Do Language Models' Words Refer?

Matthew Mandelkern
Department of Philosophy
New York University
mandelkern@nyu.edu

Tal Linzen
Department of Linguistics and
Center for Data Science
New York University
linzen@nyu.edu

What do language models (LMs) do with language? They can produce sequences of (mostly) coherent strings closely resembling English. But do those sentences mean *something, or are LMs simply babbling in a convincing simulacrum of language use? We address one aspect of this broad question: whether LMs' words can* refer, *that is, achieve "word-to-world" connections. There is prima facie reason to think they do not, since LMs do not* interact *with the world in the way that ordinary language users do. Drawing on the externalist tradition in philosophy of language, we argue that those appearances are misleading: Even if the inputs to LMs are simply strings of text, they are strings of text with* natural histories, *and that may suffice for LMs' words to refer.*

1. Introduction

Language models (LMs) are machine learning systems trained to capture the statistical distribution of sequences of words in a collection of texts. Once trained, these systems can be used to generate new texts drawn from the same distribution. When generating those texts, these systems, especially those based on neural networks trained on large amounts of text, often *appear* to be producing series of meaningful sentences. But some have argued that LMs are really nothing more than babbling "stochastic parrots" (a term created by Bender et al. 2021). One might think, in particular, that LMs produce a mere simulacrum of meaningful human language, because their words never achieve the referential connections required for meaning (Ostertag 2024; Titus 2024).

This skepticism can be motivated in a number of ways. Here we will address one argument, which says that because LMs' only input is strings of symbols, they cannot produce referential words: *reference* cannot be derived from *form*. Here we will show that this argument is mistaken: LMs' inputs are not bare strings of symbols, but strings of symbols with certain natural histories that connect them to their referents. We draw

---

on the externalist tradition in the philosophy of language, which has convincingly argued that reference is achieved, not by having particular beliefs about, experiences with, or capacities involving words' referents (the kinds of things that LMs most plausibly lack) but rather by standing in the right kind of natural history to the referents. Hence the natural histories of LMs' inputs may be enough to ground reference for the LMs' outputs, even if the LMs have no way to grasp or access those histories.

2. Our Question

To help get clear on the question we are asking, consider a contrast based on Putnam 1981. Suppose that, at a picnic, you observe ants wending through the sand in a surprising pattern, which closely resembles the English sentence "Peano proved that arithmetic is incomplete". At the same time, you get a text message from Luke, who is taking a logic class. He writes, "Peano proved that arithmetic is incomplete".

Intuitively, the two cases are very different, despite involving physically similar patterns. The ants' patterns do not *say* anything; the ants just happen to have formed patterns which *resemble* meaningful words. Of course, *you* can interpret the pattern, just as you can interpret an eagle's flight as an auspicious augur; but these are interpretations we overlay on a natural pattern, not meanings intrinsic to the patterns themselves. By contrast, Luke's words *mean something* on their own (regardless of whether you or anyone else interprets them): namely, that Peano proved that arithmetic is incomplete. What Luke said is false: It was Gödel who proved incompleteness. But Luke *said* something, whereas the ants didn't say anything at all. In particular, he said something false *about Peano*, which means that his (use of the) word "Peano" managed to *refer* to Peano.

We are interested in the question of whether the outputs of LMs are more like the ants' patterns or like Luke's text: Do they merely *resemble* meaningful sentences, or are they in fact meaningful sentences? Hence suppose that an LM produces the sentence "Peano proved the incompleteness of arithmetic". Does this amount to the production of a *false but meaningful* sentence, like Luke's text, or is it a *meaningless pattern* which merely resembles a meaningful sentence, like the ants' pattern?

There are many ways we could make this question precise; here, we will address the question of whether the words that LMs produce **refer**, that is, achieve the word-to-world connections that ground meaningful language use. For instance, we can ask whether the word "Peano" output by the LM refers to Peano, or rather doesn't refer at all; we'll use *referentiality* for the thesis that LMs' words do refer.[1] This is intimately related to the question of whether the sentence the LM produces is meaningful, since meaning (nearly all agree) requires reference.

