chalmers 2024). These missing components include recurrent processing (the feedback loops essential for maintaining internal states) and a "global workspace" for integrating information from different modules. This is compatible with the view that future AI systems could plausibly become conscious because these limitations are contingent. For example, Chalmers (2024) proposes that "LLM+" systems could be built within a decade to overcome current obstacles by incorporating senses and embodiment, recurrent processing, global workspaces, and unified agency. Similarly, Butlin et al. (2023) conclude that while no current AI is a strong candidate for consciousness, there are no obvious technical barriers to designing a system that satisfies the indicators derived from consciousness science. In summary, even if current LLMs are poor candidates, a path toward conscious AI is often considered technologically feasible, either by augmenting current architectures or by developing new biologically inspired ones.[30]

10 | Conclusion

The advent of LLMs marks a watershed moment for the philosophy of language, mind, and cognitive science. These systems compel us to revisit fundamental questions about linguistic competence, meaning, representation, reasoning, agency, and consciousness in a new light. Rather than settling these debates, LLMs extend and sometimes complicate them: they challenge long-held assumptions about what statistical learning can achieve, blur traditional boundaries between performance and competence, and demand new frameworks for thinking about semantic content and cognitive processes in artificial systems. Philosophy—alongside computer science, linguistics, and psychology—has an important role to play in advancing these issues by clarifying explanatory targets and making explicit our background theoretical commitments. However, many downstream questions about the capacities and limitations of LLMs also require serious engagement with empirical evidence from behavioral experiments and mechanistic interpretability research. We hope this opinionated review will encourage philosophers to engage with such evidence while motivating non-philosophers to engage with philosophical theory.

Acknowledgments

The authors have nothing to report.

Funding

Raphaël Millière is funded by an AI2050 Early Career Fellowship from Schmidt Sciences.

Ethics Statement

The authors have nothing to report.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Endnotes

[1] For up-to-date data on the best performing LLMs on popular benchmarks, see https://epoch.ai/data/ai-benchmarking-dashboard.

[2] For a philosophical introduction to neural networks, see Buckner (2019a). For a brief overview of language models, see Mitchell (2024). For visual introductions, see Sanderson (2017, 2024).

[3] The Transformer architecture was introduced by Vaswani et al. (2017); for an accessible introduction, see Millière (2025).

[4] This terminology is misleading, as these are phases of the training process.

[5] See Grzankowski et al. (2025) for a discussion of the redescription fallacy that draws on an analogy with natural selection.

[6] Bender et al. (2021) famously described LLMs as "stochastic parrots," though their core concern is about linguistic meaning rather than rote memorization (see Section 5).

[7] See Chomsky (1965) for the classic distinction and Firestone (2020) for application across biological and artificial systems.

[8] See Hu and Frank (2024) on the influence of auxiliary task demands and Millière and Rathkopf (2025) for a broader defense of a bidirectional dissociation between performance and competence in LLMs.

[9] For a detailed discussion, see Futrell and Mahowald (2025) and Millière (2026). See Dupre (2021) for a defense of the skeptical position, and Piantadosi (2024) for a radical stance against it.

[10] See Belinkov (2022) for a review of probing methods and Harding (2023) for a discussion of their role in arbitrating claims about linguistic representation in language models.

[11] An algorithm is an abstract, medium-independent specification of representations and procedures; the same algorithm can, in principle, be implemented by different mechanisms. What it means for a circuit in a neural network to implement an algorithm can be made precise using the framework of causal abstraction (Geiger et al. 2025); roughly, a circuit implements an algorithm when the circuit's internal states can be mapped onto the algorithm's variables in a way that preserves counterfactual structure—that is, interventions that change a variable in the algorithm correspond to interventions on the circuit that produce the expected downstream effects. In practice, task-relevant circuits in neural networks rarely align exactly with simple, human-interpretable algorithms; instead, the typical finding is approximate alignment—the circuit's states correspond imperfectly to algorithmic variables, or the circuit implements a computation that resists characterization in simple terms (such as an aggregation of heuristics rather than a unified algorithm). Nonetheless, even approximate implementation claims can be explanatorily valuable when they support predictions about the network's behavior under novel inputs and interventions.

[12] It remains an open question how much of the more abstract theoretical apparatus of syntax (such as binding principles, island constraints, or cross-linguistic parametric variation) is learned by language models; see Futrell and Mahowald (2025) for discussion.

[13] See Millière (2026) and Lan et al. (2024) for a more detailed discussion.

[14] See Cowie (1998), Laurence and Margolis (2001), and Pearl (2022) for discussion of the POS argument.

[15] For discussion, see McGrath et al. (2023), Millière (2024), Pavlick (2023), Quilty-Dunn et al. (2023), and Russin et al. (2024).

[16] A semantic theory specifies *what* the meanings of expressions are, whereas a metasemantic theory explains the foundational facts about language users and the world *in virtue of which* those expressions have those meanings.

[17] This argument relies on the additional assumption that LLMs' outputs token the same word types as the outputs of the linguistic community whose causal-historical connections they inherit; since LLMs only process and generate tokenized text that doesn't map neatly onto words, this assumption is potentially controversial. See Stojnić (2022) for discussion of what determines which words are tokened in an utterance.

