Are AI language models scientific models of language?

James Ladyman[1] and Ryan M. Nefdt[2,3]

[1]Department of Philosophy, University of Bristol
[2]Department of Philosophy, University of Cape Town
[3]Department of Philosophy, University of Bristol
[*]contact: ryan.nefdt@uct.ac.za/ryan.nefdt@bristol.ac.uk

Abstract

There has been a recent surge of philosophical interest in AI large language models (LLMs) and their relevance for the understanding of human language. In this article we take the perspective of the philosophy of science on modelling in general to consider whether LLMs are (good) models of language. We establish that LLMs are scientific models, and that they are good scientific models of certain aspects of human language, in so far as they can capture the modal structure of natural language. We argue that they can do this even if they do not reveal (all) its causal structure, nomological structure or mechanisms.

Keywords: *scientific models*, *language models*, *deep learning*, *human language*, *philosophy of science*

1. Introduction

There has been a lot of recent work about the role large language models (LLMs) can play in linguistic theory, philosophy of language, and the cognitive science of language in general (Baggio & Murphy, 2024; Baroni, 2022; Bender & Koller, 2020; Mandelkern & Linzen, 2024; Piantadosi, 2024). These accounts start from one or more of these fields and ask whether LLMs shed light on human language processing or acquisition, or possess properties such as semantic meaning or communicative intent themselves. Along the way, questions arise over whether LLMs are truly models of language (Millière, 2024; Veres, 2022), or whether their data-driven statistical natures prevent this somehow (Dupre, 2024; Katzir, 2023; Moro et al., 2023).

This article resituates the discussion by drawing on how models are understood in the philosophy of science. Adopting this perspective avoids the confusion in the debate hitherto due to different conceptions of what it means to model a phenomenon or system. Section 2 outlines some core aspects of LLMs. Section 3 presents a number of theses about scientific modelling in general, while section 4 explains the idea that (some) scientific models can represent modal information or structure (to different extents and in different ways). We use these considerations to suggest how models can be good (or better) at modelling a target domain. Section 5 argues that in the light of the theses about scientific modelling LLMs are scientific models of language varying degrees of fit depending on what the target of the model is taken to be. We also argue that LLMs can represent important modal aspects of natural language relevant to its scientific study.[1]

2. Large Language Models

Large language models (LLMs) are AI systems based on a particular kind of artificial neural network architecture (although other architectures are possible), and on machine-learning algorithms and very large data-sets. The models defined for natural language processing (NLP) are descendents of the connectionist models of the 1980s (Elman, 1991; McClelland et al., 1986), often known as 'artificial neural networks'. The earlier symbolic AI systems are brittle in the sense that one error or problematic transition can cause a halt or crash. Fixing them is like searching for syntax errors in complex program code. Neural networks are much more robust (in the sense of complexity science, see Ladyman & Wiesner (2020)), and modern networks use a host of nonlinear functions in order to fit the data. We briefly describe the paradigm of deep learning, and the kinds of architectural features common to most forms of AI (paying particular attention (pun intended) to transformer models) and the data regime for training these systems.

[1] It should be noted that our interest is not concerned with nomenclature. The word 'model' is polysemous within and outside of scientific contexts. Rather our aim is to investigate whether certain accepted features of scientific modelling are indeed present in the context of research on and with language models.

Deep learning is a type of machine learning using artificial neural networks. The bedrock of Language models is the use of deep learning to perform the task of predicting the next word in a sequence. Words have to be encoded in the form of tokens, and there is usually more than one token per word. Hence, the task of word prediction is turned into the task of token prediction. "[G]iven a sequence of tokens, [LLMs] are tasked with predicting which token is statistically most likely to follow" (Millière & Buckner, 2024, p. 6). LLMs find patterns in the corpora on which they are trained, and predict the next token based on what preceded it. 'Autoregressive' transformers (like the generative pretrained transformer (GPT) series) predict each token solely on the basis of previous tokens. In a landmark paper, McCoy et al. (2024a) argue this feature prevents such LLMs from performing basic functions such as counting words or reversing lists as well as simple reasoning tasks.[2]

Transformer LLMs can be related to two principles in linguistic theory. The first is redolent of older computational linguistic theory and shared by some other LLMs, namely the distributional hypothesis (Lenci, 2008) according to which words with similar meanings tend to occur in similar contexts. Collocation in a corpus can be used as a proxy for meaning (or even taken to be constitutive of meaning). In LLMs, tokens are converted to numerical vectors (called 'input embeddings'), where similar vectors represent words with similar meanings. Positional encoding is then used to encode the position of each word (in the input sequence) as an ordered set of numbers along with the input embeddings.

The second principle applies to all transformers (whether or not they are LLMs) and concerns the self-attention mechanism itself. Not all the tokens are equally relevant to the predictive task. In other words, some input vectors are more connected (have stronger weights) to one another than others. The concept of attention was introduced in the seminal Vaswani et al. (2017) as a mechanism for tracking the relative importance of parts of an input sequence. Attention is basically a mechanism that focuses on selective aspects of the previous context for determining the next token in a sequence, since not all words are equally important in determining said next word. Self-attention focuses on the relationships between words in the same sentence. Two things are important for understanding self-attention, (1) the query, key, value distinction, and (2) positional encoding. To oversimplify, a query is like a question, the key is the form of the answer and the value is the content of that form. The job of the query is to find all the best matches with keys (technically, it just computes these values for all pairs of words in the sequence). This is usually measured by the cosine similarity metric or the dot product between the vectors which encode each word. These mechanisms are used to calculate the attention score which determines the importance of different parts of the input data.

Much of the information about how the models are trained is shrouded in corporate confidentiality, but we do know that models such as Llama 2 have 70 billion parameters and were trained on 2 trillion tokens (Touvron et al., 2023) while GPT-4 is assumed to have over 175 bil-

[2]In a follow up article (McCoy et al., 2024b), they claim the problem improves but persists with the last (as of writing) GPT-4o model, which was specifically optimised for reasoning.

lion parameters, with 4o having around 200 billion (Ayub, 2023). This gives some indication of the massive scale of these systems. With great data comes great controversy. The greater orders of magnitude of the data (basically the whole internet in some cases) have suggested to many that LLMs are unlikely to be good models of human language (certainly of human language learning). But some initiatives such as the BabyLM challenge have been designed to confront some of these issues by limiting data to more realistic corpora more plausibly aligned with the average child's linguistic input. Deeper issues such as potential data-contamination (where the test set is possibly contained in the training set) are more difficult to ameliorate, especially given the information hoarding by those developing this technology. Commercially available models are also 'fine-tuned' in various ways the details of which are largely kept secret (as discussed further below).

Some further aspects of LLMs are discussed below in the light of the next two sections that consider models in science in general.

3. Scientific Models

Models are the subject of a vast literature in philosophy of science, and there are very different accounts of their nature, and whether and how they represent real systems, and there are very different accounts of scientific representation in general.[3] Some philosophers deny or downplay the representational role of models (for example, see Knuuttila & Voutilainen (2003) who regard them as epistemic artefacts). van Fraasen (2008) argues that models only represent given an agent with a purpose. This is the case because models generally serve different ends and are situated epistemic tools. For Morgan & Morrison (1999) models act as mediators between theories and the world. For Knuuttila (2011, 2020) they are 'erotetic devices' that are designed not only to answer certain theoretical questions or test hypotheses, but also to generate questions (Knuuttila & Merz, 2009). Many scientific purposes require what Weisberg (2007) calls 'multiple models idealisation', involving the construction of multiple connected, but potentially even incompatible models, each of which targets one or more aspects of the target system. This strategy does not involve "expecting a single best model to be generated" (Weisberg, 2007, p. 646). The idea is that since scientific theories are used for diverse purposes and the construction of a model necessarily involves trade-offs among criteria such as predictive accuracy, simplicity, scope, explanatory power, and so on. This kind of modelling is prevalent in climate forecasting in which prediction is often paramount.

Obviously, there are as many ways for models to be good as there are ends to which they can be put. What follows is about models in so far as they are sometimes representations of real phenomena and systems, and does not presuppose or entail that is their primary role (models are used for many other purposes), that they represent independently of being used to do so, nor

[3] A foundational work on scientific models is Hesse (1963); see also Cartwright (1983), van van Fraasen (2008) and Frigg & Nguyen (2020).

that all models are used to represent reality. Despite much disagreement and debate about such matters, a lot of consensus about the role of models in science has emerged from the attention to models and modelling across the sciences over the last few decades. It is widely agreed that:

1. In general, models as well as theories are required to represent actual systems precisely and accurately;
2. Models are (usually if not always) approximate and idealised in different ways;
3. Models can represent phenomena, and may or may not represent causal structures or mechanisms;
4. Models can represent without being isomorphic to what they represent;
5. In general, there can be multiple models of the same system that represent it in different ways;
6. In general, models do not represent every aspect of a system (incompleteness), and different models can represent different aspects of the system (partiality);
7. Different models are apt for different scientific purposes; and,
8. Models can have 'surplus structure' in the sense that not all elements of the model are representational.