There are many important questions to ask about LMs, language, and meaning. Our question is just one. A closely related question, which we will not tackle head-on, is whether the LM *itself* refers (whether it achieves *speaker reference*), which can in principle come apart from the question of whether *its words* refer (Kripke 1977). While our question is narrow, it strikes us as being important to know whether LMs are more

---

1 As Hawthorne and Lepore (2011) emphasize, "word" is polysemous, and on one usage, words are partly individuated by speaker intention (Kaplan 1990). Since we don't want to assume LMs have intentions, this is not the usage of "word" we have in mind. Instead, we will use "word" as a thin sense, appropriate to this article, to refer to concrete instances of (potentially meaningless) strings of symbols, so that it is uncontroversial that LMs output words, and controversial only whether those words are meaningful.

like ants in the sand, producing strings that happen to resemble human language, or rather are producing meaningful strings of human language.

Our goal in this short article is to bring arguments to the attention of the AI and NLP communities from the externalist tradition in the philosophy of language (Kripke 1980; Putnam 1975). These arguments undermine a very tempting argument: that LMs' words cannot refer because LMs' inputs are only strings of symbols and they do not interact directly with the external world. However, we make no claim to a decisive argument that LMs' words *do* refer. There are many subtle issues in the neighborhood of those we address, and much more to say about every issue we raise. Our goal is to briefly explain externalism and begin to explore its ramifications for LMs in an accessible way that we hope will spur further exchanges and explorations of these ideas between the philosophy and AI/NLP communities.

Our paper is not the first to connect externalism to the question of referentiality. Cappelen and Dever (2021) defend referentiality on externalist grounds, but argue that standard externalist accounts don't apply to LMs because of "crippling anthropocentric biases". By contrast, we will argue that standard externalist accounts are at least prima facie compelling with respect to LMs. Butlin (2021) argues that externalism vindicates referentiality for chatbots but *not* for LMs. DeVault, Oved, and Stone (2006) argued that externalism did not vindicate referentiality for NLP as it existed at the time, but could in principle apply. The accounts most similar to ours in viewing standard externalism as providing a possible vindication of referentiality are the excellent recent papers Pavlick (2023) and Mollo and Millière (2023). We are very sympathetic to both discussions of externalism, but neither one explains the original motivations behind externalism and how those same motivations might apply to LMs, a lacuna we aim to rectify here.

3. The Grounding Problem

LMs *appear* to produce meaningful strings. Suppose an LM outputs the string "Apples are fruits"; that string appears to be saying that apples are fruits, and hence using "apples" to refer to apples. Still, you might think that the LM's output is no more intrinsically meaningful than figures accidentally sketched by ants in the sand because the LMs' use of language is not *grounded* in beliefs, experiences, and capacities involving the external world. After all, an LM trained on language alone has had no sensory contact with apples; it has never tasted or touched or smelled them. It cannot sort apples from oranges. You might also think that LMs lack any desires and beliefs about apples. Without a doubt, LMs do not have the store of apple-related memories and experiences that ordinary speakers of English generally have. So in the absence of relevant experiences, beliefs, desires, and capacities, how could they refer to apples? LMs are just too isolated from the world for their symbols to be connected to it in the way required for reference.

Another way to put the point is that *form* does not suffice to ground *reference*. Everyone will agree that the way that symbols are arranged together does not determine the referent of the arrangement. But, the thought goes, *form is all that LMs have*, and so LMs cannot achieve reference.[2]

---

2 Bender and Koller (2020) make an argument with a similar form, but it differs in subtle ways: First, they are concerned with relating forms to a broader notion of meaning than just reference; and second, their main argument is that training on form is insufficient for an LM to *learn* such relations, and they did not directly consider whether such a system could produce meaningful language.