[18] Although he is not concerned with the Kripkean objection to the externalist approach, Attah (2025) argues more broadly against the claim that LLMs lack communicative intentions. He contends that the architecture of LLMs can in fact support communicative intentions on a functionalist account that steers clear of overly demanding Gricean assumptions.

[19] Another way to extend bibliotechnism is to embrace fictionalism, biting the bullet that the outputs of LLMs are quite literally meaningless, even though they can be fictionally meaningful to us through a form of prop-oriented make-believe that allows for rational epistemic engagement (Mallory 2023).

[20] On the opposite end of the spectrum, some authors contend that genuine representation is a very demanding psychological capacity, characterized by its "offline" or stimulus-decoupled use in conscious, deliberate cognitive processes such as planning and imagination (Krakauer 2025). On this view, applying the term "representation" to the internal states of systems like LLMs is a nonstarter. Throughout this section, we follow mainstream naturalistic accounts of representation in assuming that questions about whether LLMs have representational content—and, if so, what content they have—are in fact meaningful empirical questions. See Favela and Machery (2025) and Cao (2025) for a discussion of the multiple (and conflicting) ways in which researchers in different fields use the notion of representation.

[21] As with linguistic expressions, we can distinguish between psychosemantic theories which specify *what* the contents of mental representations are, and meta-psychosemantic theories which explain the foundational facts about a system and its environment *in virtue of which* those representations have the specific contents that they do. See Dretske (1981), Millikan (1984), Neander (2017), and Shea (2018) for discussion from major proponents of naturalistic theories of representation.

[22] Teleosemantic theories that appeal to selection history are not the only naturalistic approach to grounding representational content. For example, Buckner (2021) proposes a "forward-looking" theory in which the content of a representation is determined by what the agent's own error-correction mechanisms are disposed to make it a better indicator of over time. On this view, a representation's content is what it is being shaped to better indicate through predictive learning and self-correction.

[23] This diagnostic corresponds to what Harding (2023) calls the "misrepresentation condition." However, as Chalmers (2025) notes, this is less a principle for determining mental content and more a condition of adequacy on a psychosemantic theory. We therefore characterize it as a diagnostic that supports a representational explanation conditional on an independently motivated normative framework, rather than as an intrinsic test for normativity.

[24] Mollo and Millière (2025) also argue, more speculatively, that mere pre-training in formally constrained domains such as board games can, under specific conditions, provide the requisite selection history.

[25] Though see Jenner et al. (2024) for evidence that a Transformer model trained on chess learns a mechanism to *look ahead* several moves in advance, and see Men et al. (2024) and Dong et al. (2025) for evidence that real-world LLMs learn similar look-ahead mechanisms to plan moves in text-based games and to anticipate future tokens.

[26] Many philosophers use the terms "reasoning" and "inference" interchangeably, though some reserve the former for the more demanding account outlined in this paragraph.

[27] Some philosophers also argue that we have other a priori reasons to deny thinking or reasoning to LLMs. For example, Stoljar and Zhang (2025) claims that LLMs cannot think about extra-linguistic reality because they can at best infer conclusions about worldly facts (e.g., about bananas) from premises that are exclusively about statistical properties of language (e.g., the probability of the word "bananas"). Since such inferences are evidentially unsound, they allegedly violate the constitutive rationality of thought. This leaves open the possibility that LLMs think, but only about *words*. However, as discussed in Section 6, it has been argued that LLMs can acquire representations with extra-linguistic content, in which case this argument from irrationality might be less compelling.

[28] For a discussion of associative transitions and their relevance to machine learning, see Mandelbaum and Millière (2025).

[29] For an introduction to LLM-based "agent" systems, see Millière and Buckner (2024), Section 3.1.2.

[30] This does not mean, of course, that developing conscious AI is *desirable*.

References

Ahuja, K., V. Balachandran, M. Panwar, et al. 2025. "Learning Syntax Without Planting Trees: Understanding Hierarchical Generalization in Transformers." *Transactions of the Association for Computational Linguistics* 13: 121–141. https://doi.org/10.1162/tacl_a_00733.

Attah, N. O. 2025. "Do Language Models Lack Communicative Intentions?" *Synthese* 205, no. 5: 187. https://doi.org/10.1007/s11229-025-05022-6.

Belinkov, Y. 2022. "Probing Classifiers: Promises, Shortcomings, and Advances." *Computational Linguistics* 48, no. 1: 207–219. https://doi.org/10.1162/coli_a_00422.

Bender, E. M., T. Gebru, A. McMillan-Major, and S. Shmitchell. 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. FAccT '21*, 610–623. Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922.

Bender, E. M., and A. Koller. 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*, 5185–5198. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463.

Block, N. 1981. "Psychologism and Behaviorism." *Philosophical Review* 90, no. 1: 5–43. https://doi.org/10.2307/2184371.

Block, N. 1986. "Advertisement for a Semantics for Psychology." *Midwest Studies in Philosophy* 10: 615–678. https://doi.org/10.1111/j.1475-4975.1987.tb00558.x.