For example, consider Newtonian gravitation applied to the motion of Mercury. The theory takes the form of laws about force and motion, as well as the specific force law of gravitation, and an estimate of the Gravitational constant. To apply it to the motion of a planet requires a representation of the positions of the planet and the Sun, as well as of their relative masses and initial positions and momenta – a model (1). The planet is represented as a point particle as is the Sun (since they are approximately spherically symmetric) (2). The model represents the force due to gravity (3), but notoriously does not involve a mechanism for how it is propagated. The model is not isomorphic to the solar system or even to the subsystem that is Mercury and the Sun (4), because it does not model the orbit of Mercury exactly (as discovered in the nineteenth century before the anomaly was accommodated by General Relativity). The model is heliocentric but there are also geocentric and geoheliocentric models of Mercury's orbit (5). The model does not represent everything about Mercury (incompleteness), and Mercury has a magnetic field that can be modelled with electromagnetism (partiality). (6) A Newtonian model is not apt for the scientific purpose of understanding the interaction of Mercury with the solar wind but an electromagnetic one is (7). The model represents the Sun as at rest in Euclidean space, but this can be regarded as surplus structure and does not have to be taken to represent absolute space (8).

(1) says that models are necessary, but sometimes they can also be sufficient for representation. For example, a map or diagramme is a not a theory, but it can be a model (Pincock, 2007). Going beyond (2), scientific models can be accurate while misrepresenting important features of the system. For example, the best geocentric astronomical models were predictively accurate to several significant figures for the motions of the heavenly bodies even though they

wrongly represent the Earth as being at the centre of the solar system. Brahe's geoheliocentric model was as empirically accurate as Copernicus's heliocentric model (but not Kepler's).

It is important for what follows that representing causal structure or mechanisms is not necessary for modelling (3). There are plenty of scientific models that represent measurable quantities and relations among them without representing causes or mechanisms. There are many models that are predictively successful, but which are not considered to be relevant to causation or explanation. Thus, successful, precise, quantitative prediction is sufficient for a model to be successful. The empirical adequacy is always to some degree of precision and always involves phenomena at some scale.

In addition, models can represent causal structure or mechanisms, while also misrepresenting important aspects of the system. For example, the flow of a liquid can be represented by a model that is continuous or discrete. As discussed below, this addresses some worries formal linguists have had about LLMs.

4. Modelling Modal Structure

Ontic structural realists emphasise how science represents 'modal structure' (Berenstain & Ladyman, 2012; French, 2014; Ladyman & Ross, 2007; Nefdt, 2023). Modal structure is understood as a general term that subsumes causal structure, mechanisms, nomological structure, probabilistic and statistical structures and so on. Others focus on causation (Cartwright), laws (Lange) or dispositions and powers (Bird, Chakravartty), but all these philosophers of science are realists about scientific modality in some form or other. Massimi (2022)'s recent exposition of Perspectivalism emphasises multiple models and perspectives, but is conducive to scientific realism since she takes there to be perspective-independent facts some of which are 'modally robust', and she explicitly endorses the idea of science giving us modal knowledge. Some prefer to talk about 'modal information' (Sjölin Wirling & Grüne-Yanoff, 2023), where, models providing 'modal information' usually refers to the idea of a space of possibilities, some of which involve nonactual states defined by the parameters of the models. There is a family resemblance among ideas of counterfactuality (Dohrn, 2023; Godfrey-Smith, 2006), fictions (Frigg, 2010), possibilities (Weisberg, 2007), real patterns (Dennett, 1991), and dispositional properties (Nguyen, 2020) that links these distinct (and prominent) views on the nature of scientific modelling to modality. Sometimes the idea of modal structure may be more apt than the idea of modal information or vice versa, but what follows does not turn on which term is used and those readers who do not like the idea of modal structure can think in terms of modal information instead.

The idea is that models can be taken to represent modal structure accurately in so far as they are projectible, which requires making successful predictions of measureable quantities. Assigning probability distributions to states (as is the case with the output of LLMs) generates predictive capabilities. Modal information or structure can be minimal, so, for example, even

a data model of the positions of the heavenly bodies on the celestial sphere could provide a basis for accurate prediction to some extent, simply because it may encode regularities that are projectible. Hence, even geocentric astronomy can be taken to represent modal structure to some extent (Saunders, 1993). However, in an accurate geocentric model the overall observable motion of the planets on the celestial sphere results from a complex combination of circular motions (deferents and epicycles, and these need not be taken to represent real motions. Next consider Kepler's model of the solar system. It is heliocentric and so predicts that Venus and Mercury should show a full set of phases like the Moon when viewed from the Earth. This fact is not an accident, it is a causal or nomological consequence of Venus and Mercury being inside the Earth's orbit around the Sun. To the extent that it is not an accident that these planets shows a full set of phases the model captures modal information or structure. However, this model of the solar system represents only kinematics (motions) not dynamics (forces), so does not capture the same modal structure as the central force model of gravity. The modal structure is different in Newtonian physics because the model includes causal structure in the form of the force of gravity, which makes the orbit of Mercury not exactly elliptical. Nonetheless, Kepler's laws remain a good approximation ('lossy compression') to the orbits of the planets according to Newtonian gravitation). Furthermore, Newtonian theory includes no account of how gravity propagates through space, while according to General Relativity it propagates by waves that travel at the speed of light. Hence, models in science can encode modal structure in various ways from the most minimal sense in which data models can encode projectible regularities in measureable quantities, to the thick sense of representation of a causal mechanism in great detail.

Recent work in the philosophy of physics emphasises that although the Standard Model of particle physics is highly predictively accurate and well-confirmed in a large domain of phenomena, its theories are 'effective' in the sense that they break down at higher-energies. This makes them like theories in the rest of physics and chemistry. Acknowledgement of the limitations (2, 3, 4 and 6) of models does not require antirealism and is compatible with regarding them as representing modal information. For example, the stability of the Earth's orbit and the fact that the Moon never displays its dark side to us is explained by gravitational tidal-locking in Newtonian physics, and that mechanism remains the explanation in current physics as an effective consequence of General Relativity. Baron et al. (2025) explicitly consider effective theories in physics as a source of modal information in line with 'effective realism' which is a way of thinking about scientific ontology that is not fundamental or universal (that is compatible with ontic structural realism Ladyman & Lorenzetti (forthcoming)). In sum, representation of modal information or structure is not all or nothing, and can be done by models that are predictively successful to some degree. Hence, the final thesis of scientific modelling:

9. Scientific models can represent *modal information* or *structure* about real phenomena and systems.

Successful prediction does not imply the accurate representation of modal information or structure, but persistent predictive success is taken by realists to be indicative of it. In causal modelling successful intervention and manipulation of target systems is taken to be indicative of the model representing causal structure. Knuuttila describes the importance of manipulation in modelling as follows:

> [S]cientists learn from models by constructing and manipulating them. From the perspective of learning, the epistemic value of modelling can be attributed to their manipulability instead of (a more or less) accurate representation. To be sure, something is represented in the model. (Knuuttila, 2021, p. 14)

The 'something' in question is modal structure or information. Chirimuuta (2024) develops what she characterises as a Kantian or transcendental account of neuroscientific modelling, which she terms 'haptic realism' (she believes her view is extendable to other sciences). The resulting metaphor for science moves from sight to touch.

> In other words, we should reject the traditional realist's conception of knowledge attainment as the picturing of objective facts, with its ideal of disinterestedness. Scientists learn about the world through tinkering and interacting with it, and these learning practices are bound up with their practical intentions. (Chirimuuta, 2024, p. 40)

In neuroscience modelling involves manipulation of the target system (as it does in many other sciences including chemistry and many parts of physics such as optics). Linguistics has traditionally favoured a more detached theoretical approach in which models act as mathematical representations of a cognitive system or language faculty. However, despite the initial promise of formalisation in linguistics, rigid rules and frameworks have proven an impediment to computational modelling (Müller, 2018). LLMs present new possibilities for the manipulation of language. For instance, using LLMs researchers can investigate which parts of the models are representing particular structures by intervening on activation patterns. Some of these techniques draw directly from early neuroscience, such as the concept of ablation. Prominent NLP researchers have even emphasised the model organism nature of LLMs (Tuckute et al., 2024). While in classical linguistic modelling, aphasiology is a passive process, i.e. when impairment occurs linguists study their effects, LLMs allow direct manipulation and evaluation of causes and effects on linguistic tasks. As Millière (2024, p. 20) notes "these causal interventions go a long way towards establishing the claim that language models do represent syntactic features".