4. Externalism About Reference

Drawing on the externalist tradition in philosophy of language, we will argue that this objection to referentiality is misplaced. What grounds reference is not beliefs, experiences, and discriminatory capacities of the kind that LMs obviously lack. Rather, what grounds reference is the *natural histories* of words: the causal-historical links between a speech community's use of a word and the word's referent.[3]

Recall Luke. Suppose that he had never heard of Peano before recently reading a very short article about him, comprising just the sentence "In 1931, Peano proved that arithmetic is incomplete". This is the first that Luke has heard of Peano, and hence amounts to his entire body of information about Peano.

Now, as a matter of fact, this sentence is false. Nonetheless, just by being exposed to this sentence, Luke comes to be able to use the word "Peano" to refer to Peano. How do we know? Well, if he texts you "Peano proved that arithmetic is incomplete", his text *says something false*, namely, something *about Peano*: that he proved incompleteness. So Luke's use of the word "Peano" refers to Peano. This is striking, since Luke's only substantive belief about Peano is the false belief that he proved that arithmetic is incomplete. Luke has nothing by way of experience with Peano, no fluency with regard to facts about his life—indeed, no ability to discriminate him from Gödel.

Of course, he has some trivial beliefs: that Peano is the referent of "Peano", that Peano is called "Peano" by Luke's speech community, and so on—something we'll return to below. There are other things you might think that he could infer from the sentence in question—for instance, that "Peano" was called Peano by his contemporaries. But, as Kripke (1980) points out in an analogous discussion, we can easily suppose this to have been false: For instance, suppose that both Peano and Gödel were called "Georg Kreisel" by their contemporaries, and only later dubbed "Peano" and "Gödel", respectively. Luke would still be able to refer to Peano with "Peano".

One lesson of thought experiments like this is that it is strikingly easy for a speaker to use a word referentially. This does not require substantive experience, beliefs, or capacities vis-à-vis its referent. This runs contrary to a prima facie compelling picture on which reference is achieved by having a certain body of information and experience with the referent. That picture was long popular in philosophy but was refuted by externalists like Kripke and Putnam; and we think that now-defunct idea underlies the argument about grounding sketched in the last section.

How does Luke's use of "Peano" come to refer to Peano, if not via Luke's beliefs about and experiences with Peano? There is much to say about this question. We will briefly summarize two observations. The first is that a word use's *natural history* can largely determine its reference. This is apparently what goes on with Luke's use of "Peano": The fact that Luke is part of a speech community whose history of using the word traces back to Peano, rather than Gödel, seems to be enough to make the referent of the word in Luke's usage Peano, rather than Gödel—even though Luke's *beliefs* about the referent of "Peano" uniquely pick out Gödel, not Peano.

For a striking illustration of the point, consider Putnam's famous *Twin Earth* case. Compare speakers of English in an era when they do not yet know the chemical composition of water to denizens of an alternate reality, Twin Earth, which is identical

---

to earth in all macroscopic properties, but where the chemical composition of the stuff that in every macroscopic way resembles water is not H_2O but rather XYZ. By hypothesis, there are people on Twin Earth who speak a language which is superficially indistinguishable from English. But consider what their word "water" refers to: it refers to the stuff *in their reality* which they call "water", which is XYZ, not H_2O. To see this, note that we Earthians can truly say, in our dialect, "There is no water on Twin Earth; instead, it contains a macroscopically identical substance, made of XYZ". Conversely, Twin Earthians could, in *their* version of English, truly say "On Earth there is no water, only H_2O".

What makes it the case that "water" in Earthian English refers to H_2O, while "water" in Twin Earthian refers to XYZ? It is *not* the beliefs of the speakers about the liquids' respective chemical compositions, since, by assumption, they do not yet know the chemical compositions of their respective watery stuffs. (You could argue that the meanings of the word *change* when the speakers learn the chemical compositions, but this is wrong: Scientists discovered *that water was H_2O*, they did not change the meaning of "water".) Rather, it seems to simply be the fact that the stuff in the natural history of the word "water" on earth was made of H_2O, whereas the stuff in the natural history of the word "water" on Twin Earth was made of XYZ.