Boghossian, P. 2014. "What Is Inference?" *Philosophical Studies* 169, no. 1: 1–18. https://doi.org/10.1007/s11098-012-9903-x.

Boguraev, S., C. Potts, and K. Mahowald. 2025. Causal Interventions Reveal Shared Structure Across English Filler-Gap Constructions: arXiv: 2505.16002. https://doi.org/10.48550/arXiv.2505.16002.

Borg, E. 2025. "LLMs, Turing Tests and Chinese Rooms: The Prospects for Meaning in Large Language Models." *Inquiry*: 1–31. https://doi.org/10.1080/0020174X.2024.2446241.

Bubeck, S., V. Chandrasekaran, R. Eldan, et al. 2023. Sparks of Artificial General Intelligence: Early Experiments with GPT-4: arXiv: 2303.12712. https://doi.org/10.48550/arXiv.2303.12712.

Buckner, C. 2019a. "Deep Learning: A Philosophical Introduction." *Philosophy Compass* 14, no. 10: e12625. https://doi.org/10.1111/phc3.12625.

Buckner, C. 2019b. "Rational Inference: The Lowest Bounds." *Philosophy and Phenomenological Research* 98, no. 3: 697–724. https://doi.org/10.1111/phpr.12455.

Buckner, C. 2021. "A Forward-Looking Theory of Content." *Ergo an Open Access Journal of Philosophy* 8. https://doi.org/10.3998/ergo.2238.

Butlin, P. 2021. "Sharing Our Concepts With Machines." *Erkenntnis* 88, no. 7: 3079–3095. https://doi.org/10.1007/s10670-021-00491-w.

Butlin, P. 2025. "The Agency in Language Agents." *Inquiry*: 1–21. https://doi.org/10.1080/0020174X.2024.2439995.

Butlin, P., R. Long, E. Elmoznino, et al. 2023. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness: arXiv: 2308.08708. https://doi.org/10.48550/arXiv.2308.08708.

Cao, R. 2022. "Multiple Realizability and the Spirit of Functionalism." *Synthese* 200, no. 6: 506. https://doi.org/10.1007/s11229-022-03524-1.

Cao, R. 2025. "Assessing the Landscape of Representational Concepts: Commentary on Favela and Machery." *Mind & Language* 40, no. 2: 226–232. https://doi.org/10.1111/mila.12535.

Cappelen, H., and J. Dever. 2025. Going Whole Hog: A Philosophical Defense of AI Cognition: arXiv:2504.13988. https://doi.org/10.48550/arXiv.2504.13988.

Chalmers, D. J. 2024. Could a Large Language Model Be Conscious?: arXiv:2303.07103. https://doi.org/10.48550/arXiv.2303.07103.

Chalmers, D. J. 2025. Propositional Interpretability in Artificial Intelligence: arXiv: 2501.15740. https://doi.org/10.48550/arXiv.2501.15740.

Chomsky, N. 1957. *Syntactic Structures*. Mouton.

Chomsky, N. 1965. *Aspects of the Theory of Syntax*. MIT Press.

Colombatto, C., and S. M. Fleming. 2024. "Folk Psychological Attributions of Consciousness to Large Language Models." *Neuroscience of Consciousness* 2024, no. 1: niae013. https://doi.org/10.1093/nc/niae013.

Cowie, F. 1998. *What's Within? Nativism Reconsidered*. Oxford University Press USA.

DeepSeek-AI, Guo, D., D. Yang, et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning: arXiv: 2501.12948. https://doi.org/10.48550/arXiv.2501.12948.

Descartes, R. 1985. "Discourse on the Method of Rightly Conducting One's Reason and Seeking Truth in the Sciences." In *The Philosophical Writings of Descartes*, edited by J. Cottingham, R. Stoothoff, and D. Murdoch, Vol. 1, 111–151. Cambridge University Press.

Donatelli, L., and A. Koller. 2023. "Compositionality in Computational Linguistics." *Annual Review of Linguistics* 9, no. 1: 463–481. https://doi.org/10.1146/annurev-linguistics-030521-044439.

Dong, Z., Z. Zhou, Z. Liu, C. Yang, and C. Lu. 2025. "Emergent Response Planning in LLMs." In *Forty-Second International Conference on Machine Learning*.

Dretske, F. 1981. *Knowledge and the Flow of Information*. MIT Press.

Dung, L. 2024. "Understanding Artificial Agency." *Philosophical Quarterly* 75, no. 2: pqae010–pqae472. https://doi.org/10.1093/pq/pqae010.

Dupre, G. 2021. "(What) Can Deep Learning Contribute to Theoretical Linguistics?" *Minds and Machines* 31, no. 4: 617–635. https://doi.org/10.1007/s11023-021-09571-w.

Evans, J. S. B. T., J. L. Barston, and P. Pollard. 1983. "On the Conflict Between Logic and Belief in Syllogistic Reasoning." *Memory & Cognition* 11, no. 3: 295–306. https://doi.org/10.3758/BF03196976.

