Do Language Models Have Semantics? On the Five Standard Positions

Anders Søgaard
CPAI, University of Copenhagen
Lyngbyvej 2, DK-2100 Copenhagen
soegaard@di.ku.dk

Abstract

We identify five positions on whether large language models (LLMs) and chatbots can be said to exhibit semantic understanding. These positions differ in whether they attribute semantics to LLMs and/or chatbots trained on feedback, what kind of semantics they attribute (inferential or referential), and *in virtue of what* they attribute referential semantics (internal or external causes). This allows for $2^4 = 16$ logically possible positions, only five of which have been argued for. Based on a pairwise comparison of these five positions, we conclude that the better theory of semantics in large language models is, in fact, a sixth combination: Both large language models and chatbots have inferential and referential semantics, grounded in both internal and external causes.

1 Introduction

Large language models (LLMs) are, at heart, trained to solve the so-called *cloze* task, introduced by educational psychologist Wilson L. Taylor in 1953. Each instance of a cloze test is a piece of text in which a word or subword token is omitted:

(1) Look in thy mirror and tell the face thou ____

Taylor would ask his students to guess the missing word and fill in the blank. LLMs are being trained to do the same. Some have argued that because solving cloze tasks is the proper function of LLMs, they cannot possibly acquire *semantic understanding* (Bender and Koller, 2020; Landgrebe and Smith, 2021; Hicks et al., 2024) – by which we shall mean the comprehension and, maybe, production of texts in their linguistic context, but not in their extralinguistic context. Why would anyone think such a capacity is beyond reach for LLMs? Because solving cloze tasks – what LLMs have been trained to do – is essentially a *memorization* task, the kind of thing that parrots can do *without* semantic understanding (Searle, 1980; Bender and Koller, 2020; Bender et al., 2021).

	Inferential	Int-Referential	Ext-Referential	Chat-Ext-Referential
P1				
P2	+			
P3	+	+		
P4	+		+	
P5	+			+
P6	+	+	+	

Figure 1: Five positions in the LLM Understanding Debate

However, human memory is limited. So is the memory of LLMs. While some of us remember their Shakespeare well enough to infer that the missing word in (1) is *viewest*, others will not, and most LLMs can only memorize a small fraction of the training data (Lu et al., 2024). In other words, while the proper function of LLMs is to solve cloze tasks, they do not have enough memory to do so. That means they have to adopt another strategy. They have to find a way to guess what is omitted in Shakespearean sonnets or physics textbooks. Such guessing requires semantics.[1]

In order for our intuitions to not get in our way, we should remind ourselves that committing to the view that LLMs exhibit semantic understanding does not mean you have to commit to the idea of LLMs teaching themselves semantics. The random initialization of an LLM opens up a space that is astronomically big. All you have to commit to is the idea that the space is big enough for there to be semantic understanding in it, as well as to the idea that an LLM will benefit from accidentally stumbling upon such a capacity.

There are, as far as we can see, five standard

[1]Having semantics typically mean outputting words with semantic content, having internal states with semantic content, or understanding words with semantic content. I assume that generation and production turn on the same semantic capacities and will not bake in any assumptions about internal states.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 25910–25922

positions on semantic understanding in LLMs in the literature: **(i)** LLMs are all syntax, no semantics (Searle, 1980; Bender and Koller, 2020; Landgrebe and Smith, 2021); **(ii)** LLMs have inferential semantics, but not referential semantics (Rapaport, 2002; Piantadosi and Hill, 2022; Baggio and Murphy, 2024); **(iii)** LLMs have both inferential and latent referential semantics by virtue of structural similarity or world models (Søgaard, 2022, 2023; Pavlick, 2023; Lyre, 2024); **(iv)** LLMs have both inferential and referential semantics by virtue of participation in communicative chains, but *not* by virtue of having world models (Cappelen and Dever, 2021; Mandelkern and Linzen, 2023; Lederman and Mahowald, 2024; Koch, forthcoming);[2] **(v)** LLMs have inferential semantics and may acquire referential semantics through learning from human feedback. Corollary: Only chatbots have referential semantics (Butlin, 2021; Mollo and Millière, 2023). We will argue for a sixth position, also listed in Figure 1. Here, Int-Referential refers to the idea that LLMs have referential semantics grounded in internal causes, i.e., world models; Ext-Referential refers to the idea that LLMs have referential semantics grounded in external causes such as training data, developers and annotators, model selection and evaluation, and human (responsive) agents), i.e., being part of the right communicative chains; Chat-Ext-Referential refers to the idea that referential semantics is exclusively for chatbots, because semantic understanding must be grounded in human (responsive) agents).[3] I present the following arguments for our position:

Position 1 and 2 I will argue the P1-P2 distinction is less important than it seems. The distinction between syntax and inferential semantics is sensitive to how syntax is defined, and I suspect that proponents of P1 and P2 are actually largely in agreement. The only way to uphold a position in which LLMs do not have inferential semantics, is by positing that all their inferences were memorized. This flies in the face of the ability of LLMs to generalize to unseen data (Lotfi et al., 2024), as well as the fact that memorization is a highly inefficient strategy (Michaud et al., 2023), and LLMs are generally too small to memorize *all* of their training data (Lu et al., 2024). I therefore give proponents of P1 the benefit of doubt and assume that despite of their fondness for slogans such as LLMs are 'syntax all the way down' (Searle, 1980), or 'LLMs lack semantic knowledge'[4] or 'semantic understanding' (Titus, 2024), what they really mean is (the more reasonable) P2.

Position 2 and 3 Reasons to prefer P3 over P2 turn on whether LLM representations facilitate reference or alignment with other representations. There are empirical facts – from unsupervised machine translation and unsupervised multimodal inference – that seem to support the idea that LLM representations do in fact facilitate such reference or alignment (Li et al., 2023b; Lyre, 2024; Gurnee and Tegmark, 2024; Baggio and Murphy, 2024; Jha et al., 2025). Our argument is a *modus tollens*: If LLMs only learned syntax or inferential semantics (p_1), they would *not* have made unsupervised bilingual dictionary extraction possible ($p_1 \rightarrow \neg q$).[5] Since they make such extraction possible (q), it follows that p_1 is false. I discuss whether alignment with other representations checks all the boxes of the job description for referential semantics; to keep everyone happy I distinguish between weak and strong referential semantics and argue that we at least have empirical support for LLMs exhibiting weak referential semantics.[6]

Position 4 and 5 The distinction between LLMs and chatbots – and the view that while LLMs do *not* have referential semantics, chatbots *do* – is, in my view, hard to defend. Specifically, Butlin (2021) and Mollo and Millière (2023) are forced to accept that two identical models A and B can be such that A exhibits referential semantics, whereas B does not. This happens if a pre-trained LLM (A) is not updated during fine-tuning, say, because it was already optimal. In light of how LLMs are constantly copied, forked and backed up, this quickly leads to absurd scenarios. Some copies will have semantic understanding because of their copying history; others will not because of theirs.

