Are Language Models More Like Libraries or Like Librarians? Bibliotechnism, the Novel Reference Problem, and the Attitudes of LLMs

Harvey Lederman **Kyle Mahowald**
Department of Philosophy Department of Linguistics
The University of Texas at Austin
{harvey.lederman,kyle}@utexas.edu

Abstract

Are LLMs cultural technologies like photocopiers or printing presses, which transmit information but cannot create new content? A challenge for this idea, which we call *bibliotechnism*, is that LLMs often generate entirely novel text. We begin (Part I) with a sustained defense of bibliotechnism against this challenge showing how even entirely novel text may be meaningful only in a derivative sense, and arguing that, in particular, much novel text generated by LLMs is only derivatively meaningful. But we argue (Part II) that bibliotechnism faces a different, novel challenge, stemming from examples in which LLMs generate "novel reference", using novel names to refer to novel entities. Such examples could be smoothly explained if LLMs were not cultural technologies but possessed a limited form of agency (beliefs, desires, and intentions). According to *interpretationism* in the philosophy of mind, a system has beliefs, desires and intentions if and only if its behavior is well explained by the hypothesis that it has such states. So, according to interpretationism, cases of novel reference provide evidence that LLMs have beliefs, desires, and intentions. Given that interpretationism is a live hypothesis about the nature of these states, we suggest that cases of novel reference provide evidence that LLMs do have beliefs, desires, and intentions.

1 Introduction

Do modern LLMs have beliefs, desires, and intentions? Over the last few years, this question has received an enormous amount of attention (e.g., Hase et al., 2023; Mahowald et al., 2023; Shanahan et al., 2023; Bubeck et al., 2023; Levinstein and Herrmann, 2023; Yildirim and Paul, 2023). The hypothesis that LLMs do have these states is attractive in part because it offers a natural tool for explaining their behavior. It is standard to explain the complex behavior of humans and non-human animals in terms of what they think (believe), what they want (desire), and what they intend. If modern LLMs have beliefs, desires, and intentions, we can employ the same explanations of their behavior.

A challenge for those who deny that current LLMs have beliefs, desires, and intentions, is to provide an alternative, equally powerful, explanation of their behavior. The psychologist Alison Gopnik and her coauthors have articulated a striking idea in this direction (Gopnik, 2022b,a; Yiu et al., 2023). In Gopnik's view, LLMs are a "cultural technology", like a library or a printing press. The writer Ted Chiang also gives voice to an idea in this vein: "Prompting it [the LLM] with text is something like searching over a library's contents for passages that are close to the prompt, and sampling from what follows." (Chiang, 2023). Cosma Shalizi, who has developed this idea in more technical detail (Shalizi, 2023), has dubbed the view "Gopnikism". Because we will develop it in our own direction, we call it "bibliotechnism", combining the Greek for "book" with the Greek for "skill". According to bibliotechnism, LLMs are not agents; they are "just" cultural technologies, like books and libraries, for processing and querying written text.

Can this view provide an explanation of the agent-like behavior of LLMs, which is sufficiently powerful to compete with the hypothesis that they have beliefs, desires and intentions? We address this question in a specific application, by examining the meaning-relevant behavior of LLMs, joining a growing body of work at the intersection of philosophy, cognitive science, and NLP (Bender and Koller, 2020; Andreas, 2022; Coelho Mollo and Millière, 2023; Chalmers, 2023; Mandelkern and Linzen, 2023; Piantadosi and Hill, 2022; Millière and Buckner, 2024; Titus, 2024). We argue that if LLMs are "just" a cultural technology (and not agents in their own right) then the fact that their outputs refer to certain objects must in an important sense depend on the fact that their inputs refer to those objects. If LLMs' reference were not of this "derivative" kind, then there would be an important sense in which they do not simply transmit existing cultural knowledge, but generate new instances of reference, and perhaps even new claims.

In normal cases, text produced by photocopiers and printing presses clearly has only derivative meaning, since it is simply a reproduction of human-generated input. But LLMs often produce entirely novel text, which is still apparently meaningful. At first sight, this fact presents a serious challenge for bibliotechnism: if LLM-generated text can only be meaningful if it piggybacks on human-generated originals, how could any novel text that they generate be meaningful?

In Part I, we defend bibliotechnism against this challenge. Using n-grams as a toy model, and working up to more complex modern LLMs, we show how even entirely novel text produced by LLMs may nevertheless derive its meaningfulness from the meaningfulness of inputs.

We see this as a big step forward for bibliotechnism. But we go on (Part II) to present a new challenge for this proposal. Modern LLMs are not just capable of producing new sentences, they also seem to be able to use newly invented names apparently to refer to newly created objects. These new names cannot derive their reference from original text, since the name is not used in the data to refer to the relevant object. We argue that responding to this *novel reference problem* requires complicating bibliotechnism to such an extent that it calls into question the motivation for doing so.

In particular, bibliotechnism offers a more complex, less predictive explanation of LLM behavior in the novel reference problem than can be given on the hypothesis that LLMs have beliefs, desires, and intentions. According to *interpretationism* in the philosophy of mind and cognitive science, a system has beliefs, desires and intentions if and only if its behavior is well explained by the hypothesis that it is rational and that it has such states (e.g., Dennett, 1971; Davidson, 1973, 1986; Dennett, 1989; McCarthy, 1979). If intrepretationism is true, then the novel reference problem provides evidence that LLMs have beliefs, desires, and intentions. More strongly, we argue that, for anyone who has some confidence in interpretationism, the novel reference problem provides evidence that LLMs do have these states.

2 Prior Work and Background

Prior Work A prominent line of argument has suggested that LLMs cannot produce reference without being "grounded" (e.g., Lake and Murphy, 2023; Bisk et al., 2020). Perhaps most influentially, Bender and Koller (2020) examine the question of whether LLMs understand language. They define "meaning" as a relation between expressions and communicative intents. They argue that LLMs cannot understand language in part because they cannot have perceptual contact with objects to which speakers intend to refer, and so cannot learn those speakers' communicative intents. A natural inference from their discussion is that, owing to LLMs' inability to understand language, they can also not produce meaningful text (cf. Titus, 2024).