Everaert, M. B. H., M. A. C. Huybregts, N. Chomsky, R. C. Berwick, and J. J. Bolhuis. 2015. "Structures, Not Strings: Linguistics as Part of the Cognitive Sciences." *Trends in Cognitive Sciences* 19, no. 12: 729–743. https://doi.org/10.1016/j.tics.2015.09.008.

Favela, L. H., and E. Machery. 2025. "The Concept of Representation in the Brain Sciences: The Current Status and Ways Forward." *Mind & Language* 40, no. 2: 215–225. https://doi.org/10.1111/mila.12531.

Feng, J., and J. Steinhardt. 2023. How Do Language Models Bind Entities in Context?: arXiv: 2310.17191. https://doi.org/10.48550/arXiv.2310.17191.

Firestone, C. 2020. "Performance vs. Competence in Human–Machine Comparisons." *Proceedings of the National Academy of Sciences* 117, no. 43: 26562–26571. https://doi.org/10.1073/pnas.1905334117.

Fodor, J. A. 1975. *The Language of Thought*. Harvard University Press.

Fodor, J. A., and Z. W. Pylyshyn. 1988. "Connectionism and Cognitive Architecture: A Critical Analysis." *Cognition* 28, no. 1: 3–71. https://doi.org/10.1016/0010-0277(88)90031-5.

Frank, M. C. 2023. "Baby Steps in Evaluating the Capacities of Large Language Models." *Nature Reviews Psychology* 2, no. 8: 451–452. https://doi.org/10.1038/s44159-023-00211-x.

Futrell, R., and K. Mahowald. 2025. How Linguistics Learned to Stop Worrying and Love the Language Models: arXiv: 2501.17047. https://doi.org/10.48550/arXiv.2501.17047.

Geiger, A., D. Ibeling, A. Zur, et al. 2025. Causal Abstraction: A Theoretical Foundation for Mechanistic Interpretability: arXiv: 2301.04709 [cs]. https://doi.org/10.48550/arXiv.2301.04709.

Godfrey-Smith, P. 2016. "Mind, Matter, and Metabolism." *Journal of Philosophy* 113, no. 10: 481–506. https://doi.org/10.5840/jphil20161131034.

Goldstein, S., and B. A. Levinstein. 2024. Does ChatGPT Have a Mind?: arXiv: 2407.11015. https://doi.org/10.48550/arXiv.2407.11015.

Goodale, M., and S. Mascarenhas. 2023. Fodor and Pylyshyn's Systematicity Challenge Still Stands: A Reply to Lake and Baroni (2023).

Grzankowski, A. 2024. "Real Sparks of Artificial Intelligence and the Importance of Inner Interpretability." *Inquiry*: 1–27. https://doi.org/10.1080/0020174X.2023.2296468.

Grzankowski, A., S. M. Downes, and P. Forber. 2025. "LLMs Are Not Just Next Token Predictors." *Inquiry*: 1–11. https://doi.org/10.1080/0020174X.2024.2446240.

Harding, J. 2023. "Operationalising Representation in Natural Language Processing." *British Journal for the Philosophy of Science*: 728685. https://doi.org/10.1086/728685.

Harnad, S. 1990. "The Symbol Grounding Problem." *Physica D: Nonlinear Phenomena* 42, no. 1: 335–346. https://doi.org/10.1016/0167-2789(90)90087-6.

Herrmann, D. A., and B. A. Levinstein. 2024. "Standards for Belief Representations in LLMs." *Minds and Machines* 35, no. 1: 5. https://doi.org/10.1007/s11023-024-09709-6.

Hu, J., and M. C. Frank. 2024. "Auxiliary Task Demands Mask the Capabilities of Smaller Language Models." In *First Conference on Language Modeling*.

Jenner, E., S. Kapur, V. Georgiev, C. Allen, S. Emmons, and S. Russell. 2024. "Evidence of Learned Look-Ahead in a Chess-Playing Neural Network." *Advances in Neural Information Processing Systems* 37: 31410–31437. https://doi.org/10.52202/079017-0987.

Jones, C. R., I. Rathi, S. Taylor, and B. K. Bergen. 2025. "People Cannot Distinguish GPT-4 From a Human in a Turing Test." In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency. FAccT '25*, 1615–1639. Association for Computing Machinery. https://doi.org/10.1145/3715275.3732108.

jylin04, JackS, A. Karvonen, and Can. 2024. OthelloGPT Learned a Bag of Heuristics.

Kambhampati, S., K. Stechly, and K. Valmeekam. 2025. "(How) Do Reasoning Models Reason?" *Annals of the New York Academy of Sciences* 1547, no. 1: 33–40. https://doi.org/10.1111/nyas.15339.

Kambhampati, S., K. Stechly, K. Valmeekam, et al. 2025. Stop Anthropomorphizing Intermediate Tokens as Reasoning/Thinking Traces!: arXiv: 2504.0976. https://doi.org/10.48550/arXiv.2504.097622.

Keeling, G., and W. Street. 2025. "On the Attribution of Confidence to Large Language Models." *Inquiry*: 1–27. https://doi.org/10.1080/0020174X.2025.2450598.