[2] Lederman and Mahowald (2024) are arguably agnostic about semantic understanding, but have an explicitly externalist story about semantic intelligibility by virtue of being chained to training data: 'LLMs can produce novel text which is nevertheless derivatively meaningful, because they copy individual tokens from their [training data], and assemble them in ways that are causally sensitive to the high-level feature of intelligibility in their [training data].'

[3] Arguably, Ext-Referential and Cha-Ext-Referential are mutually exclusive, making the set of valid positions 16-4=12.

[4] https://tinyurl.com/3y6vyhej

[5] Or think of multi-modal/cross-model alignment, or results that representations are structurally similar to knowledge bases (Gammelgaard et al., 2023).

[6] Baggio and Murphy (2024) refer to weak referential semantics as having aboutness.

Position 3 and 4 If our arguments are on track, we are left with two options, P3 and P4. There are, it seems, good arguments for both. I will discuss whether P4 is too permissive and consider a possible reason to reject Kripkean externalism: Pure externalism cannot account for the influence of random initialization on LLMs. It is unclear whether externalists have to account for this.

I will argue for a sixth position (**Position 6**), a hybrid position. Some concepts can be expressed *both* in terms of internal representations (e.g., world models) and external factors (e.g., communicative chains). I will argue that the internal representations emerge naturally from use.[7] Some concepts, e.g., *color*, are naturally fixed through world models; others, e.g., proper names, more naturally through communicative chains.

2 Five Positions

An **LLM** is, for our purposes, a neural network fitted to predict text in context – typically the next word given the preceding context, in which case the LLM is called *autoregressive*. Such a neural network is a complex mathematical function. If the function is sufficiently expressive, it may memorize the corpus; if not, it will have to compress the information somehow, e.g., by modeling it. A **chatbot** is an autoregressive LLM specialized in (that is, fine-tuned on) two-way dialogue, sometimes also from high-level feedback from human users. Chatbots are thus also complex mathematical functions, but their coefficients may have been altered a bit to make them more suitable conversation partners or personal assistants. **Syntax** is patterns of which word forms or morphemes follow which word forms or morphemes in a language. As such, LLMs are first and foremost trained to perform a syntactic task. The question is whether semantics falls out of that. **Inferential semantics** is the part of semantics that concerns the relations between snippets of text. When LLMs are evaluated on text-only data, we are concerned with inferential semantics. In contrast, **referential semantics** concerns the reference-fixing relationship between words or texts and (entities and events in) the world (or some representation thereof). Referential semantics, in other words, concerns the *aboutness* of language. For a machine to understand a human language it must possess concepts which are coreferential with at least some commonly-used words in that language (Butlin, 2021). We generally do not have direct access to entities and events in the world, but some representation thereof, i.e., the world *für Uns*. I refer to referential semantics with respect to some such representation as *weak* referential semantics – whereas I refer to referential semantics with respect to the world *an Sich* as strong referential semantics.

P1: All Syntax, No Semantics LLMs are trained with a syntactic objective (or, at least, a *syntactically defined* objective), i.e., predicting the next token, word, or sentence, or predicting a missing word mid-sentence. The end result is a complex function with a predefined number of coefficients. Neither the coefficients or the neural layers they form come equipped with semantic translation functions. That P1 follows naturally from these two facts, has been proposed by Searle (1980) in his response to the Connectionist Reply to the Chinese Room Argument, as well as by Bender and Koller (2020) in their Octopus Argument .[8]

Consider the argument: *The LLM A has a purely syntactic objective.* $\rightarrow$ *The LLM A is syntax all the way down.* What exactly is meant by (pure) syntax? Syntax concerns the distribution of forms. Syntax governs how sequences of tokens are composed. Regardless of our exact definition of syntax, it is clear that we need to make a hidden premise explicit for the above argument to go through. Churchland and Churchland (1990) propose the following premise will do the job:

> 'Syntax by itself is neither constitutive of nor sufficient for semantics.'

This premise, of course, is an empirical claim. It may be true, it may be false. P1 therefore comes down to whether something whose proper function is syntax, can reliably induce semantics. Proponents of P1 say no. Proponents of P2 say yes, but only with respect to *inferential* semantics. P3 and P4 say yes and include referential semantics, too.

P2: Inferential Semantics Only Inferential semantics identifies the meaning of an expression with its relationship to other expressions. Following Marconi (1997), I will assume that seman-

[7] This observation echoes the good regulator theorem (Conant and Ashby, 1970), but in artificial intelligence, the idea goes back to Elman (1990) and was recently explicated by Ha and Schmidhuber (2018) and Gurnee and Tegmark (2024).

[8] John Searle famously wrote that 'no program, by itself, is sufficient for intentionality' (p. 424); Bender and Koller (Bender and Koller, 2020) writes that 'learning meaning requires more than form' (p. 5193).

tics comes with inferential and referential properties. Piantadosi and Hill (2022) and Rapaport (2002) adopt this view. Piantadosi and Hill (2022) claim that LLMs induce inferential semantics, but *not* referential semantics; 20 years earlier, Rapaport (2002) said the same of NLP in general. Proponents of Position 1 would, if hard pressed, probably also subscribe to this view. Why? It is simply odd to deny that LLMs induce inferential semantics. These models exhibit human-level-and-beyond performance on textual entailment, synonymy and hyponymy detection, relation extraction, analogical reasoning, etc. (Lewkowycz et al., 2022). They clearly seem to acquire adequate inferential semantics. Proponents of P1 would argue, however, that this is all syntax and no (proper) semantics. For P2, the question thus becomes: Are the inferential capacities of LLMs best described as syntactic or semantic? Note how this question is orthogonal to the robustness of what is learned. LLMs may learn spurious correlations or pick up on signals that are causally relevant. However, both spurious and relevant signals can be syntactic or semantic. So how do we tell them apart?

