Proceedings of the 5th International Conference on Interdisciplinary Humanities and Communication Studies
DOI: 10.54254/2753-7064/2026.39365

Beyond Language Form: Semantic Understanding in Large Language Models

Xize Liu

Shenzhen College of International Education, Shenzhen, China
s24276.liu@stu.scie.com.cn

Abstract. The rapid development of large language models has led to a profound and challenging question: is it enough to prove that these models understand the meanings of language if they perform well in linguistic tasks? This paper argues that the success of the task alone does not fully demonstrate its complete semantic understanding ability. Based on the analysis of the distinction between linguistic form and meaning proposed by Bender and Koller and the evidence of statistical shortcuts, this paper explores why successful task performance does not necessarily demonstrate semantic understanding. However, since human-produced text reflects interactions with the world, models trained based on such data might develop conceptual relations and even reference connections through human language practices. Nevertheless, these abilities are still not enough to constitute a complete understanding. In this way, semantic understanding should be regarded as a multidimensional structure, covering conceptual relations, reference, contextual interpretation, and communicative intention. Each of those dimensions should be assessed separately through tests that exclude surface-level shortcuts, rather than inferred from overall task accuracy. The resulting question is which aspects of semantic understanding language models can demonstrate, and what evidence would be needed to assess each of them.

Keywords: large language models, semantic understanding, language form, reference, communicative intention

1. Introduction

1.1. Background and research question

Large language models can now produce fluent conversations, translate between languages, answer questions, and solve tasks that appear to require reasoning. Some models also achieve high scores on standardized tests and professional knowledge benchmarks. These capabilities have made the question of whether language models understand the meanings expressed by language increasingly important.

However, consistent success on linguistic tasks does not by itself show whether a model understands the meanings expressed by language. This paper asks two related questions. First, to what extent does successful performance on linguistic tasks provide evidence of semantic

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-7064/2026.39365

understanding? Second, which dimensions of semantic ability can text-only training plausibly support, and which remain uncertain?

1.2. Linguistic form and semantic understanding

Bender and Koller distinguish linguistic form from meaning. Linguistic form includes observable words, sentences, symbols, and the patterns in which they occur. In contrast, meaning focuses on the relationship between the linguistic expression and the object it refers to, the contexts in which they are used, and the content that the speaker intends to convey [1].

Large-scale text training can produce representations that capture syntactic patterns, conceptual relations, and information that appears to have semantic structure. The issue therefore lies not in whether text training can produce complex linguistic and conceptual structure, but in whether such structure is sufficient for full semantic understanding.

The complete semantic understanding referred to in this article involves several related but distinct abilities: representing conceptual relations, establishing reference, interpreting expressions in context, and identifying communicative intention. Success in one of these areas does not necessarily imply success in the others.

1.3. The current debate

An influential perspective highlights the gap between linguistic performance and meaning. Bender and Koller argue that learning the distribution of linguistic forms cannot by itself establish the connection between language and communicative intention [1]. Empirical studies also show that models can sometimes achieve correct answers by relying on linguistic shortcuts [2]. Such results create a problem for evaluations that treat benchmark accuracy as direct evidence of understanding.

Other studies challenge the stronger claim that text-only learning can produce no semantic content at all. Piantadosi and Hill argue that meaning does not always depend on direct perceptual contact with objects. Relations between concepts themselves may contribute to meaning [3]. Merrill et al. further show that statistical patterns learned through next-word prediction can contain information about semantic relations such as entailment [4]. On this view, statistical patterns in text can carry traces of the semantic relations that shaped the text in the first place.

The question of reference creates a similar complication. Harnad's symbol grounding problem and Putnam's semantic externalism suggest that relations among internal symbols may not be sufficient to determine what those symbols refer to [5, 6]. However, Mandelkern and Linzen argue that language models may inherit reference through the causal and linguistic history of the texts on which they are trained [7]. A model may therefore have some connection to reference even without directly perceiving the objects described.

1.4. Position and structure of the paper

This paper therefore takes a more cautious position. Text-only training can support conceptual and semantic structure, and may even allow models to inherit certain referential connections through human linguistic practices. However, these abilities do not mean that success on linguistic tasks is sufficient to demonstrate full semantic understanding.

Therefore, evidence for one form of semantic ability should not be treated as evidence for all other forms. Conceptual relations, reference, contextual interpretation, and communicative intention need to be evaluated separately.