Keysers, D., N. Schärli, N. Scales, et al. 2019. "Measuring Compositional Generalization: A Comprehensive Method on Realistic Data." *International Conference on Learning Representations*. https://openreview.net/forum?id=SygcCnNKwr.

Kim, G., M. Valentino, and A. Freitas. 2025. A Mechanistic Interpretation of Syllogistic Reasoning in Auto-Regressive Language Models: arXiv: 2408.08590. https://doi.org/10.48550/arXiv.2408.08590.

Kim, N., and T. Linzen. 2020. "COGS: A Compositional Generalization Challenge Based on Semantic Interpretation." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9087–9105. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.731.

Koch, S. 2025. "Babbling Stochastic Parrots? A Kripkean Argument for Reference in Large Language Models." *Philosophy of AI* 1: 19–33. https://doi.org/10.18716/ojs/phai/2025.2325.

Kokotajlo, D., S. Alexander, T. Larsen, E. Lifland, and R. Dean. 2025. AI 2027. https://ai-2027.com/.

Kosinski, M. 2024. "Evaluating Large Language Models in Theory of Mind Tasks." *Proceedings of the National Academy of Sciences* 121, no. 45: e2405460121. https://doi.org/10.1073/pnas.2405460121.

Krakauer, J. W. 2025. "Where Did Real Representations Go? Commentary on: The Concept of Representation in the Brain Sciences: The Current Status and Ways Forward by Favela and Machery." *Mind & Language* 40, no. 2: 239–242. https://doi.org/10.1111/mila.12534.

Kripke, S. 1980. *Naming and Necessity*. Harvard University Press.

Lake, B. M., and M. Baroni. 2018. "Generalization Without Systematicity: On the Compositional Skills of Sequence-to-Sequence Recurrent Networks." In *Proceedings of the 35th International Conference on Machine Learning*, 2873–2882. PMLR.

Lake, B. M., and M. Baroni. 2023. "Human-Like Systematic Generalization Through a Meta-Learning Neural Network." *Nature* 623, no. 7985: 1–7. https://doi.org/10.1038/s41586-023-06668-3.

Lampinen, A. K., I. Dasgupta, S. C. Y. Chan, et al. 2024. "Language Models, Like Humans, Show Content Effects on Reasoning Tasks." *NAS Nexus* 3, no. 7: pgae233. https://doi.org/10.1093/pnasnexus/pgae233.

Lan, N., E. Chemla, and R. Katzir. 2024. "Large Language Models and the Argument From the Poverty of the Stimulus." *Linguistic Inquiry*: 1–28. https://doi.org/10.1162/ling_a_00533.

Lasri, K., T. Pimentel, A. Lenci, T. Poibeau, and R. Cotterell. 2022. "Probing for the Usage of Grammatical Number." Long Papers In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Vol. 1, 8818–8831. https://doi.org/10.18653/v1/2022.acl-long.603.

Laurence, S., and E. Margolis. 2001. "The Poverty of the Stimulus Argument." *British Journal for the Philosophy of Science* 52, no. 2: 217–276. https://doi.org/10.1093/bjps/52.2.217.

Lederman, H., and K. Mahowald. 2024. "Are Language Models More Like Libraries or Like Librarians? Bibliotechnism, the Novel Reference Problem, and the Attitudes of LLMs." *Transactions of the Association for Computational Linguistics* 12: 1087–1103. https://doi.org/10.1162/tacl_a_00690.

Leong, C. S.-Y., and T. Linzen. 2024. Testing Learning Hypotheses Using Neural Networks by Manipulating Learning Data: arXiv: 2407.04593. https://doi.org/10.48550/arXiv.2407.04593.

Levinstein, B. A., and D. A. Herrmann. 2024. "Still No Lie Detector for Language Models: Probing Empirical and Conceptual Roadblocks." *Philosophical Studies* 182, no. 7: 1539–1565. https://doi.org/10.1007/s11098-023-02094-3.

Lewis, M. A., and M. Mitchell. 2024. "Using Counterfactual Tasks to Evaluate the Generality of Analogical Reasoning in Large Language Models." *arXiv preprint arXiv:2402.08955*. https://arxiv.org/abs/2402.08955.

Li, K., A. K. Hopkins, D. Bau, F. B. Viégas, H. Pfister, and M. Wattenberg. 2023. "Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task." In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1–5, 2023*. OpenReview.net.

Li, Z., H. Liu, D. Zhou, and T. Ma. 2023. "Chain of Thought Empowers Transformers to Solve Inherently Serial Problems." In *The Twelfth International Conference on Learning Representations*.

Li, Z. Z., D. Zhang, M. L. Zhang, et al. 2025. "From System 1 to System 2: A Survey of Reasoning Large Language Models." *arXiv preprint arXiv: 2502.17419*. https://arxiv.org/abs/2502.17419.

Linzen, T., E. Dupoux, and Y. Goldberg. 2016. "Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies." *Transactions of the Association for Computational Linguistics* 4: 521–535. https://doi.org/10.1162/tacl_a_00115.