Having inferential semantics amounts to having a *world model*. By world model we mean something that can be represented as a graph over concepts or proxies for concepts. On such an account, a database such as WikiData or Princeton WordNet can be a world model. The vector space induced by a neural network can also be a world model. In order for such a graph to be a world model, it needs to be *broad-coverage*, and its relations have to be somewhat *interpretable*. By *broad-coverage*, we mean that a world model has to cover a sizeable chunk of our day-to-day concepts. There is no principled cut-off, of course, but in the same way we would not call a list of two word pairs [⟨ dog, Hund⟩, ⟨ cat, Katze⟩] an English-German dictionary; a world model must facilitate more than just a few inferences. Note that it is the broad-coverage requirement that guarantees that water clocks cannot be said to understand time; or that the spindles of Watt's centrifugal governor do not understand speed (Bechtel, 1998). By *interpretable*, I simply mean that some fraction of the relations map onto recognizable conceptual dimensions, such as temporal distance, spatial distance, IS-A relations, metonymy, etc.

With the idea of world models in place, my intuition should be clear enough. In order to know that buying a car implies buying a vehicle, you need to know that cars are vehicles. You need an ontology to make non-trivial inferences. Proponents of P2 will emphasize how world models emerge from training LLMs. The observation that world models can do that, goes back to early work by Elman (Elman, 1990), and the thesis has seen considerable support over the last two decades (Mikolov et al., 2013; Ha and Schmidhuber, 2018). For recent articulations, see Vulić et al. (2020); Søgaard (2022); Gurnee and Tegmark (2024); Huh et al. (2024). Several proponents of P1 have claimed that LLMs cannot induce world models (Bender and Koller, 2020; Landgrebe and Smith, 2021), but I suspect that this is because they believe the world models induced by LLMs and earlier work somehow do not qualify as *proper* world models (whatever that means).[9]

P3: Referential Semantics through Mapping The convergence of world models facilitate alignment (Huh et al., 2024) across different languages or different modalities, e.g.: a) LLMs exhibit very similar geometries across languages (Vulić et al., 2020); b) LLMs exhibit very similar geometries across LLM architectures (Hartmann et al., 2018; Jha et al., 2025); c) LLMs exhibit very similar geometries across modalities (Li et al., 2023b). If the geometries are sufficiently similar, they can be aligned by unsupervised means (Artetxe et al., 2018a; Xu et al., 2018). This facilitates what I call *weak* referential semantics. It is important to understand what exactly is being posited to hold between LLM representations and external representations (from other LLMs, computer vision models, or cognitive data). Geometries are induced by the word representations learned over time by LLMs. Geometries can be similar in a first-order or second-order sense. Similarly, representations can be first-order or second-order isomorphic. First-order similarity compares the pairwise representations: Does the representation for 'dog' share any properties with $[\![dog]\!]$? P3 does not subscribe to there being any such similarities between the sign and the signified, only to second-order similarities across how concepts are organized (Abdou et al., 2021; Merullo et al., 2023a; Søgaard, 2023; Li et al., 2023b). The vector representations do

[9] Proponents of P1 may, for example, ask that world models are grounded in proper functions or are validated with peers. Some teleo-semanticists and some Griceans would have it this way. The question is what this extra baggage buys us, and whether it tracks with our intuitions about what systems exhibit understanding.

not have interesting properties on their own, but the geometries they induce do. These second-order similarities are explained by the phenomenon often noted by proponents of P2, i.e., that world models emerge from training LLMs. The extra claim made by proponents of P3 is that this is, at least in principle, sufficient to get referential semantics.

P4: Referential Semantics through Communicative Chains If my name was James, it would likely not be because I fit the description of a James, but because I was once baptized and people around me picked up on the name. Is semantic understanding of a word in the same way a matter of being in the right communicative chain rather than of forming the right idea? While engineers have long inspected the internal representations of LLMs in search for world models, biases, or explanations – several philosophers have – inspired by recent trends in philosophy of mind (Burge, 1989; Stalnaker, 1990) – voiced concern that this is perhaps too limited a view, and that researchers ought to consider instead more external factors (Sullivan, 2022; Cappelen and Dever, 2021; Butlin, 2021; Mollo and Millière, 2023; Mandelkern and Linzen, 2023). These authors often refer to the so-called Teleological Argument against similarity-based accounts of semantic understanding; see §3.[10]

P5: Referential Semantics for Chatbots Only LLMs generally do not interact with human end users during training. They interact with developers and (very large volumes of) text. The developers implement learning architectures, set hyperparameters and evaluate the performance of LLMs across benchmarks and performance metrics. The text is predicted, sometimes memorized, sometimes compressed. Some philosophers have felt that this was insufficient to get semantic understanding off ground (Butlin, 2021; Mollo and Millière, 2023; Baggio and Murphy, 2024). Chatbots are different, though. They rely on LLMs, but have been fine-tuned in a feedback loop with human conversation agents. They are wheels in the causal chains that fix reference on some accounts; or, in Butlin's words, they are involved in the 'language-mediated concept-sharing described by social externalism about mental content' (Butlin, 2021).

[10] See also Carnap (1967), §154.

3 Pairwise Comparisons

I consider on what grounds we would prefer one position over another, taking stock of the evidence that is currently available.

P1 and P2: A Continuum? What is the test that discriminates between P1 and P2? What would have to be true for P2 to hold, and not P1? Or vice versa. Syntax and inferential semantics are sometimes presented as distinct (Higginbotham, 1987), sometimes as forming a continuum (Rapaport, 1995). Syntactic categories are distillations of semantic ones, and our inferences are biased by selectional restrictions. The debate around semantic understanding in LLMs has often focused on world models. Is this a test for inferential semantics?

> P1-P2 TEST An LLM A has inferential semantics (falsifying P1) iff A has induced a world model.

It is not trivial to decide whether an LLM has induced a world model or not. An LLM that has learned to distinguish verbs from nouns, may look very much like a model that has learned to distinguish processes from things. Syntactic distinctions – even phonetic ones – correlate with the distribution of what is normally considered semantic features, such as animacy or gender. The debate between P1 and P2 may not reduce to a war of words, but even then, it may not be easily settled.