There are different ways to spell out the notion of natural history here. Kripke (1980) talks evocatively of an initial baptism event, where a word is associated with a referent ("Let's use this word 'water' to refer to this stuff over here"), which is connected to current usage by myriad causal-historical chains. That is clearly an idealization; the origin stories of word reference can be diffuse and complicated (see Evans [1982] for a different approach). But it provides a helpful schematic picture of how words can refer in ways that go beyond or even contradict the speaker's representation of the referent.

In other cases, what determines reference is *deference* to experts. Consider another case of Putnam's: an ordinary speaker's use of the words "elm" and "beech". Most speakers of English (let us suppose) don't know the difference between elms and beeches: They don't have beliefs that distinguish the two types of trees; nor could they distinguish them if they were asked to; nor, in some cases, have they had any experiences of either. Still, "beech" in the mouth of such speakers refers to beeches, not elms; while "elm" in the mouth of an ordinary speaker refers to elms, not beeches. To see this, suppose such a speaker pointed to a tree that happened to be an elm, and said "This is a beech": We would take them to be saying something false, namely that the tree before them is a beech. So "beech" still refers to beeches in their mouth. Likewise for elms.

How can this be? Part of the story is, again, the natural history of the word: "Beech" in this speech community has historically been used to refer to beeches, and that is enough for ordinary speakers to use it that way, even if their beliefs and experiences don't connect them to beeches in any particular way. Another part of the story, Putnam argues, is *deference*: Ordinary speakers use "beech" to refer to *whatever tree experts take it to refer to*, and that is enough to refer to beeches.

5. The Upshot for LMs

Luke's use of "Peano" refers to Peano, not because of his beliefs about Peano, but thanks to the fact that his use of "Peano" traces back via causal-historical lines to Peano himself. By contrast, the ants' pattern does not refer to Peano, because it doesn't trace back to Peano via such causal-historical lines.

Now back to LMs. Suppose that an LM is trained on the very same text about Peano that Luke saw, and then generates the (false) sentence "Peano proved the incompleteness of arithmetic". Does the LM's word "Peano" succeed in referring to Peano?

If you thought that the difference between the ants and Luke was that Luke's words, but not the ants', are grounded in substantive beliefs, capacities, and experiences involving Peano, then you should think that LMs are more like ants (the skeptical position described in §3). But we have now seen strong reason to think that this is *not* what accounts for the difference; rather, the difference is that Luke's word, but not the ants', is connected to Peano by the right kind of causal-historical chains.

So does the LM's use of "Peano" trace back to Peano via the right kind of causal-historical chains to secure reference? This is a tricky question, and we don't have decisive considerations one way or another. We will presently consider some reasons to think the answer is "no", and argue that none of them is especially convincing. But first, let us emphasize that *this* is the central question to ask, for if the answer is yes, then LMs' use of "Peano" refers to Peano: No beliefs, capacities, or experiences with Peano are required for reference in general, and so there is no particular worry in the case of LMs either.

One way to formulate the skeptical argument from grounding is that LMs have access only to form, and form underdetermines reference. Now we can see what is wrong with this argument. *The inputs to LMs are not just forms, but forms with particular histories of referential use.* And those histories ground the referents of those forms, whether or not those histories are known or accessible. For instance, Luke wrongly believes that the causal history of his use of "Peano" traces back to the discoverer of the incompleteness proof; nonetheless, the simple fact that the *actual* history of his use of "Peano" traces back to Peano, not Gödel, suffices to make Peano, not Gödel, the referent of his use of "Peano".

The key question for referentiality is thus whether an LM's use of "Peano" traces back to Peano via the right kind of causal-historical chains to secure reference. While a strong case could be made that LMs lack substantive experiences, capacities, and perhaps also beliefs about their words' referents, prima facie considerations tell in favor of counting contemporary neural network LMs such as ChatGPT as part of our speech community in a way that secures the right kind of causal-historical chains between words and referents: Namely, they speak and interact in apparently linguistically ordinary ways. On the spectrum between ants and Luke, they *look* closer to Luke. This is, of course, only prima facie evidence: There might be compelling reasons to conclude that LMs are *not* part of a speech community in the relevant sense.[4] We won't try to argue decisively against this position. But we will briefly address some of the most obvious arguments.

