What do Large Language Models Represent?

Quentin Ruyant
Departamento de filosofía, lógica y filosofía de la ciencia
Universidad de Sevilla

Abstract

If large language models (LLMs) are models, what do they represent exactly? I address this question by applying the various conceptions of epistemic representation that have been entertained by philosophers of science to this case. After discarding the idea that they would represent the structure of natural languages, I argue that LLMs do not directly represent the world by linguistic means either, but rather a certain class of appropriate linguistic production that is constructed by their makers. They are mostly used to generate fictitious instances of this class, although these fictions can be performative when appropriated by humans. The main implication of this analysis is that conversing with a LLM chatbot amounts to engaging in a fiction, where our own imagination plays a central role. I examine some ethical consequences of this conclusion regarding how LLMs are currently deployed in numerous places to serve as "AI assistants".

Keywords: Large language models, Artificial intelligence, Scientific representation

Large language models (LLMs) have come to the fore in recent years for their capacity to produce intelligible texts and mimic conversations with humans in a way that makes them hard to distinguish immediately from human speakers. This has generated various debates on their possible status as agents and on the meaningfulness of their production.

In this article, I wish to approach these issues from a different angle. If we take seriously the idea that LLMs are, as their name suggests, *models*, what do such models represent? This is the question that I will attempt to answer, by taking as input the literature on scientific representation. I believe that answering this question can shed light on the nature of LLMs and inform various issues regarding their use, but also incidentally serve to evaluate the merits of the different conceptions of epistemic representation that have been entertained by philosophers of science.

In section 1, I briefly present these conceptions of representation, distinguishing three main categories: those based on content, intentions and norms. In section 2, I examine the internal structure, creation process and typical uses of LLMs. In section 3, I bring the two together to determine what LLMs represent. My conclusion is that they are meta-representational: they represent a certain class of “appropriate” linguistic productions. In practice, instances of this class are mostly imaginary, but they can be performative if appropriated and deployed by human agents. In section 4, I draw implications of this conclusion for broader issues, notably ethical issues associated with automated deployments of LLMs as “AI agents”, which are becoming increasingly common.

2 What is epistemic representation?

Scientific models, as well as city maps, are epistemic vehicles. They have content that is accessible to their users, and this content is *about* another entity. They thus allow their users to make inferences about this entity without the need to access it directly. The represented entity is often called the *target* of representation, as opposed to the *source*, which refers to the model or vehicle. The target of representation of a city map is the configuration of a city. In science, the target is typically some object or phenomenon under a description, which is to say that the model focuses on specific aspects, properties or contrast classes of this object assuming certain background conditions. For example, the model of the pendulum represents the position of a bob attached to a string and subjected to a uniform gravitational field along a particular axis in the absence of wind. Philosophers of science have been interested in explaining what kind of relationship is involved between sources and targets of representation when such models are used. They put a lot of emphasis on scientific representation, but it should be noted that their accounts often aim at addressing epistemic representation in general, including, for example, city maps and the like.

My aim in this article is to analyse LLMs by taking them to be epistemic vehicles. By this, I do not necessarily mean that they are scientific models: although LLMs can be used by scientists, they are more often deployed in ordinary contexts to serve as chatbots than in scientific contexts. However, this does not mean that they should not be treated as epistemic vehicles, in the sense that they allow their users to make inferences or to learn something in a broad sense. My working hypothesis is that they can indeed be treated as epistemic models, and I hope that this hypothesis can give us insight into their nature by revealing some features that are obscured when we only treat them as potential agents, as is often done in the philosophical literature.

Taking seriously the idea that LLMs are epistemic representations, the question arises: what do they represent exactly? Their name suggests that it has something to do with language, but we could wish for a more precise characterisation. In order to inquire into this issue, let us start by asking what determines the target of representation of epistemic models in general. The responses philosophers have offered to this question can be divided into three broad categories[1].

[1] Theories of representation are often classified into two categories, for example substantive vs. deflationary (Suárez 2024, ch. 4) or informational vs. functional (Anjan Chakravartty 2010), depending on whether they assume that a specific source—target relation is required for representation to take place, or if a functional role is enough. Substantive/informational accounts roughly correspond to what I call “content-views” here. I make a further distinction between intention-views and norm-views within the function/deflationary category. I think it comes out naturally when the question is framed as “what exactly determines the target of representation?”,

Content-views

According to what I will call *content-views*, what determines the target of representation of a model is mainly its content. A picture typically looks like the object it represents and allows us to know about the appearance of this object without directly seeing it. Similarly, a model (or vehicle) would fulfil its epistemic function in virtue of being, in relevant respects, a copy of what it represents. In this view, for a model to represent something, the right kind of relation must necessarily hold between the source and target of representation: typically, similarity or some type of morphism (isomorphism, homomorphism, etc.). For the sake of simplicity, I will only talk about similarity in what follows, assuming that morphisms are structural similarities.

We could make similarity not only necessary, but also sufficient for representation, implying that if something is similar to another thing, then it is a representation of it. However, there are very few defenders of such a liberal view, if any. The main reason is that representation is generally thought of as directional (the vehicle represents the target, not the other way around) in a way that is not captured by similarity relations. Most authors who claim that some form of similarity is necessary for representation also invoke other essential components, such as user intentions and denotation, to ensure this wanted directionality (Mundy 1986; Watson 1995; Chakravartty 2010; Poznic 2016b). We could say, building on (Bartels 2006), that what is identified by means of similarity is only *potential* representation, in the sense that any similarity can be exploited for representation purposes, but that *actual* representation only takes place when an agent or some causal mechanism picks or produces a relevant similarity relation. In so far as similarity is still required for representation to take place, this is still a content-view.

In this approach, a single model *potentially* represents many entities simultaneously. As such, model content per se is naturally associated with kinds of objects, those identified by the content of the model, rather than with instances. For example, the model of the simple pendulum would potentially represent every simple pendulum in the world, assuming that they all share a relevant similarity to the model. Users would typically pick a concrete instance for their purpose in context. By analogy with linguistic representation, potential representation relations could be considered a matter of semantics, having to do with the context-independent "meaning" of the model, and actual representation relations a matter of pragmatics, having to do with the contextual intentions of model users.

Intention-views

Content-views have been criticised on the ground that no relevant similarity is actually necessary for representation, because misrepresentation is possible. Just like an inaccurate map that fails to perfectly reproduce the configuration of a city would still be considered a map of this city, a model can be a bad representation, but still a representation of its target (Suárez 2003). Some prefer to deny this and argue that representation is a "success term" (Poznic 2018; Chakravartty 2010, 209–10). Others, however, see a defect in this inability to account for misrepresentation. One remedy is to adopt what I shall call an *intention-view*, where the intentions of model users are considered sufficient to determine targets of representation.

On one approach, users or modellers postulate that there is a relevant relation such as similarity between the model and a denoted target, even though it might not actually hold (Giere

which is what interests us here.