P2 and P3: Evidence for Alignability I adopted the distinction between inferential and referential semantics to get us off ground. How important referential semantics is, is an empirical question. After all, 'many terms that are meaningful to us [...] have no discernible referent at all' (Piantadosi and Hill, 2022). *To the extent* referential semantics plays an important role, we must ask whether LLMs can instantiate referential semantics. In other words: Is the world knowledge induced by LLMs sufficiently reliable that it can be used to ground semantic understanding?

Several applications of LLMs suggest their world models *are* reliable enough to facilitate weakly supervised or unsupervised alignment, including unsupervised machine translation (Artetxe et al., 2018b; Vulić et al., 2020), vision-and-language models (Li et al., 2023b), knowledge graph embeddings (Christiansen et al., 2023), and brain decoding (Mitchell et al., 2004; Li et al.,

2023a). Common to all these studies is the observation that LLMs induce representations that are structurally similar to independently obtained representations of other languages, other modalities, knowledge bases, or neural response measurements; see Appendix A.1.

Now wait a minute. We already know – from the evidence for P2 – that LLMs encode world knowledge, so what's new here? What's new is the existence and quality of easy-to-learn linear mappings between the vector of LLMs and possible referent spaces, e.g., image representations or neural response measurements. If these mappings are sufficiently accurate, they allow for high-quality decoding. That is, they allow for semantic understanding. Since Newman (1928) and Carnap (1967), philosophers have routinely argued that structural similarities, even isomorphisms, are insufficient for semantic understanding. Their arguments revolve around the alleged triviality of isomorphisms, the asymmetry of meaning-grounding representations, as well as the role played by adaptivity in biological systems. Some of these arguments are more challenging than others,[11] but I will briefly address the role played by adaptivity or *proper function*. Shea (2018), for example, talks about whether a structural correspondence is 'used by the system' (p. 119) or not. An *exploitable structural correspondence* is a correspondence between two relations A (on representations) and B (on real-world entities) such that a system S is systematically sensitive to A, and B is of significance to S. Clearly, an LLM (S) is sensitive to relations (A) between representations. An LLM is systematically sensitive to A and relies on it for inference (Garneau et al., 2021; Merullo et al., 2023b), and its success turns on its ability to make inferences about B. Structural correspondences between LLMs and perceptual or cognitive spaces are, on Shea's definition, *exploitable*. Exploitable structural correspondences need to satisfy one more condition to be sufficient for semantic understanding: The structural correspondences need to play an unmediated role in explaining the system's performance. This is an empirical question about LLMs, and one that, fortunately, already has been investigated. Several studies have shown that the more structurally similar LLMs are to each other across languages, or to knowledge bases, the better they perform on downstream tasks (Vulić et al., 2020; Garneau et al., 2021; Merullo et al., 2023b). To Shea, such 'unmediated explanatory structural correspondence is a sufficient condition for having content' (exhibiting semantic understanding). Shea (2018) discusses how the unmediated explanatory functions form a proper subset of the exploitable ones; exploitability is too loose a criterion and instead proposes that 'unmediated explanatory' structural similarity can facilitate referential semantics. Unmediated explanatory functions include, but are not limited to, proper functions. One challenge for Shea, pointed out by Egan (2020), is that the definition of what is unmediated explanatory, seems pragmatic, appealing to Occam-like simplicity. For Egan, it is hard to see how pragmatic considerations can play a content-determining role in a naturalistic theory of representation. Egan therefore finds the set of unmediated explanatory functions too permissive, preferring instead proper functions. On his view, there are too many objective correlations that might be appealed to in framing causal-explanatory generalizations of A's performing the task function. We derive the following empirical test:

> P2-P3 TEST A has referential semantics (falsifying P2) if A has induced a world model that is sufficiently rich to align the two, *by virtue of unmediated explanatory or proper functions*.

Saying LLMs can establish reference through structural similarity by virtue of proper functions, implies two things (Rubner, 2022), namely, that LLMs can establish reference by structural similarity that is a result of the function of LLMs, not a mere accident; and LLMs can malfunction with severe performance costs, when structural similarity is disturbed. Cosine distances in LLM vector spaces *are* directly used by LLMs to draw inferences. Not just nearest neighbor inferences, but also to draw analogies. Such inferences have been demonstrated multiple times (Merullo et al., 2023b; Wijesiriwardene et al., 2023, 2024). Merullo et al. (2023b) showed that LLMs use simple vector arithmetic in order to encode abstract relations. That is, they use vector offset to make inferences, especially in out-of-distribution contexts. What this means, is that LLMs make active use of their vector

[11] P3 proponents can refute Newman's objection, since isomorphisms between the representations of LLMs and what they refer to, are non-trivial, because the metric (cosine distance) is fixed. Moreover, world models in LLMs are homomorphisms at best, so they do not inherit the problems of isomorphisms discussed elsewhere in the literature. Newman's objection generally does not apply to world models, since, by definition, they are graphs over interpretable relations.

spaces, and the properties that make these vector spaces (nearest neighbor) graph isomorphic to perceptual and physical spaces. Well, but the proper function of LLMs is to predict the next word, no? In a sense, yes. But the representations induced are a direct result of this objective,[12] and their structure serves next-word prediction directly. Finally, structural similarity with the physical-social does indeed seem to track performance (Vulić et al., 2020; Garneau et al., 2021; Merullo et al., 2023b).

P4 and P5: Privileged Chatbots? The Teleological Argument has led many philosophers to explore the Kripkean idea that LLM understanding can be grounded in participation in communicative chains (Cappelen and Dever, 2021; Butlin, 2021; Mollo and Millière, 2023; Mandelkern and Linzen, 2023). Some of these philosophers (Butlin, 2021; Mollo and Millière, 2023) have come to the conclusion that only chatbots can be said to instantiate understanding on such grounds. Butlin (2021) gives chatbots special status: 'this mechanism will work for chatbots, because they will rely very heavily on what their interlocutors say to learn about what is going on in particular parts of the world.' The human feedback in conversation (participation in communicative chains) grounds understanding. Butlin's argument for attributing understanding to chatbots is that chatbots satisfy three criteria of understanding. The first criterion is 'deference to experts'. If a chatbot learns that experts do not take certain entities to fall under a concept, it must be disposed to stop representing those entities as falling under that concept. The second criterion is 'apparent *de jure* coreference', which falls out of linguistic generalization; the third criterion is covariation. However, LLMs also seem to satisfy all three criteria. When experts are generalized to include authors, LLMs are also trained to 'stop representing [inappropriate] entities as falling under [a] concept', derive *de jure* coreference by generalization and compositionality, and learn from correlational information.