One tempting thought is that being part of a speech community that uses a word to refer in a certain way requires being disposed to draw the particular inferences from that word that are drawn by members of that community. So, for instance, even if you can't distinguish elms from beeches, there are lots of things that you can still infer from "This is an elm", for instance, "This is a tree", or "This is an object". Similarly, even though Luke doesn't know anything about Peano, he is likely to infer certain things about him, based on his general knowledge about logic and the world. Clearly inferences aren't *enough* to determine reference. Earthians and Twin Earthians will associate the same

---

4 From the other direction, you might think that LMs are part of *many* speech communities, given the diversity of their training data. That seems right to us; see our discussion below about disambiguation.

inferences with "water", but the words in their respective languages refer to different things. But inference drawing of some kind may still be a necessary condition on using a word to refer (see, e.g., Putnam's [1975] discussion of stereotypes).

But it is hard to see how the capacity to draw appropriate inferences would be a problem for the claim that LMs' words can refer. What seems potentially hard for LMs is "word-to-world" connections, since LMs are in some clear sense isolated from the external world. By contrast, "word-to-word" connections, and in particular inferences, should in principle be possible for a language model to acquire, since connections between words just are its stock in trade (see Potts 2020; Merrill, Warstadt, and Linzen 2022; Piantadosi and Hill 2022 for relevant discussion).

Another skeptical worry is that LMs don't have the right kinds of *intentions* to produce referring words. The worry is: To produce an instance of "Peano" that refers to Peano, the speaker must intend to use "Peano" to refer to Peano, and LMs can't do that.

There is a thick and a thin version of this thesis. The thick version is that a speaker can produce a word that refers to x iff she intends to use it to refer to whatever verifies some cluster of her substantive beliefs about x. This is a tempting picture, but we have seen that it is refuted by the externalist considerations we've seen so far. Luke intends to use "Peano" to refer to whoever proved that arithmetic is incomplete (since he intends to speak truly); but he in fact succeeds in using "Peano" to refer to *Peano*, not Gödel. Speaker intentions in this thick sense are thus neither necessary nor sufficient for reference.

A thinner version of the view is more plausible, given externalist considerations: a speaker can produce a referential word iff she intends that word to refer to whatever its natural history in her speech community determines as its referent. Again we can consider more and less demanding versions of this thinner view. On a more demanding version, referential production requires high-level reflection about the notion of speech community and reference; you need to think something like, "I'm using this word to refer to whoever 'Peano' refers to in my speech community" in order to produce a referential instance of "Peano". That is obviously wrong; many referentially successful speakers never think *explicitly* about reference or speech communities. A less demanding version says that what is required is only some kind of *implicit* representation of oneself as part of a speech community and intention to refer in conformity with the word's natural history in that speech community—the kind of implicit representation that can be ascribed to, say, a child who successfully produces referential words, but may be unable to speak coherently about reference or speech communities.

Then the key question is whether LMs can have this kind of lightweight intention to refer in conformity with a speech community. To be clear, although lightweight, this is not a trivial criterion to meet. Ants don't meet it; Luke does. LMs are a much harder case, and it would take a lot of work in philosophy of mind to explore whether they can intend to refer in conformity with their speech community in this implicit, sub-personal way. Certainly ordinary speakers are happy to talk about machines acting intentionally (we'll happily say that a robotic arm intends to pick up a cup of water). And it is plausible to argue that LMs could, and increasingly do, learn to model communicative intention in order to accurately predict upcoming words (Andreas 2022). But there is still plenty of room to deny that LMs intend to refer in line with their speech community (see Ostertag [2024] for one development of this position). For instance, you might think that intending requires consciousness, or theory of mind, or self-representation, or intrinsic goals, or something else, and then argue that LMs lack these things. We leave this as an open question. But in each case, we warn against holding LMs to a higher standard than we hold ordinary speakers, like children, to whom we are happy

to ascribe referential success even though they may not have clear representations of their speech community.