Liu, H., J. Liu, L. Cui, et al. 2023. "LogiQA 2.0—An Improved Dataset for Logical Reasoning in Natural Language Understanding." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31: 2947–2962. https://doi.org/10.1109/TASLP.2023.3293046.

Luong, T., and E. Lockhart. 2025. Advanced Version of Gemini with Deep Think Officially Achieves Gold-medal Standard at the International Mathematical Olympiad.

Machamer, P., L. Darden, and C. F. Craver. 2000. "Thinking About Mechanisms." *Philosophy of Science* 67, no. 1: 1–25. https://doi.org/10.1086/392759.

Mallory, F. 2023. "Fictionalism About Chatbots." *Ergo an Open Access Journal of Philosophy* 10. https://doi.org/10.3998/ergo.4668.

Mandelbaum, E. and R. Millière. 2025. "Associationist Theories of Thought." In *The Stanford Encyclopedia of Philosophy* Fall 2025, edited by E. N. Zalta and U. Nodelman. Metaphysics Research Lab, Stanford University.

Mandelkern, M., and T. Linzen. 2024. "Do Language Models' Words Refer?" In *Computational Linguistics*, 1–10. https://doi.org/10.1162/coli_a_00522.

Marks, S., C. Rager, E. J. Michaud, Y. Belinkov, D. Bau, and A. Mueller. 2024. "Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models." *arXiv* 2403: 19647. https://doi.org/10.48550/arXiv.2403.19647.

McCoy, R. T., S. Yao, D. Friedman, M. D. Hardy, and T. L. Griffiths. 2024. "Embers of Autoregression Show How Large Language Models Are Shaped by the Problem They Are Trained to Solve." *Proceedings of the National Academy of Sciences* 121, no. 41: e2322420121. https://doi.org/10.1073/pnas.2322420121.

McGrath, S., J. Russin, E. Pavlick, and R. Feiman. 2023. Properties of LoTs: The Footprints or the Bear Itself?. https://doi.org/10.31234/osf.io/tdw34.

Men, T., P. Cao, Z. Jin, Y. Chen, K. Liu, and J. Zhao. 2024. "Unlocking the Future: Exploring Look-Ahead Planning Mechanistic Interpretability in Large Language Models." In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, edited by Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, 7713–7724. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.440.

Merrill, W., and A. Sabharwal. 2024. The Expressive Power of Transformers with Chain of Thought: arXiv: 2310.07923. https://doi.org/10.48550/arXiv.2310.07923.

Millière, R. 2024. "Philosophy of Cognitive Science in the Age of Deep Learning." *WIREs Cognitive Science* 15, no. 5: e1684. https://doi.org/10.1002/wcs.1684.

Millière, R. 2025. "Transformers." *Open Encyclopedia of Cognitive Science*. https://doi.org/10.21428/e2759450.d3acfbfb.

Millière, R. 2026. "Language Models as Models of Language." In *The Oxford Handbook of the Philosophy of Linguistics*, edited by R. Nefdt, G. Dupre, and K. Stanton. Oxford University Press.

Millière, R., and C. Buckner. 2024. A Philosophical Introduction to Language Models – Part II: The Way Forward: arXiv: 2405.03207.

Millière, R., and C. Buckner. 2026. "Interventionist Methods for Interpreting Deep Neural Networks." In *Neurocognitive Foundations of Mind*, edited by G. Piccinini, 190–221. Routledge.

Millière, R., and C. Rathkopf. 2025. "Anthropocentric Bias in Language Model Evaluation." *Computational Linguistics*: 1–10. https://doi.org/10.1162/COLI.a.582.

Millikan, R. G. 1984. *Language, Thought, and Other Biological Categories: New Foundations for Realism*. MIT Press.

Mitchell, M. 2024. "Large Language Models." *Open Encyclopedia of Cognitive Science*. https://doi.org/10.21428/e2759450.2bb20e3c.

Mitchell, M., and D. C. Krakauer. 2023. "The Debate over Understanding in AI's Large Language Models." *Proceedings of the National Academy of Sciences* 120, no. 13: e2215907120. https://doi.org/10.1073/pnas.2215907120.

Mollo, D. C., and R. Millière. 2025. The Vector Grounding Problem: arXiv: 2304.01481. https://doi.org/10.48550/arXiv.2304.01481.

Morris, J. X., C. Sitawarin, C. Guo, et al. 2025. How Much Do Language Models Memorize?: arXiv: 2505.24832. https://doi.org/10.48550/arXiv.2505.24832.

Musker, S., A. Duchnowski, R. Millière, and E. Pavlick. 2025. "LLMs as Models for Analogical Reasoning." *Journal of Memory and Language* 145: 104676. https://doi.org/10.1016/j.jml.2025.104676.

Nanda, N., A. Lee, and M. Wattenberg. 2023. "Emergent Linear Representations in World Models of Self-Supervised Sequence Models." In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2023, Singapore, December 7, 2023*, edited by Y. Belinkov, S. Hao, J. Jumelet, N. Kim, A. McCarthy, and H. Mohebbi, 16–30. Association for Computational Linguistics. https://doi.org/10.18653/V1/2023.BLACKBOXNLP-1.2.

