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 do generate entirely novel text. We begin by defending bibliotechnism against this challenge, showing how novel text may be meaningful only in a derivative sense, so that the content of this generated text depends in an important sense on the content of original human text. We go on to present a different, novel challenge for bibliotechnism, 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. In line with this view, we argue that cases of novel reference provide evidence that LLMs do in fact have beliefs, desires, and intentions, and thus have a limited form of agency.

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; Bender and Koller, 2020; Shanahan et al., 2023; Bubeck et al., 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, that is, if they are agents, we can employ the same explanations of their behavior.

A challenge for those who deny that current LLMs are agents in this sense 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). 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 were not simply transmitting existing cultural knowledge, but generating 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 apparently meaningful text, which is entirely novel. At first sight, this fact presents a serious challenge for bibliotechnism: if LLMs' outputs can only be meaningful by piggybacking on human-generated originals, how could novel text that they generate be meaningful?

We begin by responding to this challenge for bibliotechnism. Using n-grams as a toy model, and working up to more complex modern LLMs, we show how even entirely novel sentences produced by LLMs may nevertheless derive their meaningfulness from the meaningfulness of their inputs and thus be "derivatively meaningful".

We see this as a big step forward for bibliotechnism. But challenges for the proposal remain. Modern LLMs are not just capable of producing new sentences, they can also generate novel reference, using newly invented names apparently to refer to newly created objects. These new names cannot derive their reference from original text, since the name and its object are not associated in the data (if they occur there at all). 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's answer to the Novel Reference Problem is a worse explanation of LLM behavior than the hypothesis that they have beliefs, desires, and intentions. According to many views in the philosophy of mind, this fact provides evidence that LLMs do have beliefs, desires, and intentions. This is most obviously true given *interpretationism*, according to which a system has beliefs, desires and intentions if and only if its behavior is best 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). But the point holds for other prominent philosophical views as well. So, the problem of novel reference provides evidence that LLMs do have beliefs, desires and intentions, and thus are agents of a sort.

2 Prior Work on Meaning in LLMs

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 use language meaningfully. They define "meaning" as a relation between expressions and communicative intents. They argue that LLMs cannot produce meaningful expressions because they cannot have communicative intents concerning objects they have not had perceptual contact with.

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, 2023).

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.

These authors do not consider how questions about the meaningfulness of LLM-generated text relate to bibliotechnism. They also do not address how, if at all, simpler models like n-gram models can produce meaningful text. In fact, to the extent that they deliver verdicts about n-grams, their theories do not clearly imply that n-grams can pro-

duce meaningful text. By contrast, we will argue that n-gram models *can* clearly produce meaningful words. We will build on this account in the case of n-grams to offer a novel extension of bibliotechnism, showing how the view can accommodate the meaningfulness of entirely novel text generated by LLMs. We then introduce the "Novel Reference Problem", a new challenge for bibliotechnism.

3 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 will 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, normal uses of the expression "was born" is meaningful, but it does not refer.

Second, there is a difference between the word "Shakespeare" and particular *inscriptions* or *tokens* 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 will 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]

4 From Cultural Technology to Derivative Reference and 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, most notably, 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 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

[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 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* reference and meaning, while the other inscriptions are instances of *derivative* reference and meaning. We stipulate, as part of the definition of these terms, that it is only agents (which we understand to mean entities with 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 are not agents. So, according to bibliotechnism, they 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.

5 Causal Connection and Derivative Reference: The Case of N-grams

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 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 (with n sufficiently large, e.g. 1000, that they copy long stretches of their input) can produce meaningful inscriptions. Since such an n-gram model copies their input, their output can be meaningful, just as the photocopier's is.

Matters are less straightforward for unigram models, which sample from a distribution of single words. We can think of this model as implemented by taking all of its `PrimaryData`, choosing word-inscriptions from the `PrimaryData` at random, and then copying the chosen inscription. The inscriptions 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 (and possibly refer), but it does not seem that inscriptions of complex expressions formed from these words will be. The vast ma-

jority 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 as an exact copy of an original string. But even in these latter cases, the fact that a meaningful string (or even a copy) is produced is a fluke, a complete accident. Since the model produces "appropriate" expressions in such a chaotic way, 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, this inscription does not have a meaning. The generated text is not appropriately causally connected to the `PrimaryData` to inherit "glue" that binds the complex expression together. One way to put this point would be to say that, while the inscription it produces 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 words blown together by the wind, or sand dunes blown into the shape of a sentence—or like Mandelkern and Linzen's "ants formicating meaninglessly in the sand".

6 LLMs do Produce Derivatively Meaningful Complex 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 there is another route to producing derivatively meaningful inscriptions of complex expressions, where the causal chain which traces from the inscription of individual words back to the original inscriptions is distinct from the causal chain which is responsible for the expression-level features which make a whole complex expression meaningful.