The idea that only chatbots have referential semantics, also quickly leads to absurd conclusions. To see this, remember how chatbots are fine-tuned LLMs. Consider the scenario in which the original LLM works really well on dialogue, and no updates are made to its parameters during the fine-tuning stage. Call the original LLM A and the fine-tuned (but fully identical) chatbot B. Butlin now claims that B understands language, but A does not. This is in spite of the fact that they are fully identical, parameter to parameter, connections to connections. Imagine next that someone puts A and B on Github, HuggingFace or another website for LLM sharing. Copies of A and B are soon abundant, but only some (namely the copies of B) can be said to understand language. Other researchers now make copies of the copies, some of which understand language, some of which do not. Even if all these models are identical (to the decimal), only some of them understand language. This, to me, seems absurd. The P4-P5 test would be:

> P4-P5 TEST Only chatbots have externalist referential semantics (falsifying P4) iff there exist an LLM that is not a chatbot *and* has referential semantics *by virtue of external factors*.

but as our thought experiment shows, such a test becomes circular: On P5, an LLM is or is not a chatbot by virtue of the same external factors in virtue of which it would have referential semantics.

P3 and P4: Too Permissive? Too Inexpressive? Cappelen and Dever (2021) call for an externalist account of reference, yet do not discuss LLMs explicitly. They list possible external factors contributing meaning to a risk estimation model: (i) a generic initial neural net is given samples from a large pool of training cases; (ii) each training case has been hand-coded ('high risk' versus 'low risk'), for example by the programmers; (iii) the AI's output for the training case is then compared to the hand coding using some scoring function evaluating how well the AI classified the training case; (iv) that score is then used to update the weightings of the node connections in the neural net. External factors generally come in four flavors: training data, developers and annotators, model selection and evaluation, and human agents. In LLMs, there are no human agents to learn from; in chatbots, there are. Human agents are not a necessary condition for understanding on P4, however. If being part of a casual chain of training with supervision from a programmer or an author is sufficient to understand the meaning of the labels provided by those, irrespectively of what is learned, we are left with a very permissive criterion for machine un-

[12] LLMs induce world models because they do not have enough memory to memorize their training data. If your only job in life is to memorize data, but you do not have enough memory to do so, you need to become good at guessing. That's what a world model is for.

derstanding: Basically, any LLM trained in this way for a sufficient amount of time can be said to understand language. Many things can go wrong when training an LLM. Hyper-parameters can be set poorly, training may suffer from vanishing gradients, the model may get stuck in a poor local optimum, and some architectures may suffer mode collapse. None of this seems to depend on how much they were part of the relevant causal chain during training. Or consider two LLMs A and B trained on exactly the same data, but initialized differently. Both A and B will be able to predict that the next word in 'Benny slams the ____' is 'door', but whereas A (correctly) relies on the association between *slam* and *door*, B (wrongly) relies on an association between *Benny* and *door*. If this was the only door-slamming example in the training data, we would have no (externalist) reason to say that A understands language better than B, even if this intuitively seems to be the case. If being causally connected to text is sufficient for understanding, any LLM can be said to understand. This would also hold for an LLM that has *only* learned to memorize the training data, without any form of generalization, or a model that has learned nothing but the fact that dots are often followed by capital letters. Such a notion of understanding does not track our intuitions about understanding systems.

A related argument against P4 externalism is that being causally connected to external factors, may exert very limited influence. Much of the training data has little influence on the final model, and many properties of the developers and the evaluation methods may have little to no impact either. Will P4 externalism nevertheless force us to keep all this information around because it grounds semantic understanding? Consider the following thought experiment to see what is at stake here: A copy A' of an LLM A is value-aligned (fine-tuned) by one provider to be non-sexist, while another copy of A called A" is value-aligned by another provider to be socialist. For various reasons, the two value-aligned modifications of A – A' and A" – turn out to be identical. Would identical output from the two models now mean different things?

Proponents of P4-5 may brush this off as a case of external factors not exerting causal influence on A'. This, however, leads to apparent absurdity. If A' does not exhibit semantic understanding, because the external factors exerted insufficient influence on A', it means there are models A' and A" that differ in infinitely small ways (say in one parameter out of a billion, and only on the 100th decimal) such that A" has semantic understanding, but A' does not.

We turn to a third challenge to P4 externalism: LLMs – like other neural networks – are pseudo-randomly initialized (Aggarwal, 2018), and learning amounts to using data to identify a local optimum relative to the network's pseudo-randomly chosen initial state. LLMs are *not* guaranteed to converge to a global optimum. In the absence of strong convergence guarantees, a network's final state thus depends on its initial state. If a system's state depends on pseudo-random initialization, it is under-determined by external factors. I call this argument the Initialization Argument and discuss it in more detail in the Appendix A.2.

Whether co-occurrence statistics is sufficient to reliably induce the privileged status of deictic expressions, say, is an open, empirical question. Given the jury is still out, it seems that P6 has an upper hand over P3 in leaving the door open for external factors exerting influence on semantic understanding in LLMs.

The question is what factors semantic understanding turns on. If the set of appropriate factors are exclusively internal, e.g,. grammars, world models, projections, this would suggest P3 is on the right track; if the factors are mostly external, P3 is not. P6 is the more plausible position if we find both sets of factors contribute. We suggested above that this is likely to be the case, and we therefore feel we have good reason to take P6 seriously.

> P3-P4 TEST A has referential semantics by virtue of external factors only (falsifying P3) if A has referential semantics, but no explanatory world model.

What it means to have an explanatory world model can be defined more rigorously; see the idea of *unmediated exploitable structural correspondence* in Shea (2018) for one such attempt. One common alternative is to rely on proper functions (Egan, 2020), but this is complicated by the ambiguity of that concept.[13]

[13]The proper function of humans is taken to be set of adaptive traits passed on by our genome (Griffiths, 1993; Matthewson, 2020). The proper function of artefacts is what they were designed for (Preston, 1998). This creates an ambiguity around LLMs, which are trained with a specific objective function (next token prediction), but are also selected (in a more evolutionary sense) because they have certain capabilities.