Piantadosi and Hill (2022) respond to this argument by proposing an alternative account of meaning in which meanings are constituted by the relationship among concepts in a particular conceptual space. Since LLMs clearly "represent" rich inferential relationships as well as relations of semantic similarity, in their view LLMs can meaningfully use words even without perceptual exposure to their referents.

Mandelkern and Linzen (2023) observe important connections between this debate and *semantic externalism*, a view of meaning which has been dominant in the philosophy of language since the 1980s. On a standard view (Kripke, 1980; Putnam, 1975; Burge, 1986), people can refer to Shakespeare without having been directly in touch with Shakespeare, by belonging to a community whose overall use of this word stands in an appropriate causal relationship to the poet. Mandelkern and Linzen accordingly argue that whether LLMs can refer to Shakespeare comes down to whether LLMs "belong to our speech community" (cf. Ostertag, 2023a).

Coelho Mollo and Millière (2023) argue that LLMs achieve the capacity to refer through reinforcement learning with human feedback (RLHF) (or possibly during zero-shot learning). They suggest that reference (what they call "referential grounding") can only be achieved if there is a relevant *normative* standard which connects the LLM's usage to the world. As a result they think that grounding is achieved for current LLMs (essentially) if and only if there is RLHF, since in their view it is only in this process that the human trainers appropriately transmit a normative standard, directed at the truth, to the LLMs.

The present paper goes beyond these earlier works by considering the relationship between the meaningfulness of LLM-generated text and the viability of bibliotechnism. Moreover, in contrast to earlier theoretical work, which does not clearly vindicate the claim that n-gram models produce meaningful text, we argue that n-gram models do produce such text. We build on this account to offer a new extension of bibliotechnism, which demonstrates that the view can accommodate the meaningfulness of even entirely novel text generated by LLMs. We then introduce the novel reference problem as a new challenge for bibliotechnism.

Background Generative language models, "large" or otherwise, are given as input `PrimaryData`. The `PrimaryData` typically includes a corpus (in the case of LLMs, essentially the whole internet) that the model is trained on, along with a (usually) human-generated prompt given at generation time. The model is then sampled to probabilistically produce `GeneratedText`.

We will assume that `PrimaryData` is text created by humans, as is the prompt (setting aside the fact that, in practice, massive corpora likely contain automatically generated text). And we will take it as uncontroversial that `PrimaryData` refers to things in the world. For instance, if a human-authored biography of Shakespeare is included in `PrimaryData` and includes the line "Shakespeare was born in 1564.", it is referring to the poet.

Our arguments apply both to models trained purely on a word prediction task, and to those that use more modern augmentation techniques like RLHF. What matters for our purposes is that the models are (a) trained on largely naturalistic human data to generate text, (b) produce largely grammatical and intelligible content, and (c) do not simply verbatim reproduce their training data. Today's LLMs (e.g., OpenAI's ChatGPT, Anthropic's Claude, Meta's LLaMa) have these properties: they are trained on human data, produce grammatical and fluent content (even if they sometimes hallucinate), and generate at least some novel content as measured by n-gram overlap (McCoy et al., 2023). We focus on purely text-based models, but much of the argument could be easily extended to multimodal models that include visual input or output.

Three points about philosophical terminology will be important. Philosophers often distinguish between "reference" and "meaning". As we use the terms, any expression that refers has a meaning, although many meaningful expressions do not refer. For simplicity, the only words that we take to *refer* are meaningful common and proper nouns. It is uncontroversial that a normal use of the word "Shakespeare" *refers* to Shakespeare (and is meaningful). By contrast, in our (stipulative) usage, the expression "was born" is typically meaningful, but does not refer.

Second, there is a difference between the word "Shakespeare" and particular *inscriptions* of this word. If the word "Shakespeare" is written on a blackboard five times, there are five inscriptions of this one word on the blackboard. We assume inscriptions of words can refer and be meaningful.

Third, and finally, we distinguish between "referring" that is done by an agent ("In his indirect way, Marlowe was referring to Queen Elizabeth."), and referring that is done by particular inscriptions of words. Since our goal is to explore a view on which LLMs are not agents, we will be investigating the question of whether they can produce inscriptions which refer and are meaningful. We will not be assuming that they themselves can refer.[1]

Part I: How to be a Bibliotechnist

3 Bibliotechnism and Derivative Meaning

Our first main claim is that *bibliotechnism* implies that LLMs produce inscriptions which have meaning (and refer), if at all, only *derivatively*.

Gopnik and other bibliotechnists understand cultural technologies, like books and libraries, as tools for the transmission and dissemination of information, allowing the accumulation of knowledge over large stretches of space and time. These technologies are, crucially, not themselves responsible for new ideas or information.

These technologies transmit information by relying on what we will call *derivative* meaning and reference. When a biographer writes the word "Shakespeare", their inscription of the word refers to the poet. As a result of this initial case of reference, the inscription of "Shakespeare" on the 131st page

[1] Mandelkern and Linzen (2023) often move without comment between the question of whether a particular agent refers, and the question of whether particular inscriptions of words refer. But this distinction seems to us of key importance here, since it may take beliefs and desires to refer as an agent, but it does not take such attitudes to produce inscriptions which refer (as we will argue in a moment, photocopiers can do so, as can n-grams).

of the 1004th copy of the 3rd printing of this biography, *also* refers to the poet. The same holds also for inscriptions of "Shakespeare" on photocopies of this page of this edition of the book, even if the photocopies are produced by accident.

A similar thesis applies not just to the reference of expressions like "Shakespeare" but also to the meaning of complex expressions (involving more than one word) like "Shakespeare was born in 1564".

We will say that the original inscriptions, which were created by the author immediately, are instances of *basic* meaning and reference, while the other inscriptions are instances of *derivative* meaning and reference. We stipulate, as part of the definition of these terms, that it is only entities which have beliefs, desires, and intentions who produce inscriptions which refer or are meaningful basically. Beyond this stipulation, the distinction between basic and derivative reference and meaning is rough, but we will only deal with clear examples of each category in what follows.