DOI: 10.54254/2753-7064/2026.39365

Section 2 develops this problem through Bender and Koller's distinction between form and meaning and through empirical evidence concerning heuristics and linguistic probability. Section 3 discusses the objections regarding the possibility that the statistical relationships within human-produced texts may retain semantic information. Section 4 examines what may still be missing from full understanding, focusing on grounding, reference, context, and communicative intention. The final section brings these arguments together and outlines a multidimensional approach to evaluating semantic understanding.

2. Why linguistic success alone does not establish full semantic understanding

2.1. Argument overview

This section argues that the success of linguistic tasks can depend on multiple internal mechanisms, so relying solely on accuracy would be insufficient for full semantic understanding. The distinction between linguistic form and meaning propose by Bender and Koller provides the basis of this discussion [1].

2.2. The Octopus Test and its implications

Bender and Koller apply the Octopus Test in order to illustrate the distinction between form and meaning. The thought experiment involves two English speakers talking over an underwater cable, where an intelligent octopus learns patterns by listening to their conversation. Upon adequate learning, the octopus could give the response that sounds quite logical and fitting [1].

However, the octopus can only access the messages themselves, but will have no information about the contexts in which the messages are uttered. Should one of the speakers introduce something novel and ask for help conditioned by the current setting, the octopus may lack enough information to respond appropriately. The octopus may know that certain expressions usually appear at the same time, but lack a clear understanding of their specific meanings in the current context [1].

Thus, the experiment demonstrates that merely successfully modeling the language form does not provide the contextual information necessary for reference and communicative intentions.

2.3. Multiple explanations for task success

One can find this problem in benchmarks as well, because the evaluation of the tests scores usually focuses on the outcomes given by the models, while a correct answer does not uniquely determine the path used in generating these results.

For example, recognizing the relation between the premise and hypothesis enables the model to solve the natural language inference (NLI) task. However, high benchmark scores might also result from using heuristic algorithms associated with the right answers. McCoy et al. found that instead of reasoning, NLI models could rely on heuristic methods based on lexical overlap and syntactic structure [2].

Therefore, while the findings of McCoy et al. do not negate the possibility of actual reasoning, they make it quite understandable why correct answers should not be considered as the only evidence for the process. Since both semantic reasoning and linguistic heuristics may yield the same result, accuracy alone cannot explain which process is playing a role. Therefore, the testing should be more targeted in order to separate these possibilities.

DOI: 10.54254/2753-7064/2026.39365

2.4. Empirical evidence: adversarial tests and cipher experiments

The adversarial testing clearly demonstrates the point: when surface patterns are deliberately disturbed, the accuracy is considerably reduced, indicating that its previous success seem to rely on linguistic shortcuts, not the real reasoning processes [2].

Further proof of this claim is given by McCoy et al. who performed the cipher-decoding experiments. Even though one could obtain the correct answer using some fixed decoding rules, the experimental results show a significant disparity: when the correct output has a high linguistic probability, the accuracy of GPT-4 can reach 51%; and if the correct output is a low-probability sentence, the accuracy drops down to only 13% [8]. The above mentioned disparity shows that even though in such kind of tasks the linguistic probability of the output may have nothing to do with the decoding rule, it may still affect model performance.

However, this effect may also be because the model is not good at generating low-probability sentences. Thus, the experiment does not prove the lack of reasoning abilities but demonstrates that the linguistic probability may affect the model's performance regardless of its relevance for the task.

2.5. Summary

The evidence provided in this section allows reaching a limited conclusion: task success alone cannot prove full semantic understanding. However, this conclusion further raises a question: if the human-produced texts reflect meaningful interactions between human beings and the world, is there any part of semantic understanding retained in the statistical properties of them? This issue is discussed in the next section.

3. Objection: relations in text may already constitute a form of meaning

3.1. Purpose of the Section

Human-produced texts involve interactions between humans and the outside world; therefore, their statistical structure could retain semantic and conceptual information. If so, then the idea that text-only learning cannot capture any semantics whatsoever would be less plausible.

3.2. Piantadosi and Hill: meaning without direct perceptual reference

Piantadosi and Hill challenge the idea that meaning always requires direct perceptual contact with an object. Human language features numerous abstractions, such as ideas about justice and wit, that do not have simple perceptual referents [3].

This suggests that meaning can be generated through connections among conceptual objects. Take postage stamps as an example, comprehending its concept goes beyond recognition of their visual shape since it requires knowing the concept of letters, mailing, and delivery [3].