1988; Weisberg 2013). I will refer to these views as *intended-similarity-views*. On another approach, which I will call *intended-interpretation-views*, modellers provide an interpretation of the model in terms of the target so as to make inferences about the target. This can involve a simple denotation relation between model and target, or a more complex set of denotation relations between component parts of the model and target, including property denotation, in the way particular symbols on a city map denote metro stations for example (Hughes 1997; Callender and Cohen 2006; Contessa 2007; Frigg and Nguyen 2020). In both cases, misrepresentation is possible: it arises when the model fails to be relevantly similar to the denoted target, or when once interpreted, the inferences it warrants lead to incorrect conclusions about its target.

A problem for intention-views is that they put very few constraints on what can represent what: it is up to users to decide. This problem can be analysed in terms of a failure to account for *misuse*, by analogy with content-views failing to account for misrepresentation (Ruyant 2025). If I am intentionally using a map of Lisbon to navigate in Brussels, then I am misusing the map, because the map does not actually represent Brussels (which is not the same as representing it inaccurately). Intuitively, we would even say that the map represents its target, Lisbon, independently of any particular use, and so do scientific models or epistemic vehicles in general. If the target of representation were determined solely by use, then misuse would be impossible, which seems wrong.

Another problem for intention-views is that models are generally taken to represent kinds, in the sense that they can be used to represent any instance of the kind. The model of the pendulum, for example, does not represent one particular pendulum in the world. We saw that content-views can account for this aspect. Intention-views do it less easily, since users do not have the capabilities to denote all instances of the kind at once.

Norm-views

A way out of these problems is what I shall call *norm-views*, according to which the target of representation of a model is determined by norms of appropriate use rather than by actual uses. These norms can be established by model makers, or when practice becomes entrenched or institutionally licensed (which does not mean that they are purely conventional: they are also responsive to epistemic values and aims). The Lotka-Volterra model, for instance, would represent a prey–predator system in virtue of a norm within the scientific community to the effect that it is appropriate to use this model for representing prey–predator systems. So, for something to represent something else, it must be "introduced into a normative representational practice" (Suárez 2024, 160); it must be "licensed by practice" (Boesch 2017, see also Boesch 2022); there must be "norms of appropriate use" in place (Ruyant 2025) (see also Bailer-Jones 2003, 72; Toon 2010, 80). Such norms concern not only which entity the source of representation can represent (what Suárez calls its "representational force"), which is what primarily interests us here, but also how the source should be manipulated and how the result of these manipulations should be interpreted.

Misuse naturally comes out from norm-views as a failure to respect norms of use, and misrepresentation is accounted for too, since norms of use do not guarantee adequacy or any relevant similarity between the model and its target. Finally, norm-views can share with content-views the idea that at the semantic level, models are primarily about kinds of entities, while they represent instances at the pragmatic level (Ruyant 2025). Indeed, norms of use can take the form of licensed *kinds* of interpretations for inferences, as understood in intention-views, associated with a *kind* of target, and users could pick or produce the relevant concrete target for their purpose and contextually interpret the model in terms of it in a way that complies with norms of use, or even imagine a fictional target if none is present.

The main difference between norm-views and content-views lies in the deflationary nature of the former: they do not require any substantial similarity relation between source and target for representation to take place, but only that a norm be in place. Similarity is not necessarily out of the picture with such views, but it cannot be considered constitutive of representation anymore: it is at best a by-product, or a criterion involved in the process by which norms of representation become established.

One approach that can be classified as a norm-view is to consider models as props in a game of make-believe: they would function as rules that prescribe users to imagine a given system in a certain way. This can mean either that the targets of representation are entirely fictional, even though they may be compared with real systems via a second representation relation (Godfrey-Smith 2007; Frigg 2010) or that a real target of representation is imagined in certain fictional ways (Toon 2010; Levy 2015). The main aim of fictionalist theories is to account for the way models in science are typically presented: as descriptions of systems that need not be actual or that have unrealistic features, such as being massless or frictionless. In so far as these prescriptions for imagination would be enforced at the communal level, this can be considered a norm-view.

Fictionalism can have difficulties in doing justice to the epistemic function of models (Poznic 2016a). However, it is possible to avoid these difficulties by considering that representation of fictional objects merely corresponds to one type of use among many appropriate ones for a given model (Ruyant 2025), or to view fictional aspects as facilitating inferences (Suarez2010).

3 What are large language models?

Taking stock from the previous section, we have three potential ways of determining what the target of representation of a LLM is:

- Ask what its content or structure is relevantly similar (or morphic) to;
- Ask how their users interpret it: either what they take it to be relevantly similar to, or what they take it or its parts to denote, and which concrete inferences they draw from them as a result;
- Enquire into norms of use and interpretation, as specified by its makers, some institutions, or entrenched practice.

Relevant for these options are the internal structure of LLMs, how they are used and how they are designed by their makers. This is what we will examine in this section.

The internal structure of LLMs

The basic function of LLMs, the one that is centrally involved in its various uses, is to "predict" the next word (as GPTs do) or any missing word (as BERT does) in a text (I will focus on predicting the next word in what follows for the sake of simplicity). What is processed by the

LLM is not actually a raw text, but a *tokenised* text, that is, a text cut into tokens selected in advance on the basis of statistical patterns in target languages. Tokens are usually words, pieces of words or punctuation marks, and although they often correspond to etymological or linguistic units, this is not always the case ("undesired" is tokenised as "und-es-ired" by some LLMs for instance). Just to give an idea, there are about fifty thousand different tokens in GPT-3. A LLM takes a sequence of tokens in input, and outputs a distribution of weights for the possible next tokens, representing their likelihood, in a sense to be discussed.

The models that made LLMs famous are based on the transformer architecture. Their internal structure typically comprises hundreds of billions of parameters. These parameters are distributed as follows[2]:

- A **word embedding table** assigns vectors to every linguistic token in a high-dimensional vector space (over ten thousand dimensions in GPT-3);
- A series of layers (about a hundred) are used to transform the vectors of the input sequence into an output sequence in a contextualised embedding space. Each layer consists of:
  - **attention heads**, where input vectors can influence each other contextually in some respect, depending on their respective values and positions within the input sequence (each attention head has a few millions of parameters that determine the "respect" in which input vectors influence each other and associated *attention patterns*, and in each given layer, about a hundred different heads are run in parallel and their contributions summed);
  - **feed forward neural networks** that transform each input vector independently of the others (there is only one per layer, but each can have more than a billion parameters: they contribute to about two thirds of the total parameters of the LLM in GPT-3).
- An **unembedding table**, usually the transposed matrix of the embedding table, converts the last vector of the output sequence into a weight distribution over possible tokens.

It can be intuitive to view embedding tables as encoding the meaning of tokens in a semantic space, attention heads as encoding conceptual relations between them as well as grammatical and near-pragmatic rules governing their linguistic use, and feed forward networks as storing the factual knowledge needed to predict the next word of a text. We can also postulate that the first layers attend to more local structures and concrete aspects of language, while the last ones attend to more global structures and abstract aspects. As will be discussed later, research attempting to extract linguistic information from LLMs tend to give results that go in these directions (see references in section 4, subsection "according to content views"). However, we must be careful in our interpretations and remember that LLM parameters are not fixed intentionally with a specific interpretation in mind by modellers. Indeed, what sets machine learning models like LLMs apart from other kinds of models is that they are not produced directly by human agents, but by an algorithm that "trains" these models on huge textual datasets.