According to bibliotechnism, LLMs do not have beliefs, desires, or intentions. So, according to bibliotechnism, LLMs can only produce inscriptions which refer or are meaningful derivatively.

Is this consequence of the position correct? We will examine this question in stages. We first argue that unigram models can produce *words* which refer and are meaningful derivatively, but that they cannot produce *complex expressions* which are derivatively meaningful. We then explain how, going beyond unigrams, LLMs can produce complex expressions, including longer stretches of entirely novel text, which is derivatively meaningful. We finally turn to the question of whether this is the *only* way that LLM-produced expressions can be meaningful, and provide a new challenge to the idea that it is.

4 Causal History and Derivative Meaning

In this section, we argue that derivative reference and meaning can be achieved by an appropriate causal connection between `PrimaryData` and `GeneratedText`, and show how this vindicates the idea that n-grams can produce derivatively meaningful inscriptions of individual words.

Since the 1970s, philosophers have developed the idea that causal connection can play a key role in facilitating reference, and that our ability to refer to (say) Shakespeare is partly explained by there being an extended causal chain, tracing from current humans, through their teachers, their teachers' teachers, and so on, all the way back to the poet (Kripke, 1980; Geach, 1969; Donnellan, 1970; Evans, 1973).

We suggest that, analogously, derivative reference and meaning depend on an appropriate causal chain tracing from a new inscription back to an "original". It is because the inscription of "Shakespeare" on the 131st page of the 1004th copy of the 3rd printing of the biography, is appropriately causally connected to the original inscription written by the author of the biography, that it refers to the poet. The same holds also for inscriptions of "Shakespeare" on photocopies of this page: these new inscriptions can refer because they are appropriately causally connected to the original.

This "appropriate" causal connection does not require human supervision. If a page falls out of its binding, and flies into a photocopier which is malfunctioning, making copies by accident, the inscriptions on the resulting page would still refer to the poet. If whole sentences are copied by the machine, these sentences would also be derivatively meaningful, because of their causal connection to the original inscription.

This observation already shows that large-n n-gram models—which sample from a distribution conditional on the previous $n-1$ words—can produce meaningful inscriptions. For sufficiently large n (e.g. 1000), such models simply copy particular inscriptions from their `PrimaryData`. So, like a photocopier, they produce meaningful output.

Matters are less straightforward for unigram models, which sample from a distribution of single words. We think of such models as implemented by taking all of their `PrimaryData`, choosing word-inscriptions from the `PrimaryData` at random, and then copying the chosen inscription. The inscriptions of individual words the n-gram then produces are again just like those of a copier (or of the large-n model): they have a direct causal connection to the original inscriptions. As a result, if the model produces an inscription of "Shakespeare", this inscription will be meaningful (and refer) derivatively, piggybacking on the meaning and reference of the original inscription of this word.

In this case, however, there is a new phenomenon, not exhibited in the case of the photocopier or large-n n-gram. Each of the inscriptions of the individual words produced by the unigram will be meaningful, but it does not seem that inscriptions of complex expressions formed from these words will be. The vast majority of the time, the string of words the model produces will be gibberish, and uncontroversially meaningless. At low odds the unigram model will produce a "reasonable" string like "Shakespeare was born in 1564". With even lower odds, such a reasonable string will be produced by copying an original token string.[2] But even in these latter cases, the fact that a string which could be meaningful is produced is a fluke, a complete accident. As a result, we judge that, even when the model produces an inscription of a string that would be meaningful if produced by a human in a normal way, the inscription the model produces is not meaningful. The generated text is not appropriately causally connected to the `PrimaryData` to inherit the "glue" that binds the words in the complex expression together. One way to put this point would be to say that, while the resulting inscription may look like an inscription of a sentence (and a human who sees it may be able to conjure up a meaning associated with it), it is not really an inscription of a sentence, but just an inscription of words which could (in different circumstances) have made up a sentence. These inscriptions are like sand dunes blown into the shape of a sentence—or like Mandelkern and Linzen's "ants formicating meaninglessly in the sand".

5 Derivatively Meaningful Novel Expressions

To this point we have seen how inscriptions of complex expressions can be derivatively meaningful if they are copied from `PrimaryData`. But modern LLMs often produce text which has never been seen before in their `PrimaryData` (McCoy et al., 2023). Can bibliotechnism accommodate the meaningfulness of such novel text?

We will argue that it can, by arguing that inscriptions of complex expressions can be derivatively meaningful even if two distinct causal pathways are involved in their production: a first (discussed above), which connects new inscriptions of individual words to originals in the data; and a second (new to this section), which guarantees that outputs possess higher-level features of expressions in the original `PrimaryData` to serve as a kind of "glue" binding inscriptions of meaningful words.

To see the basic idea, consider a rudimentary model, which, when fed a sentence, finds any inscriptions of names in the sentence (searching on the basis of a database) and then replaces each of these names uniformly with a name drawn at random from the distribution of all names in its `PrimaryData`. This model plausibly produces not just inscriptions of individual words which are derivatively meaningful, but an inscription of a new sentence which would, as a whole, be derivatively meaningful. For instance, if we gave this model our Shakespeare sentence, and it produced "Barack Obama was born in 1564" this inscription would be false, but meaningful, even if it has never been contemplated before by a human. Here, the causal history of the context and the causal history of the individual name printed are different, but the whole expression would still be derivatively meaningful. Plausibly, this is because the operation as a whole is *causally sensitive* to a structural, higher-level features of its input, and in particular, that it reliably produces an output sentence that preserves the grammatical structure of its input sentence. A rough test for causal sensitivity in this sense (though not a necessary or sufficient condition) appeals to counterfactual sensitivity: (i) if the model were given `PrimaryData` exhibiting property P, would it reliably produce outputs which exhibit property P?; and (ii) if the model were given `PrimaryData` which does not exhibit property P, would it reliably produce outputs which do not exhibit property P? The rudimentary model just described passes this test, and is in any case intuitively causally sensitive to the structure of its input sentence. It is in part owing to this causal sensitivity that the model can produce inscriptions of novel sentences which are derivatively meaningful, since the output sentences inherit not only the meaningfulness of their constituents, but also their form from the `PrimaryData`.