5. LLMs as Scientific Models

The idea that LLMs are scientific models has received some treatment in the literature.[4] Meskhidze (2023) argues that machine-learning can provide understanding in cosmology. Cichy & Kaiser (2019) make a compelling case for why deep learning networks (ANNs) in general should be considered scientific models. Their view resembles Weisberg's 'multiple model idealisation'. Their case for the explanatory power of ANNs in terms of teleology, and their general optimism with relation to opacity, is more questionable. It is also not clear that "DNNs [ANNs] are effectively used in the same way as traditional mathematical-theoretical models" (Cichy & Kaiser, 2019, p. 311) given their engineering goals and industry-related objectives. They do not consider transformer-based LLMs directly which are the subject of much of the current controversy.

LLMs can be trained both to predict missing words in a document (the cloze or occlusion task) and the next word in a text, and they can become highly predictively successful at both tasks (and outperform humans at them). In this minimal sense they are good scientific models of a particular (massive) dataset of text in natural language. Most discussion concerns whether they are models of language in the stronger sense of being good models of language itself. There are two possibilities: (1) LLMs model human linguistic competence/cognition; and, (2) LLMs model human linguistic phenomena in some more abstract sense. We argue that (2) is more likely if we interpret this abstract sense as modal structure or information. In so far as LLMs predict plausible human linguistic responses to questions and other expressions they capture this modal structure, but they only provide indirect access to the possible rules of grammar in the cognitive sense given their opacity and the potential disconnect between grammar rules and probabilities in performance.

In the following, we discuss three different conduits to modal structure. The first is a direct mapping or modelling of the pathways in a system (or dataset), leading to predictive successes. In section 5.1, we argue that LLMs are good models in this latter sense of the corpora on which they are trained – corpus models – because they are highly predictively successful with respect to them. The second conduit concerns counterfactuality. Here models can delimit (or expand) the space of possibilities by testing hypotheses at scale. Thus, in section 5.2 we argue that they have and can serve as good models of language learning in this suitably qualified sense. The difference between how LLMs are trained and how humans learn leads to more indirect structural relationships. Lastly, we explore the extent to which LLMs model language itself, either as an abstract space of alternative forms or as an internal system of competence. We suggest cautious optimism about whether LLMs are models of modal structure in this stronger sense.

[4] We do not consider the possibility that LLMs are scientific theories here. Although see Baroni (2022) for that view. We tend to side with Potts (2024) that LLMs clearly have some contribution to make to linguistic theory but fall short of full theoretical status.

5.1 Good models of corpora

On the most minimal conception of LLMs as scientific models, they are models of the data and nothing more. Veres (2022) provides a thorough critique of certain theoretical claims concerning LLMs. First, he mounts an innovative argument against claims that ANNs bypass symbolic representation as the programming code underlying their construction is based on phrase structure grammar. We do not consider that argument here. However, a corollary of this line of reasoning concerns the nature of language models themselves. Veres claims that, given that they reflect selectional biases of the corpora upon which they are trained, they should rather be considered 'corpus' models. He states that "since the data is their only resource for learning, they can be susceptible to the characteristics of the data set" (Veres, 2022, p. 8). Thus, the idea is that the correct answer to the question of what LLMs are models of, is 'linguistic corpora' and not language itself. Specifically, he concludes:

> [W]e would suggest a clarification in terminology, and propose a change from the theory-laden term *language model* to the more objectively accurate term *corpus model*. Not only does the term *corpus model* better reflect the contents of models, it also provides transparency in discussing issues such as model bias. (Veres, 2022, p. 8)

Of course, the philosophical questions remain the same whether you call LLMs 'language models' or 'corpus models', although admittedly the former might invite more assumptions. Nomenclature aside, Veres could mean that LLMs are merely descriptions of data. In discussing the theory crisis in psychology, van Rooij (2022) makes a similar point: "[i]t [LLM] models the regularity in a set of data points. It can be a convenient summary of data" (van Rooij, 2022, p. 128). She insists they such models fail to explain why the pattern is the way it is. [5] Again, much of the literature in the philosophy of science concerning real patterns stress the compression role that scientific models play (Millhouse, 2022).

Thus, we might have arrived at the first point of applicability. LLMs are good models of the very large corpora on which they are trained in the sense of being predictively successful when tested on it. However, the corpora are themselves samples of the bigger data set of all actual linguistic output, and LLMs are good models of that to the extent that they are predictively successful when applied to data-sets from beyond the corpora on which they are trained. If its training corpus is biased then the LLM will inherit that bias, but it can still be a model of all actual linguistic output to some extent (given 2 and 4). Some of the statistical relationships among the elements of the corpora are to some extent found more generally.

Some argue that LLMs do not exceed the bounds of their training corpora (Bender & Koller, 2020). However, this account of their limitations neglects a number of components of modern

[5] If corpora themselves are taken to be models of language, in some sense, then LLMs are models of models of language. Arguably, corpora are closer to representative samples than models.

LLMs. It is not true that 'the data is their only resource for learning'. Transformer-based models such as GPT-4 are not just trained on corpora, but also, as noted above, fine-tuned. The standard fine-tuning process is called 'reinforcement learning with human feedback' (RLHF) and filters for honesty, helpfulness, and harmlessness. This involves direct human interaction with the model with the aim of improving performance, and also arguably grounding the model in real-world facts. Other techniques like Retrieval Augmented Generation (RAG) in which model performance is improved by allowing LLMs to access and use online tools in real-time for better accuracy also takes them beyond mere data models and allows worldly states of affairs to affect directly the linguistic output of LLMs.

In so far as models represent real systems they are approximate and idealised (2), need not be isomorphic to them (4), and do so partially and incompletely (6). Not all the features of the model reflect that system, nor are models usually designed to capture all the features of the system. For example, statistical models in Big Data cosmology are not models of the solar system or particular galaxies in any direct way, rather they are models of the vast amounts of amassed data from telescopes and other astronomical instruments. Physicists look for patterns in the statistical models that reflect aspects of the universe or verify theoretical posits. This job can, and has been, outsourced to appropriately parameterised models (in many cases ANNs see Meskhidze (2023)). But of course it would be wrong to insist that those models are not models of the universe because they may incorporate features not present in the target system. In the philosophy of science, such features are called 'artifacts' of the model or 'surplus structure' (8). Their existence does not preclude the possibility that models reflect, represent, or capture aspects of the target systems (i.e. human language in our case) they model. Two things seem clear at this stage, the corpus has modal structure otherwise it wouldn't be predictable (that's what we mean by modal structure) and the modal structure of corpus can't be completely independent of the modal structure of language (even considering noise).

5.2 Decent language learning models

The next possible sense in which LLMs are good models of language is based on the kind of computational models they are. LLMs (and ANNs generally) are learning models. Specifically, they are machine learning models. The central idea behind such models is that a given dataset is split between training and testing. The algorithm is then trained on the first set to find patterns applicable to the discovered on features in the set upon which it was not trained. Most modern LLMs are unsupervised learning models in that they do not received labelled or annotated input. The overall point under this interpretation of the scientific modelling of LLMs is that they can tell us something theoretically relevant about the process of language learning given their apparent success in producing fluent text without supervision or strong inductive biases.

One early (and often misinterpreted) result on language learnability is found in Gold (1967). Gold's theorem has been taken as vindication of the rationalist interpretation of language ac-

quisition and as proof of the poverty of stimulus argument in generative linguistics (Chomsky, 1986). The basic set up is as follows: a *teacher* provides a *learner* with some sequence of strings of a formal language (stringset=language). Learning here is idealised as an infinite process. The learner's task is to find a representation or formal grammar for each element of the string sequences it encounters incrementally. In other words, "[w]e may represent a learner as any function that takes finite initial sequences of an environment as input, and yields as output a guess as to the target language" (Johnson, 2004, p. 574). For Gold, the learner only has access to the text or strings of the languages (and whatever built in biases it comes with). 'Identification in the limit' is the idea that a learner will correctly guess (or converge on) the target language L after some time t_n upon exposure to the environment of strings of L. This inference is monotonic and should not change after t_n. The essence of the theorem is then that there are some classes of languages for which no amount of exposure will result in a correct guess by the learner.[6]

The general idea of this result or learning model is that any learner (human or artificial) cannot learn a language only by exposure to the actual sentences from the language. Naturally, rationalists about language have taken this to mean that without negative evidence, some innate structure must be assumed of child language acquisition (Marcus, 1993). However, as Pearl (2021, p. 428) notes "Gold's result was based on assumptions about the nature of the learning process that are unlikely to be true for children". Specifically, she claims that the model drastically idealised away the learning situation children actually face. For example, the environment Gold takes to be the learning environment, basically just exposure to positive strings, is not a realistic picture of the information available to a human child.