[2] All orders of magnitude mentioned in this section are taken from the largest GPT-3 model. I give them so that the reader can get an idea of the dimensions that are typically involved, but one should keep in mind that this is a rapidly evolving technology and that these numbers can vary greatly from model to model.

How LLMs are produced: the training process

It is noteworthy that the datasets from which LLMs are trained are not brute datasets, but that they are cleaned up. In particular, (1) duplicated paragraphs are removed to improve performance, (2) texts containing blacklisted words, notably words with sexual connotation and insults, are excluded, and (3) “low quality” texts and texts whose language could not be identified automatically are also filtered out. The quality of a text is evaluated automatically by quantifying its similarity to texts from trusted sources such as Wikipedia (Wenzek et al. 2020; Dodge et al. 2021). Various authors note that in the future, it might be necessary to exclude texts generated by previous generations of LLMs as well, although interestingly, suitably generated texts are sometimes intentionally used to train new LLMs.

During the training phase, the learning algorithm adjusts the hundreds of billions of parameters of the LLM through *backward propagation* to optimise their ability to predict any missing word in the texts of the dataset. Backward propagation is a form of feedback where a gradient is computed that represents an optimal direction of adaptation for the many parameters of a layer to minimise predicting error. The gradient is propagated from the last layers to the first ones on each iteration.

After this first training, the model is *fine-tuned* in order to specialise it for specific tasks or audiences, such as for example handling medical files, by making it integrate domain-specific knowledge and learn acceptable behaviour and expression style. A prominent type of fine-tuning involves human inputs using *reward models*. Reward models are themselves generated by machine learning on the outputs of a first, “non-aligned” instance of a LLM using human feedback to train them. Their function is to predict how a human agent would evaluate the quality of any generated text. Then the LLM instance is trained again on a dataset, but this time optimising its capacity to generate texts that are well evaluated by the reward model. Another type of fine-tuning involves methods for inducing instruction-following capabilities (“self-instruct”), which notably ensure that the LLM, when given instructions as input, will follow them instead of attempting to complete them.

Fine-tuning techniques introduce intentional biases into the LLM. They are clearly aimed at particular uses for which the models are intended, notably chatbot applications. The basic function of a LLM is to predict the next or missing token in a text, but they have other “emergent” capabilities. In practice, they are used to simulate discussions, explore ideas, generate purpose-specific texts that follow style and content instructions, retrieve information, answer questions about a text, summarise it, improve its style, translate it to another language, etc. (Manning 2022) Ultimately, LLMs can serve as assistants for any kind of text processing activity, with more or less success depending on the task at hand. This concerns natural languages, but also programming languages when LLMs are used as software development assistants. New generations are multimodal: they integrate image and speech recognition capacities. However, I will leave these aspects aside in the following analyses to focus on text completion.

How LLMs are used in practice

In practical uses, LLMs do not merely predict the possible next tokens of a text. They are used to generate longer texts. The procedure goes as follows: the LLM is first fed with an input sequence of tokens, and it outputs a distribution of weights on all possible next tokens. These weights are not probabilities: they can take any decimal value, including negative ones. A decoding function with a “temperature” parameter transforms them into probabilities, which are then used to select one of the possible tokens randomly. Finally, the selected token is added to the original sequence, and everything is fed back into the LLM as a new input to process. This procedure is repeated until a certain amount of text has been generated.

The decoding function plays a crucial role. If the token with the highest weight is systematically selected, the procedure often gets stuck in repetitive loops of one or a few tokens instead of generating meaningful text, but if any token can be chosen based on its probability weight only, the generated text quickly becomes incoherent due to the vast number of possibilities (Holtzman et al. 2019). Various decoding methods based on repetition avoidance and tail suppression (ignoring low probability tokens) are used to prevent both outcomes. These methods can be quite sophisticated. The parameters of the decoding function, notably temperature, can be adjusted to generate more predictable or “creative” texts.

In actual applications, LLMs are typically given a *prompt*, a text that precedes user input and sets up a conversational context depending on the intended use. In chatbot scenarios, the prompt usually describes two participants in a conversation, where the user will in effect play one role while the other, usually described as a “helpful AI assistant”, will be generated by the LLM. Each time the LLM must generate its part, it receives an input consisting of the prompt and either the full conversation or only its final segment if it has become too long. Some prompt techniques have been developed to improve the quality of discussions, notably *chain of thoughts* prompts, which mention in the setup of the conversation that the AI assistant proceeds step by step and explains its reasoning process. This reduces inaccuracies in the generated texts.

It is not clear whether the prompt, parameters, or even the algorithm used in the decoding function should be considered an integral part of the LLM, since they can change from one application to another. It seems reasonable to view them as mere components involved in its use: an interface between the model and its users. If so, strictly speaking, the LLM is only responsible for calculating a distribution of weights on possible next tokens from an input sequence of tokens, and nothing more. However, a LLM is not really usable without these components. It is also worth noting that prompt and fine-tuning techniques roughly share the same goals, namely, to adapt the LLM to specific tasks, which can sometimes be achieved in one way or the other (fine-tuning is costly but gives better results).

4 What do large language models represent?

We can see that LLMs have peculiarities that set them apart from other kinds of models in science and engineering, most notably the fact that they are machine learning models, i.e. they are generated by an algorithm rather than being directly designed, which introduces a form of opacity in their content, but also the fact that they are tuned towards certain uses that typically mimic conversation with an agent. As a result, the question of what they represent is not so easy to answer. Let us apply the various theories of representation reviewed in section 1 to see if they can help us in this respect.

According to content-views

The first approach, associated with content-views, posits that a LLM potentially represents anything that has the same or a similar structure. This structure is constituted by the hundreds of billions of parameters of the LLM, qualified by their functional role in the text generation process (the fact that they are used to embed tokens, make them influence each other or transform them separately, and that many matrices act on the same vector space).

What seems to tell against content-views is that the architecture of LLMs is largely arbitrary with regards to plausible targets of representation. For example, the fact that they have, say, 96 layers and that each layer has 96 attention heads, or more or less for different LLMs, is not a matter of fitting anything in the world but just a matter of performance. Now perhaps this is not an issue: one could postulate that these parameters eventually converge towards some form of similarity to *something*, regardless of such architectural choices (maps and pictures have arbitrary resolutions as well after all). But even granting this, the question remains: a similarity to *what*? Ideally, we would like it to constitute a well-defined natural entity.

We can presume that the internal structure of LLMs reflects something that was extracted from the training dataset. One option is to claim that the LLM represents mere statistical patterns in this dataset. A more optimistic option, which can be motivated by the impressive emergent capabilities of LLMs, is to assume that they represent not just mere patterns, but rather what explains or produces these patterns: namely the structure of natural languages, reflected in the dataset. The dataset would itself be representative of a larger set of potential texts that would constitute a natural class, the class of all well-formed and conceptually sound texts of a given language perhaps, and the aim of LLMs would be to capture the essential structure of this class through some form of inductive process implemented by the learning algorithm.