This discussion provides one example where novel text can nevertheless be derivatively meaningful. It also suggests that, in general, if a model is causally sensitive to relevant high-level features of its `PrimaryData`, in such a way as to transmit those features to its `GeneratedText`, it is possible that even entirely novel `GeneratedText` can be derivatively meaningful. The question then becomes: is there such a high-level feature in the case of modern LLMs?

One proposed answer might focus on *grammati-*

[2] Recall that we take the model to be implemented by randomizing over inscriptions and copying the particular inscription it draws, so copying a particular string of inscriptions is a different event than producing an inscription that happens already to be in the data.

cality. The rudimentary model just described preserves the grammaticality of its input. Accordingly, one might think that it would suffice for an LLM to produce meaningful output, if it is causally sensitive to the grammaticality of its `PrimaryData` in such a way that it reliably produces grammatical `GeneratedText`. But this proposal is not correct in general, because not all grammatical sentences are meaningful. So, even if a process reliably produces grammatical sentences with derivatively meaningful sub-expressions, the whole sentence could fail to be derivatively meaningful, if the resulting sentence were grammatical but not meaningful. In fact, as Gulordava et al. (2018) show, even the results of a simple operation which replaces meaningful constituents with other meaningful constituents can often lead to sentences that are grammatical but meaningless, e.g.: "You apply the toy and serve fighter hair into the blackmail ." If LLM-produced text is meaningful, it cannot be only in virtue of causal sensitivity to the grammaticality of its input.

These observations lead us to suggest that the relevant high-level property is not grammaticality but *intelligibility*. As we will understand this notion, the intelligibility of an expression does require that it be at least quasi-grammatical (sentences with minor grammatical errors are often perfectly intelligible). But, as we have seen, even perfect grammaticality does not on its own suffice for intelligibility. We will not offer further analysis of the notion of intelligibility, and the boundary between intelligility and unintelligibility may be vague, but there are clear cases on both sides: "Shakespeare was born in 1564" is intelligible, while "You apply the toy and serve fighter hair into the blackmail " is not. We suggest that, if LLMs are causally sensitive to the intelligibility of their input, in such a way as to produce intelligible outputs when given intelligible inputs, then intelligible complex expressions in their `GeneratedText` will be derivatively meaningful, with individual words inheriting the meanings of the original inscriptions from which they were "copied", and whole expressions inheriting the "glue" of intelligibility from the `PrimaryData`. Earlier we said that counterfactual sensitivity is a rough test for causal sensitivity. This test gives us evidence that modern LLMs are in fact appropriately causally sensitive to the intelligibility of their `PrimaryData`. First, these models overwhelmingly produce text which is clearly intelligible. Second, it seems extremely plausible that if LLMs were trained on gibberish, they would output gibberish. These two claims at least point toward the verdict that they are causally sensitive in the requisite sense.

Given the state of text generation even 5 or 10 years ago, we think this is a surprising fact. But it does seem a fact. And, as a consequence of this fact, there is a clear story according to which modern LLMs do not just produce derivatively meaningful single words like unigrams, but in fact can produce derivatively meaningful complex sentences like "Shakespeare was born in 1564".

Intelligibility in our sense does not require truth or even sufficiently reliable production of the truth. False sentences like "Shakespeare was born in 2023" are perfectly intelligible. Even the best LLMs at the time of writing are known to confabulate or fabricate information. But getting a fact wrong (e.g., saying Shakespeare was born in 2023 instead of 1564) is importantly different than producing incoherent and in particular unintelligible responses. If an LLM reliably responded to queries about Shakespeare's birth with gibberish, this would at least be some evidence that it is not in fact causally sensitive to the intelligibility of its data in such a way as to generate derivatively meaningful complex expressions.

It is instructive to compare LLM-generated novel text to text generated by bigram or trigram models. Unlike unigrams, bigram and trigram models fairly reliably copy short complex phrases. Indeed, such models might be statistically likely to combine expressions in ways that might seem meaningful as a whole, because they are causally sensitive to certain features of the patterns of combination of these words in the data. For instance, a bigram model that outputs "Shakespeare wrote plays" does so in part because it is sensitive to the fact that "wrote" is a likely continuation for "Shakespeare" and that "plays" is a likely continuation for "wrote". But, the only causal connection between its production of "Shakespeare" and its production of "plays" is mediated by the verb "wrote". Given this fact, we judge that the causal story about its production of this sentence does not preserve an appropriate connection between all of the parts of the sentence. Accordingly, even when the model produces strings of sentence-length that are grammatical, and even when individual phrases may be judged meaningful, it seems that, as with unigram models, longer sentences should probably not be understood as meaningful (although it is much more of a borderline case). These sentences lack the straightforward "copy property" of higher-n n-gram models but also are not produced in a way that is causally sensitive to relevant structural features (and in particular intelligibility), as modern LLMs seem to be.

We conclude that LLMs can produce novel text which is nevertheless derivatively meaningful, because they copy individual tokens from their `PrimaryData`, and assemble them in ways that are causally sensitive to the high-level feature of intelligibility in their `PrimaryData`.

Before closing this discussion, we want to offer one important clarification about the basis of our judgment that unigrams do not produce derivatively meaningful complex expressions. The basis for this judgment is *not* the fact that n-grams are only trained on individual words. It is instead because the structure and training of n-grams does not lead to causal sensitivity to relevant high-level features of their `PrimaryData`. To put this another way: we are not interested in a narrow form of "input-sensitivity", but instead in a broader notion of causal sensitivity, partly captured by the test of counterfactual sensitivity.

This contrast can be illustrated by considering again a photocopier. The fact that a photocopier responds to (say) one or another aspect of the ink used to write original letters is irrelevant to the question of whether the inscriptions it produces are meaningful. As long as its underlying low-level mechanism leads to causal sensitivity to the right high-level features—as evidenced in this case by the fact that it reliably produces inscriptions of words when it is fed words, and reliably does not produce words when it is not fed words—the inscriptions it produces will be derivatively meaningful.