4 Conclusion

We identified positions on LLMs and semantics: P1–P5. We claimed there are reasons to prefer P2 over P1 (world knowledge in LLMs); and reasons to prefer P3 over P2 (non-trivial structural similarities with referent spaces). We showed P5 to be inconsistent, and that there are some, though not strong, reasons to prefer P4 over P3. Finally, we introduced a sixth, more agnostic position, P6, which permits referential semantics by virtue of both internal and external factors. We did *not* discuss all (10) combinations of positions, e.g., P2-P4.[14]

Limitations

The paper has surveyed positions at the interface of language modeling and philosophy of mind. The relevant publications are scattered across technical and non-technical venues, and we may of course have missed important ones. It was not our intention to cover all publications, but most, if not all, positions. Our discussion is grounded in a relatively uncontroversial theory of linguistic semantics, but of course, alternatives exist.

Acknowledgements

This work was funded by a Carlsberg Semper Ardens Advance Grant. Thanks to Iwan Williams for feedback.

References

Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. 2021. Can language models encode perceptual structure without grounding? a case study in color. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 109–132, Online. Association for Computational Linguistics.

Charu C. Aggarwal. 2018. *Neural Networks and Deep Learning*. Springer, Cham.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018b. Unsupervised neural machine translation. In *International Conference on Learning Representations*.

Giosue Baggio and Elliot Murphy. 2024. On the referential capacity of language models: An internalist rejoinder to Mandelkern & Linzen. *Preprint*, arXiv:2406.00159.

William Bechtel. 1998. Representations and cognitive explanations: Assessing the dynamicist challenge in cognitive science. *Cognitive Science*, 22(3):295–317.

Emily M. Bender, 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, page 610–623, New York, NY, USA. Association for Computing Machinery.

Emily M. Bender 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, Online. Association for Computational Linguistics.

Tyler Burge. 1989. Individuation and causation in psychology. *Pacific Philosophical Quarterly*, 70(4):303–22.

Patrick Butlin. 2021. Sharing our concepts with machines. *Erkenntnis*, pages 1–17.

Herman Cappelen and Josh Dever. 2021. *Making AI Intelligible: Philosophical Foundations*. New York, USA: Oxford University Press.

Rudolf Carnap. 1967. *The Logical Structure of the World and Pseudoproblems in Philosophy*. Routledge K. Paul, London,.

Jonathan Gabel Christiansen, Mathias Gammelgaard, and Anders Søgaard. 2023. Large language models partially converge toward human-like concept organization. In *NeurIPS 2023 Workshop on Symmetry and Geometry in Neural Representations*.

Paul M. Churchland and Patricia S. Churchland. 1990. Could a machine think? *Scientific American*, 262 1:32–7.

Roger Conant and Ross Ashby. 1970. Every good regulator of a system must be a model of that system †. *International Journal of Systems Science*, 1(2):89–97.

Frances Egan. 2020. Content is pragmatic: Comments on Nicholas Shea's Representation in Cognitive Science. *Mind and Language*, 35(3):368–376.

[14] Baggio and Murphy (2024) do this, for example, presenting general arguments against the externalist program as a general explanation for referential semantics, eventually advocating for P2. Baggio and Murphy do seem to agree that the externalist story is on track for certain parts of speech. My story departs from theirs in accepting empirical evidence that LLMs induce world models that are rich enough to facilitate alignment with other representations of the world, i.e., a weak form of referential semantics, turning on internal causes.

Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive Science*, 14:179–211.

Mathias Lykke Gammelgaard, Jonathan Gabel Christiansen, and Anders Søgaard. 2023. Large language models converge toward human-like concept organization. *Preprint*, arXiv:2308.15047.

Nicolas Garneau, Mareike Hartmann, Anders Sandholm, Sebastian Ruder, Ivan Vulic, and Anders Søgaard. 2021. Analogy training multilingual encoders. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12884–12892.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Paul E. Griffiths. 1993. Functional analysis and proper functions. *British Journal for the Philosophy of Science*, 44(3):409–422.

Wes Gurnee and Max Tegmark. 2024. Language models represent space and time. *Preprint*, arXiv:2310.02207.

David R Ha and Jürgen Schmidhuber. 2018. World models. *ArXiv*, abs/1803.10122.

Mareike Hartmann, Yova Kementchedjhieva, and Anders Søgaard. 2018. Why is unsupervised alignment of English embeddings from different algorithms so hard? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 582–586, Brussels, Belgium. Association for Computational Linguistics.

Michael Townsen Hicks, James Humphries, and Joe Slater. 2024. Chatgpt is bullshit. *Ethics and Information Technology*, 26(2):1–10.

J. Higginbotham. 1987. The autonomy of syntax and semantics. In Jay L. Garfield, editor, *Modularity in Knowledge Representation and Natural-Language Understanding*. MIT Press.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. Position: The platonic representation hypothesis. In *Forty-first International Conference on Machine Learning*.

Rishi Jha, Collin Zhang, Vitaly Shmatikov, and John X. Morris. 2025. Harnessing the universal geometry of embeddings. *Preprint*, arXiv:2505.12540.

Karthikeyan K and Anders Søgaard. 2021. Revisiting methods for finding influential examples. *Preprint*, arXiv:2111.04683.

Antonia Karamolegkou, Mostafa Abdou, and Anders Søgaard. 2023. Mapping brains with language models: A survey. *Preprint*, arXiv:2306.05126.

Steffen Koch. forthcoming. Babbling stochastic parrots? a kripkean argument for reference in large language models. *Philosophy of Ai*.

Jobst Landgrebe and Barry Smith. 2021. Making ai meaningful again. *Synthese*, 198(March):2061–2081.

Harvey Lederman and Kyle 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.

Aitor Lewkowycz, Ambrose Slone, Anders Andreassen, Daniel Freeman, Ethan S Dyer, Gaurav Mishra, Guy Gur-Ari, Jaehoon Lee, Jascha Sohl-dickstein, Kristen Chiafullo, Liam B. Fedus, Noah Fiedel, Rosanne Liu, Vedant Misra, and Vinay Venkatesh Ramasesh. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. Technical report.