Note that with content-views, the precise way in which the model is similar to the target of representation needs not be transparent to users: all that matters is that users can exploit this similarity to make sound inferences about the target.

This optimistic view implicitly assumes that (1) natural languages are well-defined entities with specific structures, corresponding to a natural class of linguistic phenomena, (2) the dataset is representative of this class, (3) this is sufficient for the learning process to induce the structure of this natural class, and (4) it is this specific structure that is actually exploited by users in standard uses. We could wonder to what extent these assumptions hold.

Regarding (1), it is not necessarily obvious what the structure of a natural language is and if such a thing exists. We can think of it, on a first approach, as a normative structure: a set of conceptual, semantic, near pragmatic and grammatical norms that a speaker must integrate in order to count as a competent speaker, independently of the various locutory intentions she could have. This structure should therefore ideally exclude any factual knowledge that is contingent with respect to the proper use of a language, but include conceptual knowledge. Deniers of the analytic/synthetic distinction are likely to object that there is no such clear-cut distinction between the two. Others might doubt that natural languages have such a unified normative structure: they might be more nebulous entities. But if what is represented is not a well-defined class with a specific structure, then claims of similarity become less tractable. Holders of content-views should therefore maintain that natural languages, as studied by linguistics, are natural classes.

Assuming that this is viable, the next question is (2) whether the dataset is representative of natural languages. A potential issue is that tokens do not always match proper linguistic units such as prefixes. This already creates a structural mismatch between the dataset and the target language. Another issue has to do with data cleaning. Part of it seems perfectly legitimate for the purpose of getting a representative set (eliminating the texts from menus and footers of websites for example), but does selecting high-quality texts, where specific sources like Wikipedia serve as benchmarks, and suppressing sexual content and insults make the dataset more representative of our languages overall? Perhaps this is the case, but one could suspect that it inevitably conveys a certain value-laden vision of what is meant by "the structure of our languages", by undervaluing linguistic constructions associated with specific social groups for example.

Perhaps these issues do not constitute an insurmountable obstacle if datasets are sufficiently large and varied. However, a further problem is that even very large datasets are likely to reflect much more than just the structure of natural languages. In this respect, it is noteworthy that LLMs incorporate a large amount of factual information. If you ask a LLM about the planets in our solar system, for example, it can give you the list of these planets, mention that Pluto was excluded from the list by the International Astronomical Union in 2006, and provide you with a mnemonic sentence to remember them: all this is only contingently associated with the meaning of "planet", "our" and "solar system". So, if the dataset is representative of something, we are talking about a mix of linguistic, factual, and even anecdotal aspects of linguistic productions. The question becomes whether the learning process can tell these aspects apart, so that specific component parts of LLMs end up representing genuine linguistic entities, or if what is represented is actually much more nebulous.

This leads us to assumption (3): that the learning process is capable of inducing the normative structure of natural languages from the dataset. The fact that increasing the number of parameters improves the efficiency of LLMs is not necessarily a good sign: it could mean that LLMs superficially mimic linguistic constraints by reproducing their various effects instead of actually capturing them. However, recent research gives us reasons to be more optimistic. Manning et al. (2020)[3], for example, show that specific attention heads of BERT-type LLMs attend to specific grammatical relations, such as between a verb and its direct object or between a determinant and its noun. They also demonstrate, using probing techniques, that there is a latent space in the contextual embedding space in which the vectors associated with input words have distances that systematically match their distances in syntactic trees created by human linguists. This seems to indicate that during specific uses, LLMs (at least BERT-type ones) do represent the linguistic structure of the sentences that they process. We can presume that if they have this emergent capacity, this is because they have integrated the normative structures of our languages by induction from the dataset. Another interesting result is that the first layers of LLMs are less affected by fine-tuning, which can make us think that the structure of languages are represented in these layers specifically, and that LLMs do tell them apart from other, more abstract and language-independent aspects (Hao et al. 2019).

All this makes it plausible that LLMs, or some of their component parts, are structurally similar to the normative structure of natural languages: they *potentially* represent them, according to content-views. However, remember that according to these views, similarity is

[3] I am thankful to an anonymous reviewer for drawing my attention to this research.

necessary, but not sufficient. Users must also exploit the relevant similarity when using the model.

This leads us to assumption (4). Our capacity to extract linguistic structures from LLMs using probing techniques means that it is possible, in principle, to use LLMs to make inferences about natural languages, or about specific linguistic productions. This is an interesting result. However, our focus here is not on what is possible in principle, but on the way LLMs are actually used. We have seen that the structure that is involved in normal use contains much more than mere linguistic information: it also incorporates a huge amount of factual and anecdotal content. Furthermore, fine-tuning techniques are clearly aimed at biasing the models for specific uses, such as serving as "helpful assistants" that respond to instructions. If anything, these techniques should make the overall structure of fine-tuned models less like the general structure of our languages, which become somehow buried in them. If the structure of the LLMs as a whole is similar to anything, this must be at best a mixed bag of heterogeneous rules, values, design aims and habits on top of linguistic norms, as well as factual and anecdotal information. Thinking in terms of similarity or morphism to a specific object in this context becomes quite unfit and stretches the spirit of content-views beyond recognition. This casts doubts on the appropriateness of these views when it comes to determining what LLMs represent.

According to intented-similarity-views

A way out of this difficulty is to consider it sufficient that modellers merely postulate a relevant similarity relation between LLMs and an ideal linguistic structure, in line with a version of intention-view. Users would assume, for the purpose of language processing tasks, that the model correctly represents some linguistic structure in virtue of being similar to it.

There are two main differences between intended-similarity-views and content-views. Firstly, intended-similarity-views do not require an actual similarity for the model to represent its target: the similarity only needs to be postulated. Secondly, users must be able to clearly specify the respects in which the model is supposed to be similar to its target instead of merely exploiting the similarity, otherwise the view would be empty.

The first difference might help mitigate some of the difficulties we have encountered so far. The fact that LLMs are fine-tuned for local purposes, such as serving as assistants, or that they also integrate non-linguistic aspects could imply that they misrepresent their target, which is not a problem for intented-similarity-views. One could worry that fine-tuning is intentional, so that modellers cannot faithfully believe in the existence of such a similarity between LLMs and the structure of our languages. However, this concern is less severe than it initially appears: arguably, many scientific models intentionally misrepresent reality by introducing idealisations (dimensionless massive objects, infinite gases, etc). Some of these idealisations serve to facilitate model use, making calculations tractable for instance. Perhaps the same could be said of fine-tuning techniques and of factual content: their aim would be to facilitate the use of LLMs, at the cost of introducing harmless distortions (or just surplus structure) with regards to their intended target. In this context, decoding functions and prompts can be considered idealised parts of the model after all: given the emphasis on concrete uses in intention-views, it is not particularly problematic that they can change from one use to another. Or they could be compared to the way model parameters, such as the length of the string for a pendulum, are fixed during concrete applications.