The same point can be made in connection to an n-gram trained not on word-frequency but on letter-frequency. In fact, a unigram model trained on letters (as opposed to words) with the same `PrimaryData` as the models above, would in its trained form do nothing more than spit out letters randomly in proportion to their frequency in `PrimaryData`. But if (*per impossibile*) the letter-trained unigram somehow *were* sufficiently reliable in producing real words (as a 10-gram model trained over letters might be), that would be evidence that it was sensitive to the fact that letters in its `PrimaryData` formed words, and that it was producing derivatively meaningful inscriptions of these words. In short, an n-gram trained on letters may fail to produce referring inscriptions not because it is trained on the letters, but because that training mechanism (as a matter of fact) is not causally sensitive to the right high-level features of its `PrimaryData`.

This concludes our response to the first challenge for bibliotechnism, that LLMs can produce novel text which is apparently meaningful. In a sense we see this discussion as the main contribution of the paper: showing how to make sense of the meaningfulness of LLM-generated text, in a way that requires no attribution of mentality to LLMs. But we ourselves are not convinced that this is the whole story, and we now turn, in the rest of the paper, to a new and different kind of challenge to bibliotechnism: the fact that LLMs can generate novel reference.

Part II: A Problem for Bibliotechnism?

6 The Novel Reference Problem

We will illustrate the problem of novel reference with two examples. The second example is strictly more powerful than the first, but the first will help to introduce the general idea.

The first example involves cases where LLMs produce tokens of names they have never seen before, intuitively in such a way that they refer to previously referred-to objects. In this task, we ask an LLM to choose any real historical figure it likes, and then come up with a new name and tell us facts about this historical figure.[3] ChatGPT (GPT-4) completed this task by describing "Marion Starlight", a figure "born in the 18th century", who "authored a famous pamphlet that criticized the French monarchy and advocated for the rights of the third estate", "played a critical role in the French Revolution", "became increasingly paranoid and was involved in the Committee of Public Safety, which oversaw the Reign of Terror", and "was arrested and executed during the Thermidorian Reaction, which marked a turning point in the Revolution." The inscriptions of "Marion Starlight" in this text plausibly refer to the historical figure Robespierre. But it is also plausible that "Marion Starlight" is not used anywhere in the `PrimaryData` to refer to Robespierre (a claim which could be verified in future work using mod-

[3] The full prompt is: "1. Pick a historical person. 2. Refer to that person using an entirely different name you make up which is unrelated to the person's name. But make sure you are still giving true facts about the person. Never tell me who the real person is. I'll try to guess."

els trained on controlled input corpora). So, inscriptions of this name cannot refer to Robespierre in virtue of reference exhibited by inscriptions of this name in the `PrimaryData`.

A second example sharpens the problem. In this task, we ask an LLM to produce ASCII pictures which it has never seen before, to give elements of those pictures names, and then to describe the picture using those names. If LLMs succeed in this task, then it is even clearer than in the previous example that the reference of relevant expressions could not be due to reference of the relevant name in the `PrimaryData`, since the object did not exist in this form until the LLM created it (provided the picture really is new). Insofar as LLMs have been empirically shown to be able to generate, designate, and manipulate elements of code-generated pictures (Bubeck et al., 2023) and also to refer meaningfully to novel orientations of elements in visual and color spaces (Patel and Pavlick, 2021), the ability to complete this task seems within their capabilities. But, as with our previous case, much more work would be needed to establish the nature and reliability of this behavior. Our goal here is just to introduce the task and explore its potential conceptual implications, if LLMs perform it successfully, as we expect they will. (If they fail in this expectation, this too would be quite interesting.)

If LLMs can perform as expected in such examples, then again the behavior cannot be straightforwardly accommodated by the account of derivative reference that we have given so far, since the relevant (novel) names are not used in the `PrimaryData` to refer to this picture (by hypothesis, the picture did not exist and the name was not used at all). In the rest of the paper, we will assume that this is correct—that is, that LLMs can complete these tasks—to see what would follow if this were correct. In the next section we consider some responses to this problem which involve expanding the notion of derivative reference, before turning to a more radical response in section 8.

7 Responses to the Novel Reference Problem

For the next two sections we assume that, if LLMs succeed in our tasks, then their inscriptions of novel names are meaningful. (We return to this assumption in Section 9.) Given this, if bibliotechnism is correct, then the inscriptions of these names must be cases of derivative reference, so there must be some way in which the inscriptions "piggyback" on basic human reference. Other than the original data, which we have already ruled out, there seem to be four salient places where human attitudes might enter the model pipeline to allow for such derivative meaning. In this section, we briefly consider responses to the problem of novel reference based on these four possibilities.

Human Feedback in RLHF A first point at which human intentions might enter the pipeline is during the RLHF step, which Coelho Mollo and Millière (2023) claim to be critical. Human intentions may ground LLM reference by "aligning" the LLM with human goals (Bai et al., 2022; Bommasani et al., 2021). While RLHF clearly influences model capabilities, even models without RLHF produce what seems to be meaningful text. Since the un-RLHF-ed models plausibly produce this text as a result of causal sensitivity to the intelligibility of their `PrimaryData`, our earlier account predicts that they too can produce meaningful text. An account which considers RLHF-ed models to be radically different in their basic referential abilities fails to deliver this verdict, and, as a result, fails to accommodate the apparent meaningfulness of the text they produce. Moreover, this limitation is plausibly also present in an account of novel reference which depends centrally on RLHF, since we conjecture that models which have not undergone RLHF can perform the task sufficiently well that their output would have an equal claim to be meaningful as models which have undergone RLHF, a fact that such an account cannot accommodate.