To see the basic idea, suppose we have a rudimentary model, which, when fed some text, finds any inscriptions of names in this 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`. It seems plausible that in this case, the model would produce not just individual words which are derivatively meaningful, but 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. 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 sentence-level features, in such a way as to reliably produce an intelligible sentence. This could be so, even if the resulting sentence has never been seen before.

This example shows that even novel text can be derivatively meaningful. It also suggests that, if LLMs are causally sensitive to high-level features of their `PrimaryData`, in such a way as to transmit those features to their `GeneratedText`, it is possible that even their entirely novel `GeneratedText` could be derivatively meaningful. We suggest that one such high-level feature is *intelligibility*. As we will understand the notion, intelligibility of expressions requires they they be at least quasi-grammatical (sentences with minor grammatical errors are often perfectly intelligible), but it requires more than just grammaticality, since not all grammatical sentences are meaningful (and hence not intelligible). If LLMs are causally sensitive to intelligibility, so that they produce intelligible outputs from 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`.

There is strong evidence that modern LLMs are in fact appropriately causally sensitive to the intelligibility of their `PrimaryData`. First, they overwhelmingly produce text which is intelligible to human users. This does not *prove* that they are causally sensitive to this property, but it is strong evidence that they are. 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 causally transmit the intelligibility of their data.

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 intelligible. This is important because 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 if the model produces incoherent 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 text to text generated by bigram or trigram models. Bigram and trigram models are trickier cases than unigrams because they do copy short complex phrases and might be statistically likely to combine them in ways that are more plausibly meaningful as a whole. For instance, a bigram model that outputs "Shakespeare wrote plays" in a way "knows" that "wrote" is a likely continuation for "Shakespeare" and that "plays" is a likely continuation for "wrote". But, besides the one mediated by the verb "wrote", there is no causal connection between "Shakespeare" and "plays". So, 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 this is more of a borderline case): they lack the straightforward "copy property" of higher-n n-gram models but also are not causally sensitive to intelligibility as modern LLMs are.

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 words. It is instead because this training mechanism 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.

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 tokens it produces refer—as long as this underlying low-level mechanism reliably produces inscriptions of actual words, that is, as long as the low-level mechanism leads to causal sensitivity to the right high-level features.

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. 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. We next turn to a new and different kind of challenge to this view: the fact that LLMs can generate novel reference.

7 The Novel Reference Problem: LLMs Do Not Produce *Only* Derivatively Meaningful Expressions

We will illustrate the problem of novel reference with two examples.

First, LLMs can produce tokens of names they have never seen before, intuitively in such a way that they refer to previously referred-to objects. Suppose we ask an LLM to choose any real histor-

ical figure it likes, and then come up with a new name and tell us facts about this historical figure. ChatGPT (GPT-4) completed this task by telling us about "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." We think it is clear that inscriptions of "Marion Starlight" in this text refer to the historical figure Robespierre.

But nothing in `PrimaryData` (presumably) associates Marion Starlight with Robespierre. So the LLM's inscriptions of this name cannot refer to Robespierre in virtue of original reference exhibited by inscriptions of this name in the `PrimaryData`.

A second example sharpens the problem. Suppose we ask an LLM to produce the TikZ codes for a series of pictures which it has never seen before, to give elements of those pictures names, and then to describe features the picture would have when typeset using those names. If an LLM can do this, then it is even more clear than in the previous example that the reference of these expressions could not be due to some reference in the `PrimaryData`, since the object did not exist in this form until the LLM created it (provided the picture really is new).

While we do not here present empirical experiments for these cases and rely on anecdotal examples, we think they are well within the capabilities of modern LLMs, which have been empirically shown to be able to generate, refer to, and manipulate elements of code-generated pictures (Bubeck et al., 2023) and refer meaningfully to novel orientations of elements in visual and color spaces (Patel and Pavlick, 2021).

These examples cannot be straightforwardly accommodated by the derivative reference account given so far, since there is no association between the names and their referents in `PrimaryData`. In the next section we consider some responses to this problem which involve expanding the notion of derivative reference, before turning to our own favored response in the conclusion.

8 Responses to the Novel Reference Problem

If bibliotechnism is correct, our cases of "novel" reference must in fact be derivative, so there must be some way in which the inscriptions "piggyback" on original human reference. Other than the original data, which we have already ruled out, there seem to be four other salient places where human attitudes might enter the model pipeline. In this section, we briefly consider some 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 as a central point. 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 are able to produce intelligible output and follow instructions to some extent. Thus, an account which considers RLHF-ed models to be radically different in their basic referential abilities seems not to align with the empirical data.