A more specific worry is that intention is required for disambiguation. Suppose an LM produces the string "Kim won't eat the kiwi", to use a reviewer's example. Does "kiwi" refer to the fruit or the bird? In the case of human speech, it is natural to answer this question by adverting to the speaker's referential intentions. We think, however, that homophones can be disambiguated without relying on intentions. Coenen et al. (2019) show, for instance, that BERT's contextualized vector representations of the instances of an ambiguous word can be clustered into different locations in vector space for each sense of the word (see also Garí Soler and Apidianaki 2021); as such, the sense of a particular use of a word can often be inferred from its embedding. If we only have access to the LM's output probabilities, we expect the probability of birds other than kiwis to be higher when "kiwi" refers to the bird. Even given only black-box access to the LM, we can disambiguate "kiwi" using linguistic context: What other words did the model generate in the same sentence, or in the previous or next sentence? So we do not think that disambiguation must involve the ascription of speaker intention in any thick sense.

6. Conclusion

Perhaps LMs are like ants swarming meaninglessly in the sand, producing mere simulacra of referential language. This is a natural conclusion if you think that reference is achieved by connecting a word to whatever it is that satisfies a body of substantive beliefs, experiences, and capacities involving the referent. Externalist arguments in the philosophy of language, however, have shown that reference does not work that way: reference turns out to be surprisingly disconnected from speakers' beliefs, and surprisingly easy. Simply being part of a speech community whose use of a word is historically connected to a particular referent in the right way suffices for a speaker to be able to produce a referring instance of that word.

This helps vitiate a natural worry about referentiality, the worry from grounding. Even if LMs' inputs are only strings, those strings have a natural history that connects them to referents; and those natural histories are, in principle, enough to make their words refer. Of course, there are many things that LMs can't do. They may be totally incompetent with respect to the substantive facts about a given domain. But that doesn't obviously stop them from referring, any more than Luke's erroneous beliefs about Peano stops *his* use of "Peano" from referring to Peano.

As we have seen, there might still be reasons to doubt that LMs refer based on doubts about whether they can really be seen to be part of a speech community; but it is *this* question that is crucial for the question of referentiality, not questions about whether LMs have the right kinds of experiences, beliefs, desires, and so on to ground reference.

Even if you conclude that LMs are not part of our speech community, this does not foreclose the possibility that their words still refer. For instance, Lederman and Mahowald (2024) argue that LMs are not part of our speech communities, but their words nevertheless refer because they are causally sensitive to the *intelligibility* of their data. For all we've said, this position, or yet another one, could turn out to justify the claim that LMs' words refer, even if a more straightforward externalist vindication of referentiality along the lines we have explored here ultimately fails.

Let us close with a thought experiment. Suppose that Izzy is isolated at birth in a sensory deprivation chamber; his only contact with the outside world is via an

interactive screen, directly sensitive to his brainwaves, which displays text and can send back messages. Amazingly, Izzy gradually comes to be an apparently competent user of the screen. He can send messages of varying degrees of sophistication and correctness, like "Get me out of here!" or "Peano proved that arithmetic is incomplete", and he can interact with inputs on the screen in linguistically fairly normal ways. Presumably most people's first reaction would be that Izzy has, amazingly, become a part of our speech community, and that when he writes out something like "Peano proved that arithmetic is incomplete", he has managed to refer to Peano. On the face of it, there is a deep puzzle here about how his words could refer to the external world, given Izzy's isolation from it. But externalism teaches us that there is not really a puzzle here: Izzy's words can refer to the external world even though he is profoundly isolated from it, because reference is grounded in the causal-historical chains running through a speech community connecting a use of a word to its referent. In the same way, externalism dissolves the puzzle about how an LM's use of language could be referential despite its isolation from the external world.