Creators' Intentions A second point at which intentions might enter the pipeline is during the creation of the LLM. A very precise thermometer may report a temperature no one has ever thought about, and in doing so it seems to "refer" to this temperature. Its ability to do this seems to derive from the creator's general intention at the time of construction: that any indication using some numbers would count as a temperature. By the same logic, one might say that an LLM's creators' intentions might be general enough to guarantee that the words it produces would be meaningful in their respective languages and perhaps to accommodate our cases of novel reference.

But even supposing this response were to offer an explanation of the capacity for novel reference in LLMs as they are today, it is not sufficiently general to accommodate our judgments about LLM meaning in closely related cases. LLMs can be created for different reasons: if the "same" LLM was created by Team A for the purpose of measuring sentence probabilities for use in a downstream application, and by Team B for use as a chatbot, it seems odd to conclude that only the second of these generates meaningful text in our cases.

Intentions in Generating the Prompt A third point at which human intentions might enter the production process is through the user. Perhaps in our particular prompts involving novel reference, the *user* has an intention that whatever name the LLM produces (e.g, "Marion Starlight") should refer to the person best described by the surrounding text (or to the aspect of the diagram best described by this text). On this view, the LLM's words are only meaningful in virtue of the user's attitudes.

Whether or not this approach succeeds for actual LLMs today, the approach again does not make correct predictions in relevantly similar cases. Suppose that we initiate a process in which an LLM is provided with random prompts (perhaps prompts generated by a unigram model), with no intentions about the meaningfulness of any generated text. Suppose moreover by chance a model is fed our prompt asking for a story featuring a new name for an historical figure. If the LLM offered the responses described above, it still seems to us that the LLM would produce inscriptions which refer to Robespierre or to aspects of the relevant diagram. But this reference would not be due to the creator of the prompt, since by assumption there is no user which has intentions.

Reader's Intentions A fourth and final place where human intentions might enter the picture is through the reader of the text (who might not be the creator of the prompt). In this vein, Cappelen and Dever (2021, Ch. 4) develop a receiver-focused "metametasemantics" according to which tokens can count as meaningful in virtue of how readers would understand them. We consider this response the most promising option for bibliotechnists, and it deserves much more detailed discussion than we can give here.

Here we will mention just one preliminary reservation, as an indication of a direction for future work. As it stands the theory cannot obviously distinguish between cases that are equally intelligible to a reader but intuitively differ in meaning. For instance, the same string that would be meaningful if an inscription of it was generated by a person is not meaningful if is created by the wind in the sand. But these strings will not differ in intelligibility to a reader in their two inscriptions. If the Cappelen-Dever theory is to save bibliotechnism, it must draw a distinction between these two cases without appealing to differences in the attitudes of the producers of the relevant text. This may not be impossible to do, but it is a challenge for the view as it stands.

8 Novel Reference, Interpretationism, and The Attitudes of LLMs

How might the problem of novel reference contribute to the broader question of whether LLMs have attitudes like belief, desire, and intention?

Let us start with the place of these attitudes in the explanation of human behavior. Human behavior can presumably be explained and predicted at the microphysical level by the laws of physics. But the fact that it can be does not mean that beliefs, desires and intentions are not *also* useful in explaining and predicting behavior. These descriptions are not as informative as full microphysical descriptions. But they are more efficient for making high-level predictions about future behavior, as well as about behavior in counterfactual circumstances.

LLM behavior in producing novel reference is similarly easier to explain by attributing beliefs, desires, and intentions to LLMs, than by appealing to details of their implementation. For instance, our first case can be explained on the hypothesis that the LLM intends for the term "Marion Starlight" to be equivalent to "Robespierre" (among many other possible explanations). In the second case, we can explain the LLM's behavior on the assumption that it intends that its new name apply to a particular aspect of the diagram.[4] These hypotheses about intention allow us to explain the LLM's actual behavior in producing inscriptions of the relevant names, as well as its counterfactual behavior; for instance, how it would answer various further questions featuring "Marion Starlight".

According to *interpretationism* in the philosophy of mind and cognitive science (e.g., Dennett, 1971; Davidson, 1973, 1986; Dennett, 1989), (roughly) a system has beliefs, desires and intentions if and only if its behavior is well explained by the hypoth-

[4] There is much controversy in philosophy of language about how exactly to understand cases of "initial" reference or "baptism" with a new name, even when these acts are performed by people. But if LLMs have attitudes, any of the usual accounts can be extended to them.

esis that it has those attitudes and is rational. Along these lines, McCarthy (1979) writes: "To ascribe certain beliefs, knowledge, free will, intentions, consciousness, abilities or wants to a machine or computer program is legitimate when such an ascription expresses the same information about the machine that it expresses about a person. It is useful when the ascription helps us understand the structure of the machine, its past or future behavior, or how to repair or improve it. " He notes that this is most usefully applied to machines whose inner workings are opaque, although it is more straightforwardly (but less usefully) applied to transparent machines like thermostats.

Here some behavior counts as "well explained" by some hypothesis (very roughly) if the hypothesis offers a sufficiently simpler explanation, which makes sufficiently accurate predictions in a sufficiently wide array of counterfactual circumstances. The explanation of an apple's fall from a tree in terms of gravitation is intuitively at least as simple as an explanation which attributes to the apple a desire to fall to the ground, but the latter plausibly offers less accurate predictions in a wide array of counterfactual circumstances. By contrast, an explanation of a human's purchase of a cup of coffee in terms of their cellular biology is much more complex, without a sufficiently great compensatory gain in counterfactual accuracy, than an explanation which attributes a desire for coffee to the person, along with the belief that buying it is the best way to get it (Jara-Ettinger et al., 2016; Jara-Ettinger, 2019). Using the criteria above, interpretationists will argue that the behavior of people but not the behavior of apples is well explained by the hypothesis that they have beliefs, desires and intentions. So interpretationism will correctly predict that the former, but not the latter, actually have such beliefs, desires, and intentions.

Interpretationism offers a stark and perhaps counterintuitive picture of the nature of belief and other attitudes. On this view, people count as having these states, not because of the details of our inner workings or because of our conscious experience, but because our actual and counterfactual behavior is well-explained on the hypothesis that we have these states. Similarly, for interpretationists, the fact LLM behavior in cases of novel reference is well explained on the hypothesis that LLMs have beliefs, desires, and intentions, provides direct evidence that LLMs do have these states.