Early computational learning theory (the legacy of Gold's theorem) might have been limited. However, Lappin & Shieber (2007) believe that modern grammar induction models with ANNs are more informative. In the first place, they claim that experiments based on machine learning show that the inductive biases required for learning are distributional and not categorical (which is a special case of the former). Furthermore, they claim that, based on work in unsupervised grammar induction, successful learning requires even weaker biases.

A distributional bias is a linguistically motivated bias that "words form distributional patterns by virtue of falling into classes" (Lappin & Shieber, 2007, p. 10). With such weak biases in place, the classic poverty of stimulus arguments don't get off the ground. LLMs inherit these weak biases and *can* thus tell us something about the nature of learning and *a fortiori* something about the strong theoretical claims of innateness and the rationalism which underpins it, namely that they can be questioned. Recently, Piantadosi (2024) not only endorses LLMs as plausible models of language learning and acquisition but even argues that they refute generative linguistic theory in terms of the strong innateness it assumes *inter alia*. One

[6]The details of the proof will not detain us here. However, as Johnson (2004) correctly points out, the property which generates the theorem is a relation between languages and not the learnability of individual languages, as is often assumed in the literature.

way to appreciate the advantage of LLMs for him involves "how massively over-parameterized models like these work is that they have a rich potential space for inferring hidden variables and relationships" (Piantadosi, 2024, p. 6). According to him, this aspect of LLMs allow the models to show us what is possible and their precision makes testable comparisons with human behaviour fruitful. This dovetails with Cichy & Kaiser (2019)'s 'exploration' avenue for the cognitive scientific significance of ANNs in terms of Hesse's concept of neutral analogies. For Hesse (1963), neutral analogies involve cases in which it is unknown whether the model and the target share certain properties. Thus, they offer us some possibly new hypotheses.[7]

Despite the scope for LLMs to delimit the logical space of learning, their precise training regime and internal mechanisms pose a problem for interpreting them as realistic learning models. In terms of the former, as mentioned in section 2, most language models are trained on vast amounts of data, the extent of which far outstrips any reasonable learning environment that humans might encounter in their lifetimes (or even multiple lifetimes). Then, in spite of notable progress on interpretability and explainable AI, LLMs are still largely considered black boxes with relation to their inner workings. There are, of course, many attempts within the natural language processing literature to surmount both challenges.[8] Nevertheless, what remains is not a question of the performance of LLMs on linguistic tasks (which is impressive) but rather a serious quandary as to the commensurability of their internal mechanisms and learning environments (very large sets of text) with what we know about our linguistic cognition. In a recent review, Tuckute et al. (2024) highlight further incompatibilities between human language learning and that of LLMs, including the kinds of data humans have access to, i.e. multimodal data with speech and visual signals, and the inherit length limitations of human memory that current LLMs lack. Again, future (and some current work) could possibly address these discrepancies.[9]

In this sense, LLMs might be said to be decent models of language learning under idealised and specific conditions, modally separate but not necessarily incommensurate with those we know to obtain in the human case. They map the learning space, and teach us general modal facts and structure about language indirectly. This allows us a comparison point and room for counterfactual modelling.

5.3 Language models and modal structure

Establishing the claim that LLMs model human language might be thought to require some account of what a human language is. However, (8) suggests otherwise. Models can be shown

[7] If this seems far-fetched consider some recent work on the emerging metalinguistic capabilities of LLMs (Begus et al., 2025).

[8] Such as the BabyLM challenge mentioned above. Similarly, mechanistic interpretability via causal abstraction is a promising tool for understanding the internal structures of LLMs (Geiger et al., 2024). These are just two examples of an ever-growing literature.

[9] An interesting recent study by Vong et al. (2024) involved fitting a child with a head-mounted video recorder during object identification tasks. Then an LLM was trained on the video images paired with uttered words in natural environments.

to capture modal structure by the manipulating and intervening on structures (8) and their predictive successes can implicate an explanatory path from said modal structure. For a concrete example of these points at work, consider the case study of thermometry presented in Chang (2004). Chang argues that the precise measurement of temperature predated a successful theory of temperature by decades (and even involved later debunked theories). Despite this, thermometry produced precise, systemic models of the target system. As Chang (2004, p. 160) notes "practical thermometry achieved a good deal of reliability and precision before people could say with any confidence what it was that thermometers measured". Our solar system example is another case of measurement and predictive success in lieu of theoretical certainty or accuracy.[10] Here we make the argument that LLMs might be measuring or modelling language in a more abstract modal sense of the concept. Before we motivate this picture, we need to address the still dominant competence model of language which has resisted LLMs as well their earlier ancestors in computational linguistics (Chomsky, 1957).

5.3.1 Competence as a barrier

There have been a number of recent objections to the relevance of LLMs to linguistic theory and cognitive science in general. In an influential article, Bender & Koller (2020) mount a strong argument to the effect that LLMs trained on linguistic form (i.e. text) are incapable of accessing meaning *simpliciter*. Their view of meaning requires some concept of communicative intent (paired with linguistic form), which they argue is absent in LLMs. For them the main issue is that LLMs lack an appropriate connection to the outside world. Others propose different ways LLMs have linguistic access the world, either by indirect reference or bypassing reference entirely (Mandelkern & Linzen, 2024; Piantadosi & Hill, 2022).

Another line that has been pursued starts with the claim that LLMs are general learners capable of learning non-human, so-called 'impossible languages', as well as human ones. Moro et al. (2023) argue that the difference between these two classes of languages is essential to our linguistic cognition, and that LLMs fail to distinguish them. Kallini et al. (2024) dispute the latter claim by carefully evaluating the performance of a GPT-2 small model on a continuum of perturbations of English strings. They conclude that the LLM struggles with non-human languages and shows preference for human linguistic rules. This is an instance of a more general point which is that some theoretical linguists, especially those within the generative tradition, hold that understanding human language involves studying linguistic competence *not* performance. There are a few ways in which this argument can play out. We will consider Dupre (2021)'s assessment of the issues as they offer clear philosophical arguments assumed but not often explicated by others in the literature.

The foil to Dupre's argument is the thought that linguistic performance is 'competence plus noise'. The thought is tempting, especially from an information-theoretic perspective. Con-

[10] Similarly, Chirimuuta (2024) argues that the brain is Heraclitean in that it is in constant flux and not a stable object of scientific inquiry, yet it remains a reasonable target of neuroscientific models.

sider the coding model of communication of Shannon & Weaver (1949), in which a message emanates from an information source which is converted by a transmitter into a signal. This signal is then sent to a receiver through a potentially noisy channel (Shannon proved important theorems about this process). The receiver then converts the received signal back into a message and sends it to a destination. In machine learning Widrow & Hoff (1960)'s least mean squares algorithm paved much of the way to modern artificial neural networks, pioneering error minimisation techniques and improved learning algorithms. It was initially based on work on adaptive filters in noisy communication channels.

This picture might invite the idea that linguistic competence or cognition (or whatever internal causal mechanisms responsible for language in humans) might be like the message transmitted over a noisy channel. The resulting performance system includes noise from various places (perhaps including general cognition) besides this pure information source, or generative grammar, but competence is ultimately recoverable from performance. Dupre disputes this possibility. For him, as well as for most generative linguists, competence and performance are different in kind not degree (of noise). Specifically, he insists that "mainstream generative linguistic theory assumes that linguistic expressions are not, strictly speaking, sounds, or publicly observable signs, at all" but instead they are "internal, psychological structures" (Dupre, 2021, p. 625). This is Chomsky's influential notion of an I-language or internal representation of the state of the language faculty (Chomsky, 1986). If I-language is the target of our theories, performance only indirectly provides access to it. The reason is that performance is affected by a number of linguistically irrelevant factors, such as communicative goals, dysfluencies, and memory limitations. For example, your internal grammar might license an infinite series of adverbial modifications of 'very' sentences or centre embeddings based on the rules of the grammar. But you will never be able to perform such utterances. And indeed, generative linguists have claimed that infinity is a core property of natural language cognition (Chomsky et al., 2023). Hence, corpora containing actual utterances of speech (or writing) can only tell you so much about this internal representation or state, if anything. Dupre lists a number of other examples from generative linguists that allegedly block reverse engineering competence from performance alone. Examples such as echo-questions ('James said what?') and subject dropping ('Met Poppy at the office. Doing well.') might show up in the data, but they don't map onto grammatical rules (or structures of I-language). He speculates that performance might, among other things, add its own signal. On the flip side, theoretical linguistics under this banner involves a number of hidden or unobservable posits unlikely to be deduced from performance (such as movement and invisible copies or traces left after such movement).