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 a human'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 LLMs' 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.

Even supposing this response were to offer an explanation of the capacity for novel reference in LLMs as they are today (we have our doubts), this approach 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 can generate meaningful

text in our cases.

Intentions in Generating the Prompt A third point at which human intentions might emerge is through the user. Someone might say that 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 user (in writing the prompt and interpreting the output) is crucial for generating meaning, and the LLM's words are only meaningful in virtue of user attitudes.

This approach again does not make correct predictions in relevantly similar cases. Suppose that we generate prompts randomly and provide them to an LLM (perhaps generating them by a unigram model), and that by chance a model is fed the prompt asking for a story featuring a new name for an historical figure. If the LLM offered the response above, it still seems to us that the LLM would produce inscriptions which refer to Robespierre. 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 just mention 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 will not be meaningful if is created by the wind in the sand. But the string will not differ in intelligibility to a reader in its two inscriptions. If the theory is to save bibliotechnism, it must draw this distinction 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.

9 Conclusion

We argued that bibliotechnism requires that LLMs produce inscriptions which are only derivatively meaningful. We went on to develop a notion of derivative meaning which allows that many inscriptions, even of complex expressions, produced by LLMs are in fact derivatively meaningful. But we argued that the problem of novel reference poses a challenge for the view that all of them are.

Throughout, we have focused on this notion of derivative meaning. Some proponents of a view similar to bibliotechnism might prefer to develop 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 linguistic meaning). We think it would be interesting to see how a version of bibliotechnism would look if developed using natural meaning instead of linguistic meaning. We have not taken this route here 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 were to be developed in more detail, it will still face the novel reference problem, since it is highly unclear what "fire" the LLM is indicating in those cases.

In closing, we want to consider more generally how our discussion of biblotechnism and novel reference may 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. But the fact that it can be does not mean that beliefs, desires and intentions are not *also* useful in explaining and predicting behavior. As we all know from our daily lives, they are extremely useful for these purposes.

The examples of novel reference provide an example where it is easier to explain LLMs' behavior by attributing beliefs, desires, and intentions to them, rather than by offering a complex, contorted theory of derivative reference. For instance, our first case can be explained by the hypothesis that the LLM intends for "Marion Starlight" to be equivalent to "Robespierre" (among many other possible explanations).

According to the prominent tradition of "interpretationism" in philosophy 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 best explained by the hypothesis 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. In our view, we can explain cases of novel reference in a much more straightforward way if we attribute some representational states (and possibly beliefs, desires and intentions) to the LLMs. According to interpretationism, this fact provides strong, straightforward evidence that LLMs do have beliefs, desires and intentions. But many other philosophical views of these states, including varieties of functionalism, will also take the fact that attributing these states to LLMs provides a good explanation of their behavior, to be evidence that they have these states (see, e.g., Schwitzgebel, 2023; Goldstein and Kirk-Giannini, manuscript). We cannot offer a comprehensive survey of approaches to belief, desire, and intention here, but this is a reasonably feature of such approaches, even if it is not universal.

As we emphasized earlier, the explanation of LLM behavior in terms of beliefs, desires and intentions, is not meant to replace an explanation of their behavior at a finer level of detail. Of course the LLM is also "just" sampling from a distribution over words, just as humans are presumably "just" collections of atoms. But at a high-level the LLM behavior may also be well explained (indeed, better explained) by its having some representational states.

To say that LLMs have beliefs, desires and intentions would not be to say that human or superhuman intelligence is just around the corner (*contra* Bubeck et al., 2023). Spiders, rabbits, and possibly even fish have beliefs, desires, and intentions. But these animals are not super-intelligent.

Our conclusions here are 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) or as role players (Shanahan et al., 2023) to be studied from the outside. In our view, the question of whether LLMs have representational states, and what representational states they have will 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 these states. If theories like bibliotechnism which do not attribute representational states to LLMs can only explain the behavior of LLMs by becoming thinner and more complex, a simpler, stronger explanation involving representational states becomes more attractive. By putting this kind of pressure on such alternative explanations, the novel reference problem provides some evidence that LLMs do have such representational states and, accordingly, at least a limited form of agency.

10 Acknowledgments

For helpful comments on drafts, we thank Josh Dever, Robbie Kubala, Matt Mandelkern, Gary Ostertag, and Sinan Dogramaci (who introduced the creators' intentions objection, comparing LLMs to thermometers). For helpful conversations in thinking through these issues, we thank David Beaver, Ray Buchanan, Chiara Damiolini, Katrin Erk, Steven Gross, Dan Harris, and participants in UT Austin's LIN 393 graduate seminar. 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.

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.

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.

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.

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.

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

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.

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.