Neander, K. 2017. "A Mark of the Mental." In *Defense of Informational Teleosemantics*. MIT Press.

OpenAI, Jaech, A., A. Kalai, et al. 2024. OpenAI O1 System Card: arXiv: 2412.16720. https://doi.org/10.48550/arXiv.2412.16720.

Ostertag, G. 2025. "Language Models and Externalism: A Reply to Mandelkern and Linzen." *Computational Linguistics* 51, no. 2: 651–659. https://doi.org/10.1162/coli_a_00551.

Padmakumar, V., C. Yueh-Han, J. Pan, V. Chen, and He He. 2025. Beyond Memorization: Mapping the Originality-Quality Frontier of Language Models: arXiv: 2504.09389. https://doi.org/10.48550/arXiv.2504.09389.

Partee, B. H. 1984. "Compositionality." In: *Varieties of Formal Semantics*, edited by F. Landman and F. Veltman, 281–311.

Pavlick, E. 2023. "Symbols and Grounding in Large Language Models." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 381, no. 2251: 20220041. https://doi.org/10.1098/rsta.2022.0041.

Pearl, L. 2022. "Poverty of the Stimulus Without Tears." *Language Learning and Development* 18, no. 4: 415–454. https://doi.org/10.1080/15475441.2021.1981908.

Pepp, J. 2025. "Reference Without Intentions in Large Language Models." *Inquiry*: 1–19. https://doi.org/10.1080/0020174X.2024.2448482.

Philip and Hemang. 2024. *SimpleBench: The Text Benchmark in Which Unspecialized Human Performance Exceeds That of Current Frontier Models*. Technical Report v1. SimpleBench Team.

Piantadosi, S. 2024. "Modern Language Models Refute Chomsky's Approach to Language." In: *From Fieldwork to Linguistic Theory: A Tribute to Dan Everett*, *Empirically Oriented Theoretical Morphology and Syntax*, edited by E. Gibson and M. Poliak. Language Science Press.

Piantadosi, S., and F. Hill. 2022. Meaning Without Reference in Large Language Models: arXiv: 2208.02957. https://doi.org/10.48550/arXiv.2208.02957.

Prashanth, U. S. V. S. N. S., A. Deng, K. O'Brien, et al. 2025. "Recite, Reconstruct, Recollect: Memorization in LMs as a Multifaceted Phenomenon." In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24–28, 2025*. OpenReview.net.

Putnam, H. 1975. "The Meaning of 'Meaning'." *Minnesota Studies in the Philosophy of Science* 7: 131–193. https://www.upress.umn.edu/9780816657797/language-mind-and-knowledge/.

Quilty-Dunn, J., and E. Mandelbaum. 2018. "Inferential Transitions." *Australasian Journal of Philosophy* 96, no. 3: 532–547. https://doi.org/10.1080/00048402.2017.1358754.

Quilty-Dunn, J., N. Porot, and E. Mandelbaum. 2023. "The Best Game in Town: The Re-Emergence of the Language of Thought Hypothesis Across the Cognitive Sciences." *Behavioral and Brain Sciences* 46: 1–55. https://doi.org/10.1017/S0140525X22002849.

Rogers, A. 2024. A Sanity Check on 'Emergent Properties' in Large Language Models.

Russin, J., S. W. McGrath, D. J. Williams, and L. Elber-Dorozko. 2024. From Frege to ChatGPT: Compositionality in Language, Cognition, and Deep Neural Networks: arXiv: 2405.15164.

Sanderson, G. 2017. "Neural Networks. A 3Blue1Brown YouTube Video Series." https://www.3blue1brown.com/topics/neural-networks.

Sanderson, G. 2024. Large Language Models Explained Briefly.

Seth, A. K. 2025. "Conscious Artificial Intelligence and Biological Naturalism." *Behavioral and Brain Sciences*: 1–42. https://doi.org/10.1017/S0140525X25000032.

Shea, N. 2018. *Representation in Cognitive Science*. OUP.

Shea, N. 2023. "Moving Beyond Content-Specific Computation in Artificial Neural Networks." *Mind & Language* 38, no. 1: 156–177. https://doi.org/10.1111/mila.12387.

Shea, N. 2024. *Concepts at the Interface*. Oxford University Press.

Shevlin, H., and M. Halina. 2019. "Apply Rich Psychological Terms in AI With Care." *Nature Machine Intelligence* 1, no. 4: 165–167. https://doi.org/10.1038/s42256-019-0039-y.

Stechly, K., K. Valmeekam, A. Gundawar, V. Palod, and S. Kambhampati. 2025. Beyond Semantics: The Unreasonable Effectiveness of Reasonless Intermediate Tokens: arXiv: 2505.13775. https://doi.org/10.48550/arXiv.2505.13775.

Stojnić, U. 2022. "Just Words: Intentions, Tolerance and Lexical Selection." *Philosophy and Phenomenological Research* 105, no. 1: 3–17. https://doi.org/10.1111/phpr.12781.

Stoljar, D., and Z. V. Zhang. 2025. "Why ChatGPT Doesn't Think: An Argument From Rationality." *Inquiry*: 1–29. https://doi.org/10.1080/0020174X.2024.2427061.