Dupre claims that his arguments generalise to other frameworks in linguistics but this is questionable. Dupre characterises theoretical linguistics as "the basic scientific project of describing and explaining the properties of human language" (Dupre, 2021, p. 618). However, he precisifies this as follows:

> For the purposes of this paper, I shall assume that TL [theoretical linguistics] is a

> branch of cognitive psychology, and thus that a true theory of human language will thereby provide an account of the distinctive features of human psychology that enable us to learn and use language. (Dupre, 2021, p. 618)

This is of course a much stronger characterisation of theoretical linguistics, and many linguists reject the competence-performance distinction, the modular circumscription of I-language or 'distinctive features of human psychology', and even the idea that deep universals underlie surface structures. Some prominent examples are variants of cognitive grammar such as construction grammar, model-theoretic alternatives like head-driven phrase structure grammar and more semantically infused options like dynamic syntax (see Pullum (2013) and Nefdt (2024) for some comparisons). And indeed, many of these linguists have shown more receptiveness towards the incorporation of deep learning and LLMs into their research, such as Goldberg (2024) who suggests that LLMs offer a vindication of constructional approaches.

Even if Dupre's characterisation of linguistics is correct, it cannot be that nothing of competence is extractable from performance – that there are no relevant patterns in corpora that bear theoretical questions about language – since that would not only relegate LLMs to irrelevance but most of computational linguistics itself. A more plausible interpretation of Dupre's position is that the method of extraction, e.g. statistical next-word prediction over massive amounts of text, is unlikely to reveal the mechanisms behind human linguistic performance even if it can identify certain patterns in the data. We do not dispute this claim as the next section elaborates.[11] However, even if LLMs arrive at their output via different means, they can still tell us something about the 'properties of human language', in particular about aspects of its modal structure (8) without necessarily representing the causal mechanisms behind it (3) (just as a model of the solar system can represent the modal structure of the orbits of the planets without representing the causal mechanisms behind it). On the other hand, if the issue is supposed to be with the methodology of continuous versus discrete representation (the kind that has been traditionally favoured by linguistic theory), then recall from section 3 that discrete models can model continuous systems and vice versa. Fluid dynamics provides countless examples of phenomena and processes that can be represented both continuously and discretely (Piantadosi (2024) takes the latter to be a subset of the former). An additional tantalising thought is that by distinguishing model performance from its architecture, LLMs might incorporate their own competence-performance distinction.[12]

5.3.2 Evidence for Language Modelling

There is a longstanding debate in the foundations of linguistics about the exact target of theory and the ontological underpinnings of the subject (see Scholz et al. (2024)). Oversimplifying,

[11] Although it is an open question as to the extent or possibility of convergence between the structures of LLMs and language processing in humans, see Caucheteux & King (2022) and Millet et al. (2022).

[12] See Mahowald et al. (2024) for a neuroscientifically-inspired distinction between formal and functional competence in LLMs.

traditionally Chomskyans favour a mentalistic approach in which language is identified with an internal system of forms, Platonists argue that language is a mind-independent abstract object (or languages are), and nominalists motivate a more modest approach involving conventions and public languages (see Santana (2016)).[13] As already discussed, models can operate in the absence of theoretical consensus or even clarity. Thus, our final argument will present promising evidence for the position that LLMs do capture some modal structure of language. But rather than say what language is, we discuss three features commonly assumed by cognitive scientists and linguists alike. If LLMs can accurately predict the resulting structures while or by allowing for counterfactual manipulation and intervention we should entertain the possibility that they are modelling language. For coverage, we will consider three features which present at different levels of linguistic cognition within syntax, semantics, and pragmatics respectively. The three features themselves are hierarchy, compositionality, and communicative implicature.[14]

These features cannot be understood as laws of language use in the sense of true universal generalisations about all actual usage, because they are often violated. For example, idioms are considered noncompositional, yet the principle of compositionality is still thought to apply to semantic cognition in general. In this respect, they are like many laws in science. For example, as Cartwright (1983) argues, Newton's first law is not true of any actual system because all bodies are subject to external forces to some extent, and most if not all laws in science are approximate and break down in some circumstances. Nonetheless, they can capture modal information or structure in different ways (5), while being incomplete and partial (6). Models do not have to be isomorphic to the systems they represent ((4) (see Section 3). We are in a similar epistemic situation with the cognitive science of language. The three properties above capture more than minimal modal structure and less than the isomorphism sometimes required by Chomskyans. There is some evidence that LLMs are sensitive to each of them.

Morpho-syntactic hierarchical structure dependence refers to the particular composition of clauses and/or phrases within sentences. This is a formal characteristic of expressions which ensures that they embed discrete units in certain ways. For example, a Noun Phrase (NP) in English consists of a noun and a determiner (even a null one). A Verb Phrase (VP) can consist of a verb and an NP. In general, this structure promotes a tree-like representation for sentences as follows.

The hierarchical nature of syntax has been argued to be essential to human language, both structurally and in terms of acquisition. It is the closest thing to a universal in linguistics. For example, standard accounts of the poverty of stimulus rely on the fact that young children don't make the kinds of mistakes one would expect if they were learning linear rules of composition

[13] In all likelihood, natural language is closer to a complex system than a stable object or target (Ladyman & Wiesner, 2020; Nefdt, 2023).

[14] You could choose a different set of properties and run the same argument as we do below. But these properties have received the most attention in the literature and commanded something approaching consensus in terms of their significance, at least.

(Marcus, 1993; Pearl, 2021). In fact, early critiques of stochastic models of language, so called n-gram or hidden Markov models, emphasised their inability to capture the hierarchical nature of language (Chomsky, 1956).

Recent studies have shown that LLMs can indeed be trained to be sensitive to hierarchical structure. Structure sensitivity manifests in different ways. A simple case is syntactic agreement, in which there is a matching or correspondence between parts of sentences in terms of case, gender, number, and person. Subjects agree with verbs in this way in English. In order to test whether structure and not just linear order is being appreciated, researchers introduce attractors or intervening materials between subjects and verbs in the data. Linzen et al. (2016) trained a network to predict the number of the verb on various examples, extracted from corpora in a supervised setting. In the testing phase, the network made correct predictions 99% of the time. Even in the presence of up to four attractors, the LM still achieved 82% accuracy. In fact, the reduction of accuracy tracks with human performance on similar tasks.

However, as Dupre points out above, syntactic theory (especially generative grammar) often posits hidden structural elements not present in the surface forms of sentences. Filler-gap dependencies follow this pattern. A 'filler' is a displaced word in a sentence that leaves behind a 'gap' or empty position when it moves. In the wh-question, 'who' is the filler and the place after 'met' is the gap.

(1) *Who* did Linda say Vinesh met _?

Similar analyses have been offered for relativization, clefting, and fronting (see Kroeger (2004)). Again, Wilcox et al. (2018) use a *surprisal* technique to probe whether RNNs are sensitive to violations of filler-gap constructions. Their hypothesis is that if the model is indeed sensitive to this kind of syntactic structure, it would be 'surprised' by words ungrammatically occupying gap positions by assigning a lower probability to those words. Their results indicate that this is indeed the case. Manning et al. (2020) further investigate whether unsupervised (or self-supervised) transformer models such as BERT are similarly sensitive to syntactic dependencies in language. Despite behavioural suggestions that they are (BERT outperforms humans on many agreement tasks), they delve into the internal mechanisms by means of attention and structural probes for clarity. They conclude that:

> No single attention head corresponds well to dependency syntax overall; the best head gets 34.5% accuracy, which is not much better than the right-branching baseline (26.3% accuracy). However, we find that certain attention heads specialize to specific dependency relations, sometimes achieving high accuracy and substantially outperforming the fixed-offset baseline...Explicitly incorporating syntactic information to improve attention has been an active area of NLP research ... Our results suggest that self-supervised training can cause learned syntax-aware attention to arise in NLP models. (Manning et al., 2020, p. 30050)

Thus, not only do they suggest that LLMs can process syntax in a manner sensitive to language's hierarchical structure but also that such structure is possibly emergent in LLMs within a self-supervised context.[15]

The formal link between syntactic form and semantic meaning has traditionally been modelled by the 'principle of compositionality' (Janssen, 1997; Nefdt & Potts, 2024; Partee, 1981; Szabó, 2000). This principle states that there is a functional relationship between the meaning of a complex expression (like a sentence) and the meanings of its constituent parts and their syntactic combination. As Millière (2024) notes, compositionality is not just a methodological tool or property of expressions but usually also assumed to be an essential part of linguistic competence. In many cases, other important properties such as productivity, creativity, and systematicity are explained by means of compositionality.[16] In fact, compositionality has been at the centre of the debates on language and AI for decades (Fodor & Pylyshyn, 1988).