The main problem with intended-similarity-views concerns the second difference with content-views: the fact that users must be able to specify the ways in which the model and target are supposedly similar. Hypothetical similarity is supposed to serve an epistemic function: by looking at the model, we can learn about the target, assuming that it is similar to the model in relevant respects. But, first, it is not obvious that users actually postulate that LLMs represent natural languages (or any other entity) when using them, and second, even if they wanted to make such a postulate, the notorious opacity of trained models threatens to make it impossible. In the context of LLMs, any similarity is hypothetical not only because of our ignorance of the actual structure of the target, but also because of our ignorance of the structure of the model itself, or at least the respects in which it would be similar to the target (what Emily Sullivan (2022) aptly analyses as an uncertainty in the model-target connection). This is because the model was produced through an automated process rather than by human agents directly integrating their postulates into it.

Opacity in modelling is not necessarily specific to machine learning. Physical models, such as the U.S. Army Corps of Engineers Bay Model (a hydraulic scale model of the San Francisco Bay) or the Mississippi River Basin Model (Boesch 2022), rely on physical processes for their functioning, namely water flows. These processes can be opaque to users. We do not need to know the exact laws of fluid mechanics for the model to perform its epistemic function; all that is required is to assume that the hydraulic processes in the scale model are similar to those in the target system. The same goes when mice are used as models of human organisms: there is an assumption of similarity, even though underlying processes may be opaque. However, in these cases, we can clearly identify the respects in which source and target systems should be similar: flow dynamics or virus propagation for instance. Is it the case with LLMs? Users are not able to provide any explicit mapping between their billions of parameters, or their internal representations during functioning, and specific aspects of a hypothetical grammatical or conceptual structure. If some similarity to something else is postulated by modellers, it seems too underspecified to serve any epistemic function.

Sullivan argues that opacity can be mitigated by empirically establishing robust connections between model and target by other means. The results discussed previously (such as Manning et. al.) on the possibility to extract linguistic structures from LLMs are certainly relevant in this respect. However, they do not entail that we fully understand them as it stands (not to mention the average user). Yet LLMs are used on a daily basis. This means that intended-similarity-views cannot account for the way LLMs actually represent.

According to intended-interpretation-views

A similar problem occurs for accounts of representation that require an overly substantive interpretation of the model in terms of the target, such as denotation relations between component parts of sources and targets that would support inferences. However, there are clearer denotation relations between tokens and their linguistic counterparts: the words or pieces of words they encode. They can be exploited to devise an intended-interpretation-view which is unaffected by problems of opacity.

Tokens and their embeddings are the building blocks of LLMs after all, and we can view them as the main fixed points of reference to something external to the model. This implies a more instrumentalist approach, where the text inputs and outputs of LLMs are the true vehicles of representation instead of their internal structure (and decoding functions must be considered full part of the model). This focus on concrete use aligns well with intention-views, and in a sense, this should suffice for interpreting LLMs in terms of something. But in terms of *what*?

I contend that the two main options at our disposal are the following: either LLM inputs and outputs represent the world by linguistic means, or they are meta-representational, so to speak, in that they represent linguistic entities, such as words and sentences, that might themselves represent the world. Consider the following completion example: the input is "Venus is a", and the output is "planet". In the first case, the token "planet" represents a type of astronomical object. In the latter case, it represents an English word (this difference is somehow analogous the use--mention distinction).

There is a burgeoning literature on whether the production of LLMs is meaningful and whether produced terms refer to the world, where externalist strategies are typically deployed to argue that they do (Koch forthcoming; Titus 2024; Grindrod 2024; Lederman and Mahowald 2024). Among our two options, the idea that LLMs directly represent the world implies that their production is meaningful text. However, the idea that they are meta-representational is a priori neutral: it is compatible with the idea that the words and sentences that are represented are meaningful "in themselves", as externalists have argued, as it is with the idea that they only acquire meaning when interpreted by a user.

Intended-interpretation-views would sometimes align with this assumption that sequences of tokens represent the world directly, in so far as what matters for them is how users intend to interpret models. In chatbot applications, it could seem that the outputs of LLMs are interpreted by their users as if the LLM were an agent capable of producing meaningful text, an agent *holding* an internal representation of the world and making inferences instead of *being* a representational vehicle for its user to make inferences. Interactions with LLMs are apparently used to represent the world directly, to denote real entities, to learn about a topic by asking the LLM for example. Indeed, as we have seen, prompts are purposefully designed to make users believe that they are conversing with an AI agent that is representing the world.

Whether this is a correct account of how users typically interpret their conversations with chatbots does not go without saying (we will come back to this), but in any case, I think that taking LLM outputs to be linguistic representations of the world can be classified as instances of misuse, in the same way as using a map of Lisbon in Brussels is an instance of misuse, because LLMs do not represent the world directly.

A first reason to think that LLMs do not represent the world directly is that, as we have seen, tokens do not systematically correspond to meaningful linguistic units. It is possible to represent something by representing its parts, so it is possible to represent words by means of tokens, but parts of words are not parts of the worldly entities they stand for, so tokens that correspond to parts of words cannot represent worldly entities directly. Perhaps what we should conclude is that LLMs slightly misrepresent the world, since many tokens actually correspond to proper meaningful units. However, firstly, it becomes puzzling, in this view, why LLM makers deliberately introduce misrepresentations into their products, and secondly, this would entail that these meaningless tokens do not contribute to the correct functioning of LLMs, which seems quite implausible. The most sensible interpretation is that tokens in general denote words and pieces of words rather than worldly entities.

A second reason to believe that LLMs do not represent the world directly is that they are trained on texts, not on direct "worldly" data. A machine learning algorithm trained on molecular data is naturally interpreted as representing aspects of molecules by means of its outputs, and similarly, LLMs trained on texts should be interpreted as representing something linguistic. Of course, some aspects of the world are reflected in our languages, and therefore it is possible to make inferences about the world by the mediation of inferences about language. The idea that a representation of latent information can emerge makes sense in the context of a similarity-view, but less so in the context of an intention-view: if what we wanted was a model of the world, we would have trained the model on “worldly” data directly, not on texts.

Finally, and relatedly, thinking that LLMs represent the world conflicts with how LLMs are called by their makers, which seems relevant in the context of an intention-view: they are large *models of language*, not models of the world. Thinking otherwise seems deceptive regarding their functioning and training process, which is very different from the functioning and linguistic training of human speakers who use language to represent the world (Bisk et al. 2020; Bender and Koller 2020; Lake and Murphy 2023). For example, LLMs have integrated amounts of text orders of magnitude higher than anything a human could read in their entire life, but lack sensory inputs and motor outputs; they are static and do not possess any dynamic state apart from the conversation that is recursively fed to them, etc.

These three reasons may be debatable, but prima facie, they play in favour of considering that LLMs represent some linguistic phenomena rather than the world directly, that they are “meta-representational”, that is. If this is so, their tendency to generate confabulations or produce self-inconsistent texts should not be seen as a defect, since they are not designed to meet accuracy conditions regarding the external world (this has been analysed in terms of “bullshit” (Hicks, Humphries, and Slater 2024): our analysis mostly agrees, in particular in that tracking truth is not among the functions of LLMs; see also (Boisseau 2024)).