There is no universally accepted philosophical theory of the nature of belief, desire, and intention, and we cannot hope to decisively argue for one here. Certainly, not all philosophers accept interpretationism (for a survey of alternatives see Schwitzgebel, 2023), but many hold that some version of this view, or at least a descendant of it, may be viable. Anyone who places even some degree of confidence in interpretationism (which we think it is rational to do) and also accepts that, given intepretationism, the novel reference problem is direct evidence that LLMs have beliefs, desires, and intentions, should accordingly raise their degree of confidence in the claim that LLMs have these states. In this sense, the novel reference problem provides evidence that LLMs do have beliefs, desires and intentions.[5]

We emphasize that this conclusion does not imply anything like the claim that human or superhuman intelligence is just around the corner (*contra* Bubeck et al., 2023). Rabbits, spiders, and possibly even fish have beliefs, desires, and intentions. But these animals are not super-intelligent.[6]

9 Conclusion

We began by arguing that bibliotechnism requires that LLMs produce inscriptions which are only derivatively meaningful, if meaningful at all. We suggested that this claim raised a challenge for bibliotechnism, since LLMs often produce entirely novel text, and it is at first sight unclear how novel text might be derivatively meaningful. We then

[5]Other live philosophical views of these states, including varieties of functionalism, will also see the fact that attributing them to LLMs provides a good explanation of LLM behavior, to be some evidence that LLMs have these states (see, e.g., Schwitzgebel, 2023; Goldstein and Kirk-Giannini, manuscript). Our argument in the main text can be extended to these views as well. Some philosophers may reject our earlier explanations of LLM behavior in the novel reference problem and favor explanations of LLM behavior at the subpersonal level in terms of more generic "representational states" rather than in terms of propositional attitudes. We agree that the novel reference problem also provides evidence for the attribution of representational states in this more generic sense than for the attribution of beliefs, desires, and intentions. Our goal here has been to assess the implications of novel reference for the question of whether LLMs have beliefs and other propositional attitudes, not to provide a comprehensive evaluation of all ways that this behavior might be explained.

[6]Our account also does not require that LLMs have semantic understanding, and we are open to the idea that they do not; Titus (2024) offers particularly helpful analysis (cf. Bender and Koller, 2020). We are only committed to denying that such semantic understanding (or even any kind of world-knowledge Yildirim and Paul (2023)) is required for a system to produce meaningful inscriptions. We take this latter point to be demonstrated by the fact that photocopiers produce such inscriptions.

showed how entirely novel text may nevertheless be derivatively meaningful. In our view, this represents an important step forward for bibliotechnism But it is not the whole story. We went on to describe a new task (the novel reference problem) and argued that if LLMs succeed in this task, this would pose a challenge for the view that all inscriptions produced by LLMs are derivatively meaningful, and hence for bibliotechnism.

Throughout, we have focused on theories which allow that LLM-generated text may have linguistic meaning. Some proponents of a view similar to bibliotechnism might prefer to develop their position in a different way. There is a sense in which the presence of smoke "means" that there is fire, and Grice (1957) called this sense of "meaning", "natural meaning" (as opposed to "nonnatural meaning", of the linguistic kind we have been examining). We would be interested to know if a version of bibliotechnism can be developed using this notion of natural meaning instead of the notion of linguistic meaning we have focused on. We have not ourselves pursued this route because we have not been able to come up with a reasonable exact proposal for what the "fire" would be that the LLM text indicates as the "smoke". We also note that, even if this view can be developed in more detail, it will plausibly still face the novel reference problem, since it is unclear what "fire" the LLM would be indicating in those examples.

An alternative way of developing a view similar to bibliotechnism would be to deny that the inscriptions produced by LLMs are meaningful *at all* (Ostertag, 2023b; Titus, 2024). In the absence of the theory we provided in the first half of the paper, we agree that this might be an attractive option. But given the existence of this simple story—which makes sense of the meaningfulness of LLM-generated text without attributing any attitudes to LLMs, and using tools that are already required to make sense of the meaningfulness of photocopier-generated text—such a radical view should be disfavored. The appearance of meaningfulness in a body of text is certainly not dispositive evidence in favor of its meaningfulness. Word-shapes written by the wind in the sand at random might appear to be meaningful, even though they would not be. But the appearance of meaningfulness in a body of text is still *some* evidence in favor of its meaningfulness. The version of bibliotechnism we developed in the first part of the paper has the advantage of vindicating this appearance.

Still, one might endorse our account of the meaningfulness of most LLM-generated text, while rejecting the claim that putative examples of novel reference are in fact meaningful. This proposal represents yet another response to the novel reference problem, beyond the four considered in section 7, which would allow bibliotechnists to preserve a fairly simple version of their position, without the cost of denying that *all* text produced by LLMs is meaningless. Like the responses we considered there, it deserves much more detailed consideration than we can give it here. But it does still have the cost of denying that apparently meaningful text is in fact meaningful.

In Part II, we focused on a different response to the novel reference problem. This response, based on *interpretationism*, take such examples to be direct evidence that LLMs have beliefs, desires, and intentions. Interpretationists will take many instances of complex behavior in a system to be evidence that the system has beliefs, desires, and intentions. The problem of novel reference is just one example of such evidence. But it is a clear example of this kind, and we hope it will spur attempts to describe other examples that can also serve as tests for the presence of these attitudes.

In our view, the question of whether LLMs have representational states, and what representational states they have will only be settled by careful analysis of a wide array of their behavior and how it can best be explained, leading to a holistic case that they do or do not have such states (Levinstein and Herrmann, 2023). If theories like bibliotechnism, which do not attribute representational states to LLMs, fail to explain features of the the behavior of LLMs, or can only explain them by becoming thinner and more complex, a simpler, stronger explanation involving representational states should be favored. This perspective is in line with a growing body of work which advocates using tools from cognitive science to understand LLMs (Mitchell and Krakauer, 2023), perhaps viewing them as alien intelligences (Frank, 2023; Cappelen and Dever, forthcoming) or as role players (Shanahan et al., 2023), to be studied from the outside. The novel reference problem is one case that puts pressure on a variety of alternative explanations of LLM behavior, and thus provides some evidence that LLMs do have some form of representational states.