Learning relations among concepts is what a language model is good at due to processing huge corpora of texts, whose internal representations capture various semantic relations such as similarities, taxonomies [3]. Therefore, if these relations are part of meaning, then describing text-only models as possessing "form but no meaning" becomes too strong.

Piantadosi and Hill's argument supports a more restricted point: every semantic relationship learned by the model does not necessarily stem from direct perception.

DOI: 10.54254/2753-7064/2026.39365

3.3. Evidence from Merrill et al.

Entailment is a relation in which the truth of one statement guarantees the truth of another. For example, "A dog is running" entails "An animal is moving." Merrill et al. examine whether probabilities assigned to sequences of sentences by language models can reveal such relations. Their results show that these probabilities can support entailment detection above chance level, suggesting that statistical patterns in text preserve some semantic information [4].

3.4. Summary

These arguments suggest that text-only learning can produce conceptual and semantic structure. Text reflects human interactions with objects, social behaviors, and other people.

Therefore, when the model uses this kind of material as the training source, even if it is not directly involved in the described event, it still has the opportunity to grasp some of the conceptual relationships. Nevertheless, this does not mean that its semantic representation is identical to human ones, nor does it mean that text learning is sufficient for full semantic understanding.

Bender and Koller's distinction remains valuable, but their stronger claim that linguistic form cannot support any meaning needs to be qualified [1]. The problems of reference, context and communicative intention are still unsolved.

4. Response and revision: what is still missing from full understanding

4.1. Argument overview

The remaining limitations of text-based semantic structure can be examined through four issues: Harnad's symbol grounding problem, Putnam's semantic externalism, Mandelkern and Linzen's causal-historical account of reference, and the problem of communicative intention.

4.2. Harnad and the symbol grounding problem

Harnad's symbol grounding problem asks how internal symbols become connected to things in the world. A system may contain a rich network of relations between symbols without explaining how those symbols refer to actual objects [5].

A language model may associate "water" with drinking, rivers, rain, and H_2O. These associations may encode useful conceptual information, but they do not by themselves explain why the word refers to water rather than to something else.

Harnad's argument therefore shows that internal symbolic relations may be insufficient for grounding reference, but it does not prove that direct perception is the only alternative [5]. Reference may also be inherited through the linguistic practices learned by the model [7].

4.3. Putnam's semantic externalism and Twin Earth

Putnam's Twin Earth thought experiment further develops this point. It is assumed that Earth are almost exactly the same as Twin Earth, and speakers on both planets use the word "water" in the same way. The difference is that water on Earth consists of H_2O while "water" on Twin Earth is XYZ [6]. Even if the speakers have identical internal experiences, their words refer to different substances. This suggests that reference is determined both by the external environment and the internal representation [6].

DOI: 10.54254/2753-7064/2026.39365

Thus, for language models, the weaker conclusion is that it may not be enough to completely determine the reference only by the internal language structure. However, this does not mean that any system not interacting with the external world cannot have semantic understanding. The model can still make use of the language community formed by its training data to connect with the reference [7].

4.4. Mandelkern and Linzen: the causal-historical response

According to Mandelkern and Linzen, the text that language model learned comes from real human communication, which is itself related to the real world.

For example, people use the word "water" to refer to H_2O in daily communication and writing. When the language model learns these texts, it may also indirectly establish the connection between the word "water" and H_2O in reality. In other words, without having direct experience with water, the model may inherit this referential relationship [7]. Therefore, from this point of view, language users do not need direct perception to establish references [7]. This poses a challenge to the statement that the text-only model is completely unable to refer.

However, this statement explains the way words might acquire references through their usage history, but not the way the model identifies the specific objects referred to in a particular sentence. Although the model uses the word "water" to refer to H_2O, this does not mean that it can recognize the specific object or sample of water a speaker is referring to in a given context.

4.5. The remaining problem of communicative intention

Another relevant yet unresolved question is that of communicative intention. Comprehending the usual meaning and referents of the parts of a word is often insufficient for sentence comprehension; a listener must also understand what the speaker is trying to convey in a particular context [9].

For example, the sentence "That was brilliant" may either be sincere praise or sarcastic comment depending on the context. Similarly, "It is freezing in here" may simply denote that the room is cold or may euphemistically ask to close the window. In these cases, the model might produce a relevant reply, but the response itself does not prove that it truly understands the speaker's intention or just applies the familiar context model.

Therefore, a more detailed evaluative procedure might be needed. Researchers could select sentences involving irony, metaphor, and indirect requests from conversations in dialogue datasets and build up two different but plausible contexts for the utterance, where one context uses literal sense of the sentence while the other one implies some non-literal one. Afterward, the model is asked to identify the speaker's intention in the given context and understand what cues point to the intention. To reduce reliance on superficial cues, the contexts should be equally plausible and include some examples specially designed to undermine linguistic shortcuts.