Given that LLMs are often trained on massive datasets, they contain millions (sometimes billions) of parameters, evaluating whether they use compositional generalisation or strategies requires some innovative modelling. In one such study, Lake & Baroni (2018) create a synthetic or artificial dataset (SCAN) comprising commands mapped on to simple actions (jump twice $\Rightarrow$ JUMP JUMP) to assess whether LLMs will use compositional tools when required. The problem with natural datasets is that the LLMs could use a number of heuristics to mimic compositional meaning without exploiting actual compositional procedures. "By training models on synthetic data, the aim is to evaluate whether they can productively combine known units based on a representation of their underlying structure, rather than relying solely on memorized patterns" (Millière, 2024, p. 14). The key is finding ways to distinguish interpolation from extrapolation, and then determining whether this extrapolation (if it occurs) is rule-bound. Early results were mixed in terms of the performance on the test sets while transformer LLMs have significantly improved on compositional generalisation benchmarks.[17]

In terms of counterfactual intervention, Lepori et al. (n.d.) explore model pruning techniques and ablation to provide evidence that LLMs often implement subroutines in terms of modular subnetworks. This suggests a form of what they call 'structural compositionality' or the extent to which models decompose tasks into subroutines and then use these units modularly in their inner workings.

Lastly, and perhaps somewhat surprisingly, is the issue of pragmatics. Unlike syntax and semantics, which can be modelled with relatively static mathematics, pragmatics is a dynamic

[15] It should be noted that other researchers dispute these kinds of results. See Murphy et al. (2025) for a very recent such study with the o3 model from OpenAI.

[16] *Productivity* generally refers to our ability to produce an indefinite amount of expressions from a finite vocabulary and rule base, *creativity* refers to our ability to create and understand novel expressions, and *systematicity* to the manipulation of rules and a finite vocabulary to yield distinct but related meanings (from *Susan taught Steven* to *Steven taught Susan*).

[17] The especially tricky issue with compositionality is that clear consensus on its definition is lacking in both the formal semantic and the NLP domains. See Hupkes et al. (2020) for a review and some promising possibilities as well as McCurdy et al. (2024).

enterprise which involves active participants. LLMs are not exposed to the real world outside of text masked as numbers. But they do interact with us on a daily basis in Chatbot form. Pragmatics studies the inferential logic behind natural language use. For instance, syntax and semantic (via compositionality) can be used to identify the literal meaning of expressions based on the meanings of their parts. But when we use natural language in conversational settings, we often deviate in systematic ways from literal meaning. Sarcasm, subterfuge, and metaphor all work in this kind of way. Grice (1975, 1989) identified the concept of implicature in which a meaning, distinct from the literal, is implied by a communicative act in a particular context. Specifically, Grice put forward the 'cooperative principle' which enjoins participants of a conversation to maximize their contributions to communication along the lines of four maxims (manner, quantity, quality, and relation). Violation or 'flouting' of maxims generates implicature, and calls the interlocutor to reinterpret the message. There have been many challenges and amendments to this basic picture of pragmatics (see Horn (2004) and Sperber & Wilson (1995)). Nevertheless, it is quite remarkable that systems trained only on text, with a next word objective, can seem to appreciate the subtleties of human communication in the ways that LLMs seem to do across platforms. Piantadosi & Hill (2022) even account for the possibility of LM's semantic competence in terms of their appreciation of inferential or conceptual roles, in the absence of reference to external objects (in response to Bender & Koller (2020)).

Some theorists have posited that the cognitive mechanism which underlies pragmatic phenomena (in humans) is present in LLMs. Key to the pragmatics of communication is the monitoring of other individuals' mental states or beliefs during interaction (Lewis, 1979; Stalnaker, 2002). This process is iterative (*I know that you know that I know that...*) and seems to require some sort of metacognition. Communication can be understood to involve playing games in the sense of game theory (Parikh, 1992). In cognitive science, this process is referred to as 'theory of mind' or ToM and exemplified by false belief experiments in which participants are asked to assume information states of others, distinct from their own. It forms an important area of research into social cognition and even cognitive development and language.

There has been a flurry of research in NLP and cognitive science to test whether LLMs have something akin to ToM. Kosinski (2024), and many others, shows that LLMs can solve false belief tasks. Of course, given the enormity of their training, which includes academic papers, they might have encountered the tests before on the internet and thus be parroting the patterns found there. Strachan et al. (2024) conduct a comparative study of different LM families, with different training regimes and architectures, and their respective performance on ToM tests. GPT models performed well on most tests, excluding faux pas[18] where the poor performance was attributed to a hyperparameter setting rather than a failure of inference. van Duijn et al. (2023) expand the set of models and tests (beyond false belief) as well. They confirm the superiority

[18]"The faux pas test consists of vignettes describing an interaction where one character (the speaker) says something they should not have said, not knowing or not realizing that they should not say it." (Strachan et al., 2024, p. 1288). This ability allegedly requires tracking two mental states (the speaker's ignorance and the hearer's potential offence).

of GPT models over base LLMs (and even children aged 7-10). They suggest that language development (long linked to ToM) and the GPT family's reward mechanism, i.e. fine-tuning, might promote cooperation and thus explain higher level performance on ToM. Importantly, all of the above studies caution against causal conclusions as to the possible 'mindedness' of LLMs.

Modal information need not be causal information, as stressed in section 3. LLMs, like scientific models in general, seem sensitive to the former. We see no reason why they cannot also be designed to probe the latter (although we agree with critics that off-the-shelf engineering LLMs are unlikely to already be fit for this task). Recently, Rothschild (forthcoming) posits that LLMs are special as compared to other deep neural networks because they are imbued with language, which is itself a compressed representation that makes general reasoning computationally tractable. He draws on Dennett (1991) and the notion of real patterns here. This possibility links to some views in the NLP community as to whether LLMs create 'world models' via language (Ha & Schmidhuber, 2018; Sutton & Pinette, 1985). For example, for an LM to predict the next word in a chess conversation between chess masters, the LM would apparently need to construct some representation of the rules of the game. A recent study suggested that a GPT model trained to predict legal moves in the game *Othello* (Othello-GPT) produced an 'emergent, nonlinear internal representation of the board state' without prior knowledge of the game or its rules (Li et al., 2023). However, the 'Othello World Model hypothesis' has been confronted with the more parsimonious 'Othello bag of heuristics' one (jylin04 et al., 2024). Whether or not LLMs can access modal information about more than language, i.e. reasoning or world knowledge etc., is a fascinating question but one beyond the present scope. Either way, Rothchild's argument, the world models hypothesis and even the performance on ToM tests, might suggest that the modal structure of natural language already contains modal information about other behaviour like reasoning. We reserve judgement about this possible extension. In any case, the ability of LLMs to track features within syntax, semantics, and pragmatics highlights their status as scientific models as they deliver information about the possible ways language could be, and likely is, despite probably arriving there by distinct means and mechanisms. Additionally, this suggests that they could be good models of the modal structure of language. The idea of modal structure is connected to Dennett's idea of 'real patterns' by Ladyman and Ross to understand scientific ontology. In a recent article, Futrell & Mahowald (2025, p. 33) argue that LLMs can offer a lot to linguistic theory without replacing its import. Specifically, they argue that both linguistic theory and LLMs track 'real patterns' in language. They state that "LLMs are proof-of-concept that systems can process language without having linguistic structure hardwired in. But that doesn't mean it isn't real".

6. Conclusion

We suspect that much of the tension in the current literature emanates from a different source of scientific controversy concerning the relationship between engineering and science (see Rapaport (2023) for a detailed discussion). Those who tend to dismiss the scientific credentials of language models also tend to emphasise their current engineering goals, while some of those who defend their scientific merit highlight their neurobiological origins as engineering solutions due to natural selection. This issue is to a large extant imported into AI and work on language models from computer science, and the debate is unresolved in its parent context. We have shown that recent work on scientific modelling leaves plenty of room for LLMs to be viewed as scientific models in linguistic theory.

Nothing we have said entails that LLMs are accurate models of human language, cognition or learning simpliciter. However, LLMs viewed as scientific models can provide indirect representations, counterfactual interventions, and scientific insights about the modal structure of language. In fact, they seem to be scientific tools of this ilk, neither mere corpus models nor full models of human cognition. As models of language perhaps they lie somewhere between Ptolemy's and Kepler's models of the solar system, telling us something about the kinematics of language use, without yet revealing the full *gravity* of human cognition.

Acknowledgments

Many thanks to Emanuele Ratti and Karim Thebault for comments and discussion of scientific modelling.