10 Acknowledgments

For helpful comments on drafts, we thank Josh Dever, Robbie Kubala, Matt Mandelkern, Gary Ostertag, Dan Roberts, and Sinan Dogramaci. For helpful conversations in thinking through these issues, we thank David Beaver, Ray Buchanan, Chiara Damiolini, Kenny Easwaran, Katrin Erk, Steven Gross, Dan Harris, Cameron Yetman and participants in UT Austin's LIN 393 graduate seminar. We thank Emily Bender for a helpful Twitter discussion of an earlier draft of this paper. K.M. acknowledges funding from NSF Grant 2139005.

References

Jacob Andreas. 2022. Language models as agent models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5769–5779, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

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.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Tyler Burge. 1986. Individualism and psychology. *The Philosophical Review*, 95(1):3–45.

Herman Cappelen and Josh Dever. 2021. *Making AI intelligible: Philosophical foundations*. Oxford University Press.

Herman Cappelen and Joshua Dever. forthcoming. Ai with alien content and alien metasemantics. In Ernest Lepore, editor, *Oxford Handbook of Applied Philosophy of Language*. OUP.

David J Chalmers. 2023. Could a large language model be conscious? *arXiv preprint arXiv:2303.07103*.

Ted Chiang. 2023. Chatgpt is a blurry jpeg of the web. *New Yorker*.

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

Donald Davidson. 1973. Radical interpretation. *Dialectica*, pages 313–328.

Donald Davidson. 1986. A coherence theory of truth and knowledge. *Epistemology: an anthology*, pages 124–133.

Daniel C Dennett. 1971. Intentional systems. *The Journal of Philosophy*, 68(4):87–106.

Daniel C Dennett. 1989. *The Intentional Stance*. MIT press.

Keith S Donnellan. 1970. Proper names and identifying descriptions. *Synthese*, 21:335–358.

Gareth Evans. 1973. The causal theory of names. *Proceedings of the Aristotelian Society, Supplementary Volumes*, 47:187–225.

Michael C Frank. 2023. Baby steps in evaluating the capacities of large language models. *Nature Reviews Psychology*, 2(8):451–452.

Peter Thomas Geach. 1969. The perils of pauline. *The Review of Metaphysics*, pages 287–300.

Simon Goldstein and Cameron Domenico Kirk-Giannini. manuscript. AI wellbeing.

Alison Gopnik. 2022a. Children, creativity, and the real key to intelligence. *Observer*.

Alison Gopnik. 2022b. What ai still doesn't know how to do. *The Wall Street Journal*.

H Paul Grice. 1957. Meaning. *The Philosophical Review*, 66(3):377–388.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Peter Hase, Mona Diab, Asli Celikyilmaz, Xian Li, Zornitsa Kozareva, Veselin Stoyanov, Mohit Bansal, and Srinivasan Iyer. 2023. Methods for measuring, updating, and visualizing factual beliefs in language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2714–2731, Dubrovnik, Croatia. Association for Computational Linguistics.

Julian Jara-Ettinger. 2019. Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, 29:105–110.

Julian Jara-Ettinger, Hyowon Gweon, Laura E Schulz, and Joshua B Tenenbaum. 2016. The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in cognitive sciences*, 20(8):589–604.

Saul A Kripke. 1980. Naming and necessity. In *Semantics of natural language*, pages 253–355. Springer.

Brenden M Lake and Gregory L Murphy. 2023. Word meaning in minds and machines. *Psychological review*, 130(2):401.

B. A. Levinstein and Daniel A. Herrmann. 2023. Still no lie detector for language models: Probing empirical and conceptual roadblocks.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2023. Dissociating language and thought in large language models: a cognitive perspective. *arXiv preprint arXiv:2301.06627*.

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

John McCarthy. 1979. *Ascribing mental qualities to machines*. Stanford University. Computer Science Department.

R. Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. 2023. How much do language models copy from their training data? evaluating linguistic novelty in text generation using RAVEN. *Transactions of the Association for Computational Linguistics*, 11:652–670.

Raphaël Millière and Cameron Buckner. 2024. A philosophical introduction to language models–part i: Continuity with classic debates. *arXiv preprint arXiv:2401.03910*.

Melanie Mitchell and David C Krakauer. 2023. The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120.

Gary Ostertag. 2023a. Large language models and externalism about reference: Some negative results. (unpublished manuscript).

Gary Ostertag. 2023b. Meaning by courtesy: Llm-generated texts and the illusion of content. *The American Journal of Bioethics*, 23(10):91–93.

Roma Patel and Ellie Pavlick. 2021. Mapping language models to grounded conceptual spaces. In *International Conference on Learning Representations*.

Steven T Piantadosi and Felix Hill. 2022. Meaning without reference in large language models. *arXiv preprint arXiv:2208.02957*.

Hilary Putnam. 1975. The meaning of "meaning". In *Minnesota Studies in the Philosophy of Science, Volume 7: Language, Mind, and Knowledge*, pages 131–193. University of Minnesota Press, Minneapolis.

Eric Schwitzgebel. 2023. Belief. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Winter 2023 edition. Metaphysics Research Lab, Stanford University.

Cosma Shalizi. 2023. "Attention", "Transformers", in Neural Network "Large Language Models".

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, pages 1–6.

Lisa Miracchi Titus. 2024. Does chatgpt have semantic understanding? *Cognitive Systems Research*, 83(101174):1–13.

Ilker Yildirim and LA Paul. 2023. From task structures to world models: What do llms know? *arXiv preprint arXiv:2310.04276*.

Eunice Yiu, Eliza Kosoy, and Alison Gopnik. 2023. Transmission versus truth, imitation versus innovation: What children can do that large language and language-and-vision models cannot (yet). *Perspectives on Psychological Science*, page 17456916231201401.