If this is correct, then some users misuse LLMs, and this poses a problem for intended-interpretation-views, since they cannot account for misuse.

According to norm-views

Let us take stock. Content-views seem flawed, because if there is similarity between LLMs and natural languages, it is buried in their structure and what is exploited by users is much more nebulous. Intended-similarity-views imply that user inferences would rest on a hypothesised similarity relation between LLMs and something else, which is implausible due to opacity. Intended-interpretation-views suggest that a LLM would sometimes represent the world directly, which contradicts their design and training. This leaves us with norm-views, according to which what LLMs represent is what it is appropriate to make inference about when using them, as determined by their makers, entrenched usage, or institutions.

Building on the arguments from the previous subsection, we can already narrow down appropriate uses to making inferences about something linguistic, and not directly about the world. This corresponds to how LLMs are designed, trained, named and described by their makers. Vendors might lead us astray by inciting misuse, but despite this and despite our previous remarks, taking LLMs to be about language is not necessarily at odds with entrenched usage: in practice, many people use LLMs for language processing tasks, because this is what they do best. They only converse with LLMs as if they were agents in a playful or instrumental way without taking it seriously.

In any case, norms of representation are responsive to criteria of adequacy, and in this respect, the idea that the inferential capacities afforded by LLMs are primarily about something linguistic is warranted. However, we need to say more: what exactly are these inferences about?

Considering for the moment that prompts and decoding functions are external to the models, these inferences are, strictly speaking, about what could come next after a given text. But in what sense of "could"? We have seen, when analysing content-views, that this cannot be *in virtue of the structure of natural language*, or not only, because LLMs integrate many heterogeneous aspects, notably due to fine-tuning (unless these are considered idealisations). A more plausible answer is that this "could" is essentially normative: what could come next in virtue of norms of appropriateness established by LLM makers. These norms are implemented first during data cleanup: trusted sources used to assess the quality of the texts would exemplify them and constitute a model to follow, contrary to sexual content and insults. The purpose remains quite general at this stage, and the norms are probably mostly cognitive, linguistic and conceptual. However, much more specific norms are implemented when the LLM is fine-tuned to serve as a "helpful" assistant that responds to instructions, and notably when human feedback is involved to evaluate the quality of its responses. The norms thus implemented are no longer merely cognitive and linguistic.

Our answer to what LLMs represent is thus the following: they each represent a certain class of appropriate linguistic production. We are talking about appropriateness *for a purpose*: not a natural class of linguistic phenomena, but a constructed one that depends on specific aims on top of linguistic, conceptual, factual, practical, moral and aesthetic constraints. This lack of naturalness is problematic for content-views, but less so for norm-views and their deflationary approach. We do not need to postulate any similarity to anything specific in the world: only that norms of representation be in place. Some structural similarity between component parts of LLMs and our languages (assuming they have a well-defined structure) can be involved in representation and motivate these norms, but they do not constitute the representation relation.

Remember the distinction we introduced between a semantic level, where the question is what a model represents in general independently of any context, and the pragmatic level, where we wonder what it is intended to represent by users in particular occasions. In norm-views, we can take the semantic level to correspond to a *kind* of appropriate use, and the pragmatic level to instances of this kind (in so far as the model is appropriately used). So, we can say that at the semantic level, a LLM represents, by means of its internal structure, a class of appropriate linguistic production. At the pragmatic level, users can pick or produce specific instances of this class to be represented by the inputs and outputs of the LLM (in the same way as someone can pick or create a pendulum to be represented by equations using the model of the ideal pendulum to generate these equations).

I say "pick" or "produce", but users seldom pick actual texts in order to predict their continuations, and the fact that the intended class is constructed would make LLMs unfit for this purpose: it is much more sensible to look at the original text if we want to know what comes next. In most if not all cases, the text represented by the user is *fictional* rather than actual. By this, I do not mean that the generated text represents fictional entities (it does not need to), but that the text itself is not an actual linguistic production, but an imagined one. When discussing with a chatbot, we should consider that we are engaged in a fiction: we are asked to imagine that a certain kind of conversation takes place, and the LLM is used as a tool to make inferences about how this imaginary conversation would typically unfold if the norms established by LLM makers were in place. In other words, the outputs of LLMs serve as props in a game of make-believe, and only our imagination is responsible for providing hypothetical referential links to the world. If LLMs are assistants, they assist us in our imagination. We can tune their parameters and prompts to specify better the kind of text that we are interested in imagining. This is a direct consequence of the fact that LLMs represent texts rather than the world.

Here, we reach the same conclusion as Fintan Mallory (2023): that chatbots are fictions. According to Mallory, fictionalism avoids two pitfalls: either attributing incredible intentional capacities to LLMs, or assuming that users engaging with them are massively deluded. These two pitfalls can be avoided without revising traditional meta-semantic theories by assuming that chatbot conversations are literally meaningless (because LLMs lack agency), but fictionally meaningful (so, users are not deluded)[4]. The fact that we reach the same conclusion even though we took a different approach, by analysing LLMs as representations instead of agents, reinforces this conclusion.

This is not the whole story though; we would miss an important aspect if we did not mention that fictions can be performative. This is true of scientific models, notably in engineering[5]: engineers planning to build a bridge imagine, with the help of a mechanical model, how the bridge could be configured and what characteristics, desirable or not, it could have. The bridge is a fiction until it is deployed as a real structure (and often, unexpected difficulties arise during this phase). Similarly, the fictional texts generated by LLMs can become actual texts: student dissertations, professional emails, motivation letters, computer programs, news or scientific articles or paragraphs within them, authored by the human agents responsible for their deployment.

The fact that LLMs work similarly to models in engineering rather than pure scientific models should not come as a surprise. LLM makers are engineers who aim to build technologies, not scientists seeking to acquire knowledge of the world. LLMs might be used to learn about the structure of natural languages, but this is not their primary function today. Deflationary accounts of representation are particularly well suited to account for engineering models in general, as illustrated by Suarez (2024)'s case study, as well as Boesch (2022)'s one. The latter explains in particular how constraints on the vehicle of representation are brought by manufacturers to serve specific and potentially evolving purposes, which is quite relevant to the case of LLMs.

5 Implications

We thus reach the main conclusion of this paper: LLMs represent, in general (at the semantic level), a constructed class of appropriate linguistic production whose characteristics are defined by their makers, and in context (at the pragmatic level), they are used to represent imaginary instances of this class, notably fictional conversations in chatbot use, that can be later appropriated for performative use. This is captured by norm-views, according to which what a model represents is what it is appropriate or normal to use them to make inferences about.

A similar analysis could be applied to other types of machine learning products, such as

[4] I'm grateful to an anonymous reviewer for drawing my attention to this work.

[5] A comparison with fictionalism about scientific models is also proposed by Mallory.

AlphaGo or image generators such as DALL-E. I believe that the same conclusions will hold: to the extent that they are not trained on direct “worldly” inputs, they are not agents playing Go or representing the world in images; they are models of Go games and images that can be used by human agents as supports for their imagination, and potentially become performative when adopted. This sheds light on the status of these entities, but also vindicates norm-views as the best conceptual approach for thinking about representation and models, including in engineering.