Jiaang Li, Antonia Karamolegkou, Yova Kementchedjhieva, Mostafa Abdou, Sune Lehmann, and Anders Søgaard. 2023a. Large language models converge on brain-like word representations. *Preprint*, arXiv:2306.01930.

Jiaang Li, Yova Kementchedjhieva, and Anders Søgaard. 2023b. Implications of the convergence of language and vision model geometries. *Preprint*, arXiv:2302.06555.

Sanae Lotfi, Marc Anton Finzi, Yilun Kuang, Tim G. J. Rudner, Micah Goldblum, and Andrew Gordon Wilson. 2024. Non-vacuous generalization bounds for large language models.

Xingyu Lu, Xiaonan Li, Qinyuan Cheng, Kai Ding, Xuanjing Huang, and Xipeng Qiu. 2024. Scaling laws for fact memorization of large language models. *Preprint*, arXiv:2406.15720v1.

Holger Lyre. 2024. "understanding ai": Semantic grounding in large language models. *Preprint*, arXiv:2402.10992.

Matthew Mandelkern and Tal Linzen. 2023. Do language models refer? *Preprint*, arXiv:2308.05576.

D. Marconi. 1997. *Lexical Competence*. A Bradford book. Bradford Book.

John Matthewson. 2020. Does proper function come in degrees? *Biology and Philosophy*, 35(4):1–18.

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. 2023a. Linearly mapping from image to text space. In *The Eleventh International Conference on Learning Representations*.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. 2023b. Language models implement simple word2vec-style vector arithmetic. *Preprint*, arXiv:2305.16130.

Eric J Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. 2023. The quantization model of neural scaling. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Tom M. Mitchell, Rebecca Hutchinson, Radu S. Niculescu, Francisco Pereira, Xuerui Wang, Marcel Just, and Sharlene Newman. 2004. Learning to decode cognitive states from brain images. *Mach. Learn.*, 57(1–2):145–175.

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

M. H. A. Newman. 1928. Mr. russell's causal theory of perception. *Mind*, 37(146):26–43.

Ellie Pavlick. 2023. Symbols and grounding in large language models. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 381.

Steven Piantadosi and Felix Hill. 2022. Meaning without reference in large language models. In *NeurIPS 2022 Workshop on Neuro Causal and Symbolic AI (nCSI)*.

Beth Preston. 1998. Why is a wing like a spoon? a pluralist theory of function. *Journal of Philosophy*, 95(5):215.

William J. Rapaport. 1995. Understanding understanding: Syntactic semantics and computational cognition. *Philosophical Perspectives*, 9:49–88.

William J. Rapaport. 2002. Holism, conceptual-role semantics, and syntactic semantics. *Minds and Machines*, 12(1):3–59.

Andrew Rubner. 2022. Normal-proper functions in the philosophy of mind. *Philosophy Compass*, (7):1–11.

John R. Searle. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3:417–424.

Nicholas Shea. 2018. *Representation in Cognitive Science*. Oxford University Press.

Anders Søgaard. 2022. Understanding models understanding language. *Synthese*, 200(6):1–16.

Anders Søgaard. 2023. Grounding the vector space of an octopus: Word meaning from raw text. *Minds and Machines*, 33(1):33–54.

Robert Stalnaker. 1990. Mental content and linguistic form. *Philosophical Studies*, 58(1-2):129–46.

Emily Sullivan. 2022. Understanding from machine learning models. *British Journal for the Philosophy of Science*, 73(1):109–133.

Lisa Miracchi Titus. 2024. Does chatgpt have semantic understanding? a problem with the statistics-of-occurrence strategy. *Cognitive Systems Research*, 83:101174.

Ivan Vulić, Sebastian Ruder, and Anders Søgaard. 2020. Are all good word vector spaces isomorphic? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3178–3192, Online. Association for Computational Linguistics.

Thilini Wijesiriwardene, Ruwan Wickramarachchi, Bimal Gajera, Shreeyash Gowaikar, Chandan Gupta, Aman Chadha, Aishwarya Naresh Reganti, Amit Sheth, and Amitava Das. 2023. ANALOGICAL - a novel benchmark for long text analogy evaluation in large language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3534–3549, Toronto, Canada. Association for Computational Linguistics.

Thilini Wijesiriwardene, Ruwan Wickramarachchi, Aishwarya Naresh Reganti, Vinija Jain, Aman Chadha, Amit Sheth, and Amitava Das. 2024. On the relationship between sentence analogy identification and sentence structure encoding in large language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 451–457, St. Julian's, Malta. Association for Computational Linguistics.

Ruochen Xu, Yiming Yang, Naoki Otani, and Yuexin Wu. 2018. Unsupervised cross-lingual transfer of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2465–2474, Brussels, Belgium. Association for Computational Linguistics.

A Appendix

A.1 Converging Representations

Unsupervised machine translation is a type of machine translation that does not rely on human translations to train translation models. Instead, they rely on unsupervised alignment of representations from neural LLMs, which provides a basis for then learning a translation model from monolingual corpora in the source and target languages, letting translation models in opposite directions bootstrap each other.

Unsupervised alignment was preceded by weakly supervised alignment. The idea that given just a small number of translation pairs, say the numbers from 1–50 or just a few dozen loan words, was sufficient to obtain alignments from which you could learn bilingual dictionaries or word-by-word

translation models (later to be improved by bootstrapping). How would this be possible? The short answer is: If the geometries of the word embedding spaces were sufficiently similar. If the geometries were perfectly isomorphic, for example, it would require only a small set of translation pairs to induce a perfect alignment, just like in point set registration in computer vision.

The move from weak supervision to *no* supervision was first made possible by generative adversarial networks (Goodfellow et al., 2014). The basic idea behind generative adversarial networks is to train two neural networks: a generator and a discriminator. The generator is trained to create so-called *fake* synthetic samples that are as similar as possible to *real* (actual) samples from a given dataset, while the discriminator is trained to distinguish between real and fake samples. The two networks are trained together in a process known as adversarial training, where they compete against each other to improve their performance.