References

Ayub, H. (2023). GPT-4o: Successor of GPT-4? [Published on Medium]. https://medium.com/@hamid_ayub/gpt-4o-successor-of-gpt-4-123456789

Baggio, G., & Murphy, E. (2024). On the referential capacity of language models: An internalist rejoinder to Mandelkern & Linzen. https://arxiv.org/abs/2406.00159

Baron, S., Bihan, B. L., & Read, J. (2025). Scientific theory and possibility. *Erkenntnis*, 1, 1–17. https://doi.org/10.1007/s10670-025-00939-3

Baroni, M. (2022). On the proper role of linguistically oriented deep net analysis in linguistic theorising. In *Algebraic structures in natural language*. CRC Press.

Begus, G., Dąbkowski, M., & Rhodes, R. (2025). Large linguistic models: Investigating LLMs' metalinguistic abilities. *IEEE Transactions on Artificial Intelligence*, 1–15. https://doi.org/10.1109/TAI.2025.3575745

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463

Berenstain, N., & Ladyman, J. (2012). Ontic structural realism and modality. In E. Landry & D. Rickles (Eds.), *Structural realism: Structure, object, and causality*. Springer.

Cartwright, N. (1983). *How the laws of physics lie*. Oxford University Press.

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5, 134. https://doi.org/10.1038/s42003-022-03036-1

Chang, H. (2004). *Inventing temperature: Measurement and scientific progress*. Oxford University Press.

Chirimuuta, M. (2024). *The brain abstracted: Simplification in the history and philosophy of neuroscience*. The MIT Press. https://doi.org/10.7551/mitpress/13804.001.0001

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory*, 2, 113–124.

Chomsky, N. (1957). *Syntactic structures*. Mouton.

Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. Praeger.

Chomsky, N., Seely, T. D., Berwick, R. C., Fong, S., Huybregts, M. A. C., Kitahara, H., McInnerney, A., & Sugimoto, Y. (2023). *Merge and the strong minimalist thesis*. Cambridge University Press. https://doi.org/10.1017/9781009343244

Cichy, R. M., & Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in Cognitive Sciences*, 23(4), 305–317. https://doi.org/10.1016/j.tics.2019.01.009

Dennett, D. C. (1991). Real patterns. *The Journal of Philosophy*, 88(1), 27–51. https://doi.org/10.2307/2027085

Dohrn, D. (2023). Modals model models: Scientific modeling and counterfactual reasoning. *Synthese*, 201(5), 1–22. https://doi.org/10.1007/s11229-023-04135-0

Dupre, G. (2021). (what) can deep learning contribute to theoretical linguistics? *Minds and Machines*, 31(4), 617–635. https://doi.org/10.1007/s11023-021-09571-w

Dupre, G. (2024). Acquiring a language vs. inducing a grammar. *Cognition*, 247(100), 105771. https://doi.org/10.1016/j.cognition.2024.105771

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. In *Touretzky 1991: Advances in neural information processing systems 2* (pp. 91–122). Morgan Kaufmann. https://doi.org/10.1007/978-1-4615-4008-3_5

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1), 3–71. https://doi.org/10.1016/0010-0277(88)90031-5

French, S. (2014). *The structure of the world: Metaphysics and representation*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199684847.001.0001

Frigg, R. (2010). Models and fiction. *Synthese*, 172(2), 251–268. https://doi.org/10.1007/s11229-009-9505-0

Frigg, R., & Nguyen, J. (2020). *Modelling nature. an opinionated introduction to scientific representation*. Springer.

Futrell, R., & Mahowald, K. (2025). How linguistics learned to stop worrying and love the language models [Published online 24 July 2025]. *Behavioral and Brain Sciences*, 1–98. https://doi.org/10.1017/S0140525X2510112X

Geiger, A., Ibeling, D., Zur, A., Chaudhary, M., Chauhan, S., Huang, J., Arora, A., Wu, Z., Goodman, N., Potts, C., & Icard, T. (2024). Causal abstraction: A theoretical foundation for mechanistic interpretability. https://arxiv.org/abs/2301.04709

Godfrey-Smith, P. (2006). The strategy of model-based science. *Biology and Philosophy*, 21(5), 725–740. https://doi.org/10.1007/s10539-006-9054-6

Gold, E. M. (1967). Language identification in the limit. *Information and Control*, 10(5), 447–474.

Goldberg, A. E. (2024). Usage-based constructionist approaches and large language models. *Constructions and Frames*, 16(2), 220–254. https://doi.org/https://doi.org/10.1075/cf.23017.gol

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics* (pp. 41–58, Vol. 3). Academic Press.

Grice, H. P. (1989). *Studies in the way of words*. Harvard University Press.

Ha, D., & Schmidhuber, J. (2018). World models. https://doi.org/10.5281/ZENODO.1207631

Hesse, M. B. (1963). *Models and analogies in science*. University of Notre Dame Press.

Horn, L. R. (2004). Implicature. In L. R. Horn & G. Ward (Eds.), *Handbook of pragmatics* (pp. 3–28). Blackwell.

Hupkes, D., Dankers, V., Mul, M., & Bruni, E. (2020). Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67, 757–795. https://doi.org/10.1613/jair.1.11675

Janssen, T. (1997). Compositionality. In J. van Benthem & A. ter Meulen (Eds.), *Handbook of logic and language* (pp. 417–473). Elsevier Science.

Johnson, K. (2004). Gold's theorem and cognitive science. *Philosophy of Science*, 71(4), 571–592.

jylin04, JackS, Karvonen, A., & Can. (2024, July). OthelloGPT learned a bag of heuristics [Accessed: 2025-02-19].

Kallini, J., Papadimitriou, I., Futrell, R., Mahowald, K., & Potts, C. (2024). Mission: Impossible language models. https://arxiv.org/abs/2401.06416

Katzir, R. (2023). Why large language models are poor theories of human linguistic cognition: A reply to Piantadosi. *Biolinguistics*, 17. https://doi.org/10.5964/bioling.13153

Knuuttila, T. (2011). Modelling and representing: An artefactual approach to model-based representation. *Studies in History and Philosophy of Science Part A*, 42(2), 262–271. https://doi.org/10.1016/j.shpsa.2010.11.034

Knuuttila, T. (2020). Models, fictions, artifacts. In W. J. Gonzalez (Ed.), *Language and scientific research* (pp. 143–162). Palgrave Macmillan. https://doi.org/10.1007/978-3-030-60537-7_7

Knuuttila, T. (2021). Imagination extended and embedded: Artifactual versus fictional accounts of models. *Synthese*, 198(Suppl 21), 5077–5097. https://doi.org/10.1007/s11229-019-02446-2

Knuuttila, T., & Merz, M. (2009). Understanding by modeling: An objectual approach. In H. W. de Regt, S. Leonelli & K. Eigner (Eds.), *Scientific understanding: Philosophical perspectives* (pp. 146–168). University of Pittsburgh Press.

Knuuttila, T., & Voutilainen, A. (2003). A parser as an epistemic artifact: A material view on models. *Philosophy of Science*, 70(5), 1484–1495. https://doi.org/10.1086/377424

Kosinski, M. (2024). Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45), e2405460121. https://doi.org/10.1073/pnas.2405460121

Kroeger, P. R. (2004). Filler–gap dependencies and relativization. In *Analyzing syntax: A lexical-functional approach* (pp. 165–191). Cambridge University Press.

Ladyman, J., & Lorenzetti, L. (forthcoming). Effective ontic structural realism. *British Journal for the Philosophy of Science*. https://doi.org/10.1086/729061

Ladyman, J., & Ross, D. (2007). *Everything must go: Metaphysics naturalized*. Oxford University Press.

Ladyman, J., & Wiesner, K. (2020). *What is a complex system?* Yale University Press.

Lake, B., & Baroni, M. (2018). Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. *Proceedings of the 35th International Conference on Machine Learning*, 80, 2873–2882. https://proceedings.mlr.press/v80/lake18a.html

Lappin, S., & Shieber, S. M. (2007). Machine learning theory and practice as a source of insight into universal grammar. *Journal of Linguistics*, 43(2), 393–427.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics*, 20(1), 1–31.

Lepori, M. A., Serre, T., & Pavlick, E. (n.d.). Evidence for structural compositionality in neural networks. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt & S. Levine (Eds.), *Advances in neural information processing systems 36: Annual conference on neural information processing systems 2023 (neurips 2023)* (pp. –).

Lewis, D. (1979). Scorekeeping in a language game. In *Philosophical papers, volume i* (pp. 233–249). Oxford University Press.

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., & Wattenberg, M. (2023). Emergent world representations: Exploring a sequence model trained on a synthetic task. *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=DeG07_TcZvT

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4, 521–535. https://doi.org/10.1162/tacl_a_00115

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6), 517–540. https://doi.org/10.1016/j.tics.2024.01.011

Mandelkern, M., & Linzen, T. (2024). Do language models' words refer? *Computational Linguistics*, 50(3), 1191–1200. https://doi.org/10.1162/coli_a_00464

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48), 30046–30054. https://doi.org/10.1073/pnas.1907367117

Marcus, G. F. (1993). Negative evidence in language acquisition. *Cognition*, 46, 53–85.