Acknowledgments

References

Andreas, Jacob. 2022. Language models as agent models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5769–5779. https://doi.org/10.18653/v1/2022.findings-emnlp.423

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret 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, pages 610–623. https://doi.org/10.1145/3442188.3445922

Bender, Emily M. and Alexander 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*, pages 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463

Butlin, Patrick. 2021. Sharing our concepts with machines. *Erkenntnis*, 88:3079–3095. https://doi.org/10.1007/s10670-021-00491-w

Cappelen, Herman and Josh Dever. 2021. *Making AI Intelligible*. Oxford University Press. https://doi.org/10.1093/oso/9780192894724.001.0001

Coenen, Andy, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viégas, and Martin Wattenberg. 2019. Visualizing and measuring the geometry of BERT. *Arxiv preprint arxiv 1906.02715*.

DeVault, David, Iris Oved, and Matthew Stone. 2006. Societal grounding is essential to meaningful language use. In *Proceedings of the 21st National Conference on Artificial Intelligence and the 18th Innovative Applications of Artificial Intelligence Conference, AAAI-06/IAAI-06*, pages 747–754.

Evans, Gareth. 1982. *The Varieties of Reference*. Oxford University Press.

Garí Soler, Aina and Marianna Apidianaki. 2021. Let's play mono-poly: BERT can reveal words' polysemy level and partitionability into senses. *Transactions of the Association for Computational Linguistics*, 9:825–844. https://doi.org/10.1162/tacl_a_00400

Hawthorne, John and Ernie Lepore. 2011. What are words? Comments on Kaplan (1990), on Hawthorne and Lepore, and on the issue. *Journal of Philosophy*, 108(9):486–503. https://doi.org/10.5840/2011108925

Kaplan, David. 1990. Words. *Aristotelian Society Supplementary Volume*, 64(1):93–119. https://doi.org/10.1093/aristoteliansupp/64.1.93

Kripke, Saul. 1977. Speaker's reference and semantic reference. *Midwest Studies in Philosophy*, 29:255–276. https://doi.org/10.1111/j.1475-4975.1977.tb00045.x

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

Lederman, Harvey and Kyle Mahowald. 2024. Are language models more like libraries or like librarians? Bibliotechnism, the novel reference problem, and the attitudes of LLMs. Manuscript, University of Texas at Austin.

Merrill, William, Alex Warstadt, and Tal Linzen. 2022. Entailment semantics can be extracted from an ideal language model. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 176–193. `https://doi.org/10.18653/v1/2022 .conll-1.13`

Mollo, Dimitri Coelho and Raphaël Millière. 2023. The vector grounding problem. *arXiv preprint arXiv:2304.01481*.

Morgan, Marcyliena H. 2014. *Speech Communities*. Cambridge University Press. `https://doi.org/10.1017 /CB09781139151269.002`

Ostertag, Gary. 2024. Large language models and externalism about reference: Some negative results. Manuscript, Icahn School of Medicine and Mount Sinai & The Graduate Center, CUNY.

Patrick, Peter L. 2004. The speech community. In J. K. Chambers, Peter Trudgill, and Matalie Schilling-Estes, editors, *The Handbook of Language Variation and Change*. Wiley, chapter 23,

pages 573–597. `https://doi.org/10 .1002/9780470756591.ch23`

Pavlick, Elle. 2023. Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A*, 28(381):20220041. `https://doi.org/10 .1098/rsta.2022.0041`, PubMed: 37271171

Piantadosi, Steven T. and Felix Hill. 2022. Meaning without reference in large language models.

Potts, Christopher. 2020. Is it possible for language models to achieve language understanding? Blog post. `https://chrisgpotts.medium.com/is -it-possible-for-language-models-to -achieve-language-understanding -81df45082ee2`

Putnam, Hilary. 1975. The meaning of 'meaning'. In K. Gunderson, editor, *Language, Mind, and Knowledge*, pages 131–193. Volume 7 of Minnesota Studies in the Philosophy of Science series. University of Minnesota Press. `https:// doi.org/10.1017/CB09780511625251.014`

Putnam, Hilary. 1981. *Reason, Truth, and History*. Cambridge University Press. `https://doi.org/10.1017 /CB09780511625398`

Titus, Lisa Miracchi. 2024. Does ChatGPT have semantic understanding? *Cognitive Systems Research*, 83(101174):1–13. `https://doi.org/10.1016/j.cogsys .2023.101174`