Strachan, J. W., D. Albergo, G. Borghini, et al. 2024. "Testing Theory of Mind in Large Language Models and Humans." *Nature Human Behaviour* 8, no. 7: 1285–1295. https://doi.org/10.1038/s41562-024-01882-z.

Strobl, L., W. Merrill, G. Weiss, D. Chiang, and D. Angluin. 2024. "What Formal Languages Can Transformers Express? A Survey." *Transactions of the Association for Computational Linguistics* 12: 543–561. https://doi.org/10.1162/tacl_a_00663.

Talmor, A., O. Yoran, R. Le Bras, et al. 2021. "CommonsenseQA 2.0: Exposing the Limits of AI Through Gamification." In *Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Tang, C., B. Lake, and M. Jazayeri. 2025. An Explainable Transformer Circuit for Compositional Generalization: arXiv: 2502.15801. https://doi.org/10.48550/arXiv.2502.15801.

Titus, L. M. 2024. "Does ChatGPT Have Semantic Understanding? A Problem With the Statistics-of-Occurrence Strategy." *Cognitive Systems Research* 83: 101174. https://doi.org/10.1016/j.cogsys.2023.101174.

Ullman, T. 2023. Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks: arXiv: 2302.08399. https://doi.org/10.48550/arXiv.2302.08399.

Vafa, K., J. Y. Chen, A. Rambachan, J. Kleinberg, and S. Mullainathan. 2024. "Evaluating the World Model Implicit in a Generative Model." *Advances in Neural Information Processing Systems* 37: 26941–26975. https://doi.org/10.52202/079017-0846.

Vaswani, A., N. Shazeer, N. Parmar, et al. 2017. "Attention Is All You Need." In *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. V. Luxburg, S. Bengio, et al. Vol. 30, 5998–6008. Curran Associates Inc.

Vegner, I., S. de Souza, V. Forch, M. Lewis, and L. A. A. Doumas. 2025. Behavioural vs. Representational Systematicity in End-to-End Models: An Opinionated Survey: arXiv: 2506.04461. https://doi.org/10.48550/arXiv.2506.04461.

Warstadt, A., and S. R. Bowman. 2022. "What Artificial Neural Networks Can Tell Us About Human Language Acquisition." In *Algebraic Structures in Natural Language*. CRC Press.

Webb, T., K. J. Holyoak, and H. Lu. 2023. "Emergent Analogical Reasoning in Large Language Models." *Nature Human Behaviour* 7, no. 9: 1–16. https://doi.org/10.1038/s41562-023-01659-w.

Wei, J., Y. Tay, R. Bommasani, et al. 2022. "Emergent Abilities of Large Language Models." *Transactions on Machine Learning Research*. https://openreview.net/forum?id=yzkSU5zdwD.

Wilcox, E. G., M. Y. Hu, A. Mueller, et al. 2025. "Bigger Is Not Always Better: The Importance of Human-Scale Language Modeling for Psycholinguistics." *Journal of Memory and Language* 144: 104650. https://doi.org/10.1016/j.jml.2025.104650.

Williams, I. 2025. Can Structural Correspondences Ground Real World Representational Content in Large Language Models?: arXiv: 2506.16370. https://doi.org/10.48550/arXiv.2506.16370.

Wong, L., G. Grand, A. K. Lew, et al. 2023. "From Word Models to World Models: Translating From Natural Language to the Probabilistic Language of Thought." *arXiv preprint arXiv:2306.12672*. https://arxiv.org/abs/2306.12672.

Woydt, T., M. Willig, A. Wüst, et al. 2025. Fodor and Pylyshyn's Legacy—Still No Human-like Systematic Compositionality in Neural

Networks: arXiv: 2506.01820. https://doi.org/10.48550/arXiv.2506.01820.

Wu, Y., A. Geiger, and R. Millière. 2025. "How Do Transformers Learn Variable Binding in Symbolic Programs?" In *Forty-Second International Conference on Machine Learning*.

Wu, Z., C. D. Manning, and C. Potts. 2024. ReCOGS: How Incidental Details of a Logical Form Overshadow an Evaluation of Semantic Interpretation: arXiv: 2303.13716. https://doi.org/10.48550/arXiv.2303.13716.

Wu, Z., L. Qiu, A. Ross, et al. 2024. "Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks." In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, edited by K. Duh, H. Gomez, and S. Bethard, 1819–1862. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.102.

Yao, Y., and A. Koller. 2022. "Structural Generalization Is Hard for Sequence-to-Sequence Models." In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 5048–5062. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.337.

Yildirim, I., and L. A. Paul. 2023. "From Task Structures to World Models: What Do LLMs Know?" *Trends in Cognitive Sciences*. https://doi.org/10.48550/arXiv.2310.04276.

Zečević, M., M. Willig, D. S. Dhami, and K. Kersting. 2023. "Causal Parrots: Large Language Models May Talk Causality but Are Not Causal." *Transactions on Machine Learning Research*. https://openreview.net/forum?id=tv46tCzs83.