By means of conclusion, I would like to address ethical issues associated with the increasing deployment of LLMs and similar technologies in many areas of our social life. There are important ethical concerns related to the training of LLMs, notably copyright issues, environmental issues and the exploitation of human workers for fine-tuning purposes. However, the analyses of this paper have more bearing on the ethical issues associated with the use and deployment of LLMs.

A first implication of our analyses is that values are conveyed when defining the notion of appropriateness implicit in dataset selection and fine tuning. Given that the texts generated by LLMs can be appropriated for real uses, these values will likely impact society at large, and they should be critically examined. Having said that, there is nothing specific to LLMs here: values are conveyed by many kinds of technologies. What is more specific is the uncertainty regarding whether intended values or other unintended ones are actually incorporated in the final product. LLMs are trained to mimic helpful agents that will not harm humans nor assist in criminal activities: this is the kind of linguistic production that we expect these models to be representations of, but perhaps models will always *misrepresent* this intended kind despite their training. Because of their opacity, it can be hard to know whether this is the case (Arvan 2024). Other values than these are also likely to be smuggled in through the dataset, along with racist, classist and misogynist biases for example (Yang et al. 2024).

However, the main implications of our analyses concern the delusion involved in thinking of LLMs as agents rather than models, and in assuming that they directly represent the world through language instead of representing words and sentences for their users’ imagination. We have seen that fine-tuning reinforces this delusion by making them better at taking on a role in fictional conversations. The ethical issues with this capacity are obvious when the fact that users are interacting with a LLM is deliberately concealed (something done on social media platforms for political propaganda). It has an additional harmful consequence beyond delusion: it spreads politically oriented speech in amounts that no human could produce, and potentially biases future generations of LLMs trained on social media conversations. This makes the need for a critical examination of dataset selection and cleanup practices even more pressing (the case of fake content generation to attract traffic at low cost poses similar problems).

However, more regular uses of chatbots are also problematic. Today, chatbots are commonly deployed as assistants in numerous applications, which means that users are strongly encouraged to misuse LLMs by treating them as communicating agents rather than text predicting tools. LLMs are also given relative autonomy, and sometimes used to make automated decisions, as if they were indeed agents. All this is problematic, for it promotes applications for which LLMs are fundamentally untrustworthy.

A first step to mitigate this issue is to avoid anthropomorphic talk when describing LLMs (Abercrombie et al. 2023), instead insisting on the fictional nature of their production and highlighting the crucial role our own imagination plays in the interaction process. There are also different ways to make the illusion more visible to users.

One could think that setting up a conversation with an “AI agent” in the chatbot application prompt is less delusional than if it were setup as a conversation between two humans, with the LLM filling in one of them. However, I would argue that this is quite the reverse. An LLM does not naturally “identify” with an AI agent: this is set up in the prompt, but if you ask one to complete texts such as “I am”, “Who are you?” etc. without any preceding prompt, it will not spontaneously mention any helpful AI agent in its completion. Most of the time, a fictional human protagonist will show up in the generated text, because this is what would naturally follow in their training data. So, technically, the two ways of setting up a conversation, either between two humans or with an AI assistant, are on a par: these are simply different fictional texts. But at least in the latter case, the fact that we are engaged in a fiction becomes obvious (to the extent that it is explicitly stated that we are interacting with an LLM). In contrast, presenting a conversation between an AI agent and a human deliberately obfuscates this fictional nature by making us think that the agent is real.

I believe that presenting fictional (helpful and honest) human characters instead of “AI assistants” would already be an improvement in chatbot applications. Of course, people might react differently. Imagine a scenario where a chatbot is deployed in a medical context, and that users are presented with a fictional human doctor instead of a genuine “AI doctor” that will help them solve their health issues. In this situation, users might feel deluded or infantilised by the setup and refuse to interact for that reason. But this is precisely the point: they would be entirely right that they are deluded, and this is true even when their interlocutor is presented as an “AI doctor”. The latter only makes it harder for people to realise this, by making the delusion more credible. It might be helpful and economical, assuming low stakes are involved, to make people imagine what a real doctor could say when confronted with their case, but there is no reason to delude them into thinking that an expert is genuinely handling the situation[6]. This shows that chatbot applications are fundamentally problematic, and even more so when they obfuscate their fictional nature.

I do not mean to deny the potential usefulness of LLMs’ generative capacities, but there may be more interesting applications than chatbots in which probabilistic completion plays a more transparent role, for example by allowing the user to visualise probabilities, as already allowed by some interfaces, and control the production of the LLM by choosing between possible alternatives. I believe that it is crucial to foster these alternative applications, with the goal of creating more beneficial and transparent technologies for society.

6 Acknowledgements

The author acknowledges the financial support of:

- the project PID2024-157876NA-I00 (Lógica del razonamiento médico. El papel de la abducción en las hipótesis causales), Programa Estatal para la Investigación y el Desarrollo Experimental, en el marco del Plan Estatal de Investigación Científica, Técnica y de Innovación 2024-2027

[6] Pattern recognition by machine learning can be useful for medical purposes. I’m only interested in the linguistic capacities of LLMs here.

- the project PID2024-159749NB-I00 (The Philosophy of the Statistical Modeling Natural Sciences), Programa Estatal para la Investigación y el Desarrollo Experimental, en el marco del Plan Estatal de Investigación Científica, Técnica y de Innovación 2024-2027
- the grant RYC2023-042844-I (Ramón y Cajal), funded by MCIN/AEI /10.13039/501100011033 and The European Social Fund Plus (ESF+)

7 References

Abercrombie, Gavin, Amanda Curry, Tanvi Dinkar, Verena Rieser, and Zeerak Talat. 2023. ‘Mirages. On Anthropomorphism in Dialogue Systems’. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 4776–90. Singapore: Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.290.

Arvan, Marcus. 2024. ‘“Interpretability” and “Alignment” Are Fool’s Errands: A Proof That Controlling Misaligned Large Language Models Is the Best Anyone Can Hope For’. *AI & SOCIETY*, November. https://doi.org/10.1007/s00146-024-02113-9.

Bailer-Jones, Daniela. 2003. ‘When Scientific Models Represent’. *International Studies in the Philosophy of Science* 17 (1): 59–74. https://doi.org/10.1080/02698590305238.

Bartels, Andreas. 2006. ‘Defending the Structural Concept of Representation’. *Theoria* 21 (55): 7–19.

Bender, Emily M., and Alexander Koller. 2020. ‘Climbing Towards NLU: On Meaning, Form, and Understanding in the Age of Data’. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–98. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463.

Bisk, Yonatan, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, et al. 2020. ‘Experience Grounds Language’. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8718–35. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.703.

Boesch, Brandon. 2017. ‘There *Is* a Special Problem of Scientific Representation’. *Philosophy of Science* 84 (5): 970–81. https://doi.org/10.1086/693989.

———. 2022. ‘A Concrete Example of Representational Licensing: The Mississippi River Basin Model’. Studies in History and Philosophy of Science 92 (April): 36–44. https://doi.org/10.1016/j.shpsa.2022.01.002.