Even though these experiments will still not answer whether there is a real understanding in the machine, they may provide more detailed information for researching model's judgment of communicative intention of speakers.

4.6. Summary

Inherited reference may explain the relationship between the words used by the model and objects in the world; however, this does not provide an answer as to how the model understands the speaker's intention in the given context.

DOI: 10.54254/2753-7064/2026.39365

Therefore, in order to test the model's semantic understanding, the following dimensions of understanding should be considered: conceptual relations, reference, context-sensitive interpretation, and communicative intentions. This is also consistent with the distinction between form-based and function-based language skills, proposed by Mahowald et al. [10]. Thus, a model may perform well in one dimension but not in the others.

5. Conclusion

5.1. Main conclusion

The evidence considered in this paper supports a limited conclusion: linguistic task success is not sufficient evidence of full semantic understanding, although text-only learning can still produce forms of semantic structure.

Successful performance can be compatible with different underlying processes, so the final answer alone cannot determine what kind of understanding produced it. At the same time, text-only learning can support conceptual relations and other forms of semantic structure, so the absence of direct perceptual experience does not imply the absence of all meaning.

5.2. Semantic understanding as a multidimensional problem

This conclusion has a methodological consequence: semantic understanding should be evaluated as a set of distinct capacities rather than treated as a single ability.

Current evidence provides support for the claim that conceptual relations can be learned from text.

Reference remains less settled. A model may inherit some referential connections through human linguistic practices, but this does not show that it can identify the intended referent in a particular context.

Context-sensitive interpretation should be tested by keeping linguistic form relatively stable while varying the surrounding situation and examining whether the model's interpretation changes appropriately.

Communicative intention requires separate evaluation because an appropriate response does not by itself show that the model identified what the speaker intended.

5.3. Methodological implications

Methodologically, evaluations should isolate the semantic capacity being tested while controlling for surface cues that could support shortcut solutions. For example, researchers could use paraphrases that preserve meaning while varying syntactic structure, or adversarial examples in which high lexical overlap does not indicate entailment.

This approach does not provide a single yes-or-no test of whether a model understands language. Its advantage is that it replaces an overly broad question with several more precise ones. Instead of asking only whether a language model understands language, researchers can ask which parts of semantic understanding are supported by current evidence and which remain uncertain.

5.4. Future research

Several questions remain open. Future work could develop tests that more clearly distinguish semantic reasoning from statistical shortcuts. It could also examine whether multimodal models,

DOI: 10.54254/2753-7064/2026.39365

which combine language with visual or other forms of input, perform differently on tasks involving reference and contextual interpretation.

Further research could also compare models trained in different linguistic and cultural environments. If part of semantic structure is inherited from the practices recorded in human language, differences between linguistic communities may affect what models learn and how they interpret particular expressions.

A multidimensional framework does not settle the philosophical question of machine understanding, but it makes the empirical question more precise by asking which semantic capacities are actually supported by the evidence.

References

[1] Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463

[2] McCoy, R. T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3428–3448). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1334

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

[4] Merrill, W., Wu, Z., Naka, N., Kim, Y., & Linzen, T. (2024). Can you learn semantics through next-word prediction? The case of entailment. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 2752–2773). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.161

[5] Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena, 42*(1–3), 335–346. https://doi.org/10.1016/0167-2789(90)90087-6

[6] Putnam, H. (1975). The meaning of "meaning." In K. Gunderson (Ed.), *Language, mind, and knowledge (Minnesota Studies in the Philosophy of Science, Vol. 7)* (pp. 131–193). University of Minnesota Press. https://www.upress.umn.edu/9780816657797/language-mind-and-knowledge/

[7] Mandelkern, M., & Linzen, T. (2024). Do language models' words refer? *Computational Linguistics, 50*(3), 1191–1200. https://doi.org/10.1162/coli_a_00522

[8] McCoy, R. T., Yao, S., Friedman, D., Hardy, M. D., & Griffiths, T. L. (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*(41), Article e2322420121. https://doi.org/10.1073/pnas.2322420121

[9] Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics: Vol. 3. Speech acts* (pp. 41–58). Academic Press. https://doi.org/10.1163/9789004368811_003

[10] Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences, 28*(6), 517–540. https://doi.org/10.1016/j.tics.2024.01.011