The two-player training algorithm of generative adversarial networks works as follows: (i) Initialize the generator and discriminator networks with random weights. (ii) Train the discriminator on a batch of real samples from the dataset and a batch of fake samples generated by the generator. The discriminator is trained to output a high probability for real samples and a low probability for fake samples. (iii) Train the generator to fool the discriminator by generating samples that the discriminator classifies as real. The generator is trained to minimize the discriminator's output on the generated samples. (iv) Repeat steps 2 and 3 for a number of epochs, gradually increasing the complexity of the generated samples as the generator learns to produce more realistic outputs. The training process continues until the discriminator can no longer distinguish between real and fake samples, and the generator produces samples that are indistinguishable from real samples. At this point, the generative adversarial network has learned to generate realistic samples that capture the statistical structure of the dataset.

Li et al. (2023b) reused techniques for unsupervised machine translation to investigate whether unsupervised cross-modal alignment from vision to text is possible. The two sets of models rely on very different neural architectures, disjoint data, and orthogonal task definitions, but nevertheless exhibited structural similarity enabling robust alignment with very little supervision.

Knowledge graph embeddings are numerical representations of entities and relationships in a knowledge graph. A knowledge graph is a structured representation of knowledge that captures information about entities (such as people, places, and things) and their relationships. Knowledge graph embeddings aim to encode the information contained in a knowledge graph into low-dimensional continuous vectors or embeddings. These embeddings capture the semantic meaning and contextual relationships between entities and relationships in the graph and can be directly compared to representations in LLMs. Recent work (Christiansen et al., 2023) has shown LLMs also converge toward high levels of structural similarity with such representations.

Similar experiments can be done for neural response measurements and LLM representations. Brain decoding refers to attempts to do this in the direction from brain scans to LLMs. Several studies have been presented in support of the idea that linear mappings enable brain decoding (Karamolegkou et al., 2023). While previous studies are somewhat compromised by poor performance metrics, the evidence accumulated across studies in recent years suggest non-trivial similarities between activation patterns and LLM representations.

The picture that emerges is this: LLMs converge – as they grow bigger and better – toward representations that can relatively easily be aligned with representations found elsewhere: in LLMs for other languages, in computer vision models, in knowledge graphs, and in neural response measurements. Now what do the representations of all these systems have in common? They are representations of the world for us (Huh et al., 2024).

A.2 The Initialization Argument

Let us first try to flesh out the Initialization Argument a bit. The first premise, which we take for granted, is that random initialization is part of training an LLM (or, more generally, a deep neural network). This is simply common practice (Aggarwal, 2018). The second premise is that the random initialization is not completely determined by external factors. This follows directly from the fact that the parameters are determined in an internal process from a pseudo-random number that depends (in unpredictable ways) on a single, manually spec-

ified seed value.[15] The third premise is that the random initialization has impact on the network's final state. This premise is not immediately obvious, and we return to it in a bit. We then take it that it holds in general that if A is part of a process B (I call this proposition ϕ_1), and A is not completely determined by a set of factors F (proposition ϕ_2), then B is not completely determined by a set of factors F (proposition ψ). This follows from the fact that the property of being 'in part determined by' is inherited by processes from their subprocesses. In sum, the three premises are: (1) Training an LLM begins with random initialization. (2) Random initialization is under-determined by external factors. (3) Random initialization has causal influence on the final model state. Here's an auxiliary premise we take for granted: If A is part of a process B (ϕ_1), and A is not completely determined by a set of factors F (ϕ_2), then B is not completely determined by a set of factors F (ψ). The first premise means we can replace A with random initialization and B with the process of training an LLM. The first conjunct ϕ_1 is now our first premise. The second and third premises mean we can replace F with external factors, to obtain ϕ_2, saying that random initialization is not determined by external factors. Our principle now tells us: $\phi_1 \wedge \phi_2 \rightarrow \psi$, it now follows that the training of an LLM is not completely determined by external factors. This shows that if the three premises (1–3) are true, a purely externalist account of LLMs will not provide a complete and adequate description.

One may respond that while the training of an LLM may not be completely determined by external factors, maybe its final state is. In other words, one may question our third premise. This premise is not *a priori* true. If the initialization is immaterial, and the convergence guarantees of the training procedure meant we always would end up with the same network, regardless of how it was initialized, this would, for now, vindicate a pure externalist account. Unfortunately (for such an account), LLM training *is* very sensitive to such initialization. The reason for this is that the local optimization algorithms used to train LLMs in complex manifolds, only guarantee convergence to local optima, not global ones. If all you know is how to climb to the nearest mountain top, it matters a great deal for where you end up, *where* in the Himalayas you are dropped off. In the same way, the final state of an LLM depends very much on how it is initialized. A better response, perhaps, is that the final states, while different across different initializations, need not be relevantly different. There is some work suggesting that models trained with different random seeds very often end up learning structurally similar representations of the data (Hartmann et al., 2018). The externalist could cite such results and argue that to the extent reference is fixed by second-order similarities, models over different seeds are *not relevantly* different. Ironically, the externalist is vindicated by internalist tools. The sensitivity to random initialization *does* have (other) profound effects, however. K and Søgaard (2021), for example, report that the sensitivity of leave-one-out and influence function estimates of training data influence on LLMs is extremely sensitive to random initialization. Specifically, the (average) Pearson correlation across the influence estimates of two randomly initialized models, trained in exactly the same way, is very small (2–4%).

Could externalists not argue – in response to the Initialization Argument – that random influence is random, and therefore beyond the reach of both internalists and externalists? Such a response, however, would be missing the point. For the externalist who wants to explain how a speaker fixes the reference of 'Bertrand Russell', or whether an artificial intelligence system relies on a spurious correlation or not, the random influence is invisible. The random influence is internal or leaves only an internal footprint. For the internalist, examining the internal representations of the speaker or system, the random influence is directly accessible.[16]

[15] A die-hard externalist could argue that both the random number generator and the manually defined seed number are part of the relevant causal claim. If this is a bullet externalists are willing to bite, the Initilization Argument fails as an argument against externalism. It is unclear, however, what explanatory roles causal chains that include pseudo-random generators, can play.

[16] Other problems with externalist accounts of LLM understanding include tokenization. The challenge with tokenization, previously discussed in Baggio and Murphy (2024), is that it does not necessarily track the things that have causal histories in language: Tokens, they say, do not necessarily correspond to the kinds of expressions or parts of expressions that can have causal histories, such as characters, morphemes, and 'words' [. . .] (p. 4).