Boisseau, Éloïse. 2024. ‘Imitation and Large Language Models’. *Minds and Machines* 34 (4): 42. https://doi.org/10.1007/s11023-024-09698-6.

Callender, Craig, and Jonathan Cohen. 2006. ‘There Is No Special Problem About Scientific Representation’. *Theoria* 21 (1): 67–85. https://doi.org/10.1387/theoria.554.

Chakravartty, Anjan. 2010. ‘Informational Versus Functional Theories of Scientific Representation’. *Synthese* 172 (2): 197–213. https://doi.org/10.1007/s11229-009-9502-3.

Contessa, Gabriele. 2007. ‘Scientific Representation, Interpretation, and Surrogative Reasoning’. *Philosophy of Science* 74 (1): 48–68.

Dodge, Jesse, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. ‘Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus’. In *Proceedings of the 2021*

Conference on Empirical Methods in Natural Language Processing, 1286–1305. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.98.

Frigg, Roman. 2010. ‘Models and Fiction’. *Synthese* 172 (2): 251–68. https://doi.org/10.1007/s11229-009-9505-0.

Frigg, Roman, and James Nguyen. 2020. *Modelling Nature: An Opinionated Introduction to Scientific Representation*. Vol. 427. Synthese Library. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-45153-0.

Giere, Ronald. 1988. *Explaining Science: A Cognitive Approach*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226292038.001.0001.

Godfrey-Smith, Peter. 2007. ‘The Strategy of Model-Based Science’. *Biology & Philosophy* 21 (5): 725–40. https://doi.org/10.1007/s10539-006-9054-6.

Grindrod, Jumbly. 2024. ‘Large Language Models and Linguistic Intentionality’. *Synthese* 204 (2): 71. https://doi.org/10.1007/s11229-024-04723-8.

Hao, Yaru, Li Dong, Furu Wei, and Ke Xu. 2019. ‘Visualizing and Understanding the Effectiveness of BERT’. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4141–50. Hong Kong, China: Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1424.

Hicks, Michael Townsen, James Humphries, and Joe Slater. 2024. ‘ChatGPT Is Bullshit’. *Ethics and Information Technology* 26 (2): 38. https://doi.org/10.1007/s10676-024-09775-5.

Holtzman, Ari, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. ‘The Curious Case of Neural Text Degeneration’. arXiv. https://doi.org/10.48550/ARXIV.1904.09751.

Hughes, Richard. 1997. ‘Models and Representation’. *Philosophy of Science* 64: S325--S336. https://doi.org/10.1086/392611.

Koch, Steffen. forthcoming. ‘Babbling Stochastic Parrots? A Kripkean Argument for Reference in Large Language Models’. *Philosophy of Ai*, forthcoming.

Lake, Brenden M., and Gregory L. Murphy. 2023. ‘Word Meaning in Minds and Machines.’ *Psychological Review* 130 (2): 401–31. https://doi.org/10.1037/rev0000297.

Lederman, Harvey, and Kyle Mahowald. 2024. ‘Are Language Models More Like Libraries or Like Librarians? Bibliotechnism, the Novel Reference Problem, and the Attitudes of LLMs’. *Transactions of the Association for Computational Linguistics* 12: 1087–1103.

Levy, Arnon. 2015. ‘Modeling Without Models’. *Philosophical Studies* 172 (3): 781--798. https://doi.org/10.1007/s11098-014-0333-9.

Mallory, Fintan. 2023. ‘Fictionalism about Chatbots’. *Ergo an Open Access Journal of Philosophy* 10 (0). https://doi.org/10.3998/ergo.4668.

Manning, Christopher., Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. ‘Emergent Linguistic Structure in Artificial Neural Networks Trained by Self-Supervision’. Proceedings of the National Academy of Sciences 117 (48): 30046–54. https://doi.org/10.1073/pnas.1907367117.

Manning, Christopher 2022. ‘Human Language Understanding & Reasoning’. *Daedalus* 151 (2): 127–38. https://doi.org/10.1162/daed_a_01905.

Mundy, Brent. 1986. ‘On the General Theory of Meaningful Representation’. *Synthese* 67 (3): 391–437. https://doi.org/10.1007/BF00485942.

Poznic, Michael. 2016a. ‘Make-Believe and Model-Based Representation in Science: The Epistemology of Frigg’s and Toon’s Fictionalist Views of Modeling’. *Teorema* 35 (3): 201–18.

———. 2016b. ‘Representation and Similarity: Suárez on Necessary and Sufficient Conditions of Scientific Representation’. *Journal for General Philosophy of Science* 47 (2): 331–47. https://doi.org/10.1007/s10838-015-9307-7.

———. 2018. ‘Thin Versus Thick Accounts of Scientific Representation’. *Synthese* 195 (8): 3433–51. https://doi.org/10.1007/s11229-017-1374-3.

Ruyant, Quentin. 2025. ‘Two Senses of Representation in Science’. THEORIA. An International Journal for Theory, History and Foundations of Science 39 (3): 353–71. https://doi.org/10.1387/theoria.26040.

Suárez, Mauricio. 2003. ‘Scientific Representation: Against Similarity and Isomorphism’. *International Studies in the Philosophy of Science* 17 (3): 225–44.

———. 2010. ‘Fictions, Inference, and Realism’. In *Fictions and Models: New Essays*, edited by John Woods. Philosophia Verlag GmbH. https://doi.org/10.2307/j.ctv2nrzgsf.

———. 2024. *Inference and Representation: A Study in Modeling Science*. The University of Chicago Press.

Sullivan, Emily. 2022. ‘Understanding from Machine Learning Models’. *The British Journal for the Philosophy of Science* 73 (1): 109–33. https://doi.org/10.1093/bjps/axz035.

Titus, Lisa Miracchi. 2024. ‘Does ChatGPT Have Semantic Understanding? A Problem with the Statistics-of-Occurrence Strategy’. *Cognitive Systems Research* 83 (January): 101174. https://doi.org/10.1016/j.cogsys.2023.101174.

Toon, Adam. 2010. ‘Models as Make-Believe’. In *Beyond Mimesis and Convention: Representation in Art and Science*, edited by Roman Frigg and Matthew Hunter. Boston Studies in Philosophy of Science.

Watson, Richard Allan. 1995. *Representational ideas: from Plato to Patricia Churchland*. Synthese library 250. Dordrecht: Kluwer Academic publ.

Weisberg, Michael. 2013. *Simulation and Similarity: Using Models to Understand the World*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199933662.001.0001.

Wenzek, Guillaume, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzman, Armand Joulin, and Edouard Grave. 2020. ‘CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data’. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 4003–12. Marseille, France: European Language Resources Association. https://www.aclweb.org/anthology/2020.lrec-1.494.

Yang, Yifan, Xiaoyu Liu, Qiao Jin, Furong Huang, and Zhiyong Lu. 2024. ‘Unmasking and Quantifying Racial Bias of Large Language Models in Medical Report Generation’. *Communications Medicine* 4 (1): 176. https://doi.org/10.1038/s43856-024-00601-z.