Massimi, M. (2022). *Perspectival realism*. Oxford University Press.

McClelland, J. L., Rumelhart, D. E., & the PDP Research Group. (1986). *Parallel distributed processing, volume ii: Explorations in the microstructure of cognition: Psychological and biological models* (Vol. 2). MIT Press.

McCoy, R. T., Yao, S., Friedman, D., Hardy, M. D., & Griffiths, T. L. (2024a). Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41), e2322420121. https://doi.org/10.1073/pnas.2322420121

McCoy, R. T., Yao, S., Friedman, D., Hardy, M. D., & Griffiths, T. L. (2024b). When a language model is optimized for reasoning, does it still show embers of autoregression? an analysis of OpenAI o1. https://arxiv.org/abs/2410.01792

McCurdy, K., Soulos, P., Smolensky, P., Fernandez, R., & Gao, J. (2024). Toward compositional behavior in neural models: A survey of current views. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 9323–9339. https://doi.org/10.18653/v1/2024.emnlp-main.524

Meskhidze, H. (2023). Can machine learning provide understanding? how cosmologists use machine learning to understand observations of the universe. *Erkenntnis*, 88(5), 1895–1909. https://doi.org/10.1007/s10670-021-00434-5

Millet, J., Caucheteux, C., Orhan, A., Boubenec, Y., Gramfort, A., Dunbar, E., Pallier, C., & King, J.-R. (2022). Toward a realistic model of speech processing in the brain with self-supervised learning. *Advances in Neural Information Processing Systems (NeurIPS)*. https://doi.org/10.48550/arXiv.2204.00433

Millhouse, T. (2022). Really real patterns. *Australasian Journal of Philosophy*, 100(4), 664–678. https://doi.org/10.1080/00048402.2021.1941153

Millière, R. (2024). Language models as models of language. https://arxiv.org/abs/2408.07144

Millière, R., & Buckner, C. (2024). A philosophical introduction to language models – part I: Continuity with classic debates. https://arxiv.org/abs/2401.03910

Morgan, M. S., & Morrison, M. (1999). *Models as mediators: Perspectives on natural and social science*. Cambridge University Press.

Moro, A., Greco, M., & Cappa, S. F. (2023). Large languages, impossible languages and human brains. *Cortex*, 167, 82–85. https://doi.org/https://doi.org/10.1016/j.cortex.2023.07.003

Müller, S. (2018). *Grammatical theory: From transformational grammar to constraint-based approaches*. Language Science Press.

Murphy, E., Leivada, E., Dentella, V., Gunther, F., & Marcus, G. (2025). Fundamental principles of linguistic structure are not represented by o3. https://arxiv.org/abs/2502.10934

Nefdt, R. M. (2023). *Language, science, and structure: A journey into the philosophy of linguistics*. Oxford University Press.

Nefdt, R. M. (2024). *The philosophy of theoretical linguistics*. Cambridge University Press.

Nefdt, R. M., & Potts, C. (2024). Compositionality. In M. C. Frank & A. Majid (Eds.), *Open Encyclopedia of Cognitive Science*. MIT Press. https://doi.org/10.21428/e2759450.494deacd

Nguyen, J. (2020). It's not a game: Accurate representation with toy models. *The British Journal for the Philosophy of Science*, 71(3), 1013–1041. https://doi.org/10.1093/bjps/axy078

Parikh, P. (1992). A game-theoretic account of implicature. *Proceedings of the 4th Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, 85–94.

Partee, B. H. (1981). Montague grammar, mental representations, and reality. In S. Kanger & S. Öhman (Eds.), *Philosophy and grammar: Papers on the occasion of the quincentennial of uppsala university* (pp. 59–78). Springer Netherlands. https://doi.org/10.1007/978-94-009-9012-8_5

Pearl, L. (2021). Poverty of the stimulus without tears. *Language Learning and Development*, 18(4), 415–454. https://doi.org/10.1080/15475441.2021.1981908

Piantadosi, S. T. (2024). Modern language models refute Chomsky's approach to language. In E. Gibson & M. Poliak (Eds.), *From fieldwork to linguistic theory: A tribute to Dan Everett* (pp. 353–414, Vol. 15). Language Science Press. https://doi.org/10.5281/zenodo.11351540

Piantadosi, S. T., & Hill, F. (2022). Meaning without reference in large language models. *arXiv preprint arXiv:2205.12620*. https://arxiv.org/abs/2205.12620

Pincock, C. (2007). A role for mathematics in the physical sciences. *Nous*, 41(2), 253–275. https://doi.org/10.1111/j.1468-0068.2007.00643.x

Potts, C. (2024). Characterizing English Preposing in PP constructions. *Journal of Linguistics*, 1–39. https://doi.org/10.1017/S0022226724000227

Pullum, G. K. (2013). The central question in comparative syntactic metatheory. *Mind and Language*, 28(4), 492–521. https://doi.org/10.1111/mila.12029

Rapaport, W. J. (2023). *The philosophy of computer science*. Wiley-Blackwell.

Rothschild, D. (forthcoming). Language and thought: The view from llms. In D. Sosa & E. Lepore (Eds.), *Oxford studies in philosophy of language volume 3*. Oxford University Press.

Santana, C. (2016). What is language? *Ergo: An Open Access Journal of Philosophy*, 3(19), 501–523. https://doi.org/10.3998/ergo.12405314.0003.019

Saunders, S. (1993). To what physics corresponds. In S. French & H. Kamminga (Eds.), *Correspondence, invariance and heuristics: Essays in honour of heinz post* (pp. 295–325). Reidel.

Scholz, B. C., Pelletier, F. J., Pullum, G. K., & Nefdt, R. (2024). Philosophy of linguistics (E. N. Zalta, Ed.) [Substantive revision]. *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/linguistics/

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. University of Illinois Press.

Sjölin Wirling, Y., & Grüne-Yanoff, T. (2023). Introduction to the synthese topical collection 'modal modeling in science: Modal epistemology meets philosophy of science'. *Synthese*, 201(208). https://doi.org/10.1007/s11229-023-04188-1

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition* (2nd). Blackwell.

Stalnaker, R. (2002). Common ground. *Linguistics and Philosophy*, 25(5–6), 701–721. https://doi.org/10.1023/A:1020867916902

Strachan, J. W. A., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxena, K., Rufo, A., Panzeri, S., Manzi, G., Graziano, M. S. A., & Becchio, C. (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8, 1285–1295. https://doi.org/10.1038/s41562-024-01882-z

Sutton, R. S., & Pinette, B. (1985). The learning of world models by connectionist networks. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 7, 0. https://escholarship.org/uc/item/7j2202xq

Szabó, Z. G. (2000). *Problems of compositionality*. Routledge. https://philpapers.org/rec/SZAPOC-3

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Casado, C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tuckute, G., Kanwisher, N., & Fedorenko, E. (2024). Language in brains, minds, and machines. *Annual Review of Neuroscience*, 47, 277–301. https://doi.org/10.1146/annurev-neuro-032122-022937

van Duijn, M., van Dijk, B., Kouwenhoven, T., de Valk, W., Spruit, M., & van der Putten, P. (2023, December). Theory of mind in large language models: Examining performance of 11 state-of-the-art models vs. children aged 7-10 on advanced tests. In J. Jiang, D. Reitter & S. Deng (Eds.), *Proceedings of the 27th conference on computational natural language learning (conll)* (pp. 389–402). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.conll-1.25

van Fraasen, B. (2008). *Scientific representation: Paradoxes of perspective*. Oxford University Press UK.

van Rooij, I. (2022). Psychological models and their distractors. *Nature Reviews Psychology*, 1, 127–128. https://doi.org/10.1038/s44159-022-00031-5

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. https://arxiv.org/abs/1706.03762

Veres, C. (2022). Large language models are not models of natural language: They are corpus models. *IEEE Access*, 10, 61970–61979.

Vong, W. K., Wang, W., Orhan, E., & Lake, B. (2024). Grounded language acquisition through the eyes and ears of a single child. *Science*, 383, 504–511.

Weisberg, M. (2007). Three kinds of idealization. *The Journal of Philosophy*, 104(12), 639–659.

Widrow, B., & Hoff, M. E. (1960). Adaptive switching circuits. *1960 IRE WESCON Convention Record, Part 4*, 96–104.

Wilcox, E., Levy, R., Morita, T., & Futrell, R. (2018). What do rnn language models learn about filler-gap dependencies? *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 211–221. https://doi.org/10.18653/v1/W18-5423