Why Linguistics Will Thrive in the 21st Century: A Reply to Piantadosi (2023)

Jordan Kodner, **Sarah Payne**, **Jeffrey Heinz**
Department of Linguistics, and
Institute for Advanced Computational Science (IACS)
Stony Brook University
{first.last}@stonybrook.edu

1 Introduction

We present a critical assessment of Piantadosi's (2023) claim that "Modern language models refute Chomsky's approach to language," focusing on four main points. First, despite the impressive performance and utility of large language models (LLMs), humans achieve their capacity for language after exposure to several orders of magnitude less data. The fact that young children become competent, fluent speakers of their native languages with relatively little exposure to them is the central mystery of language learning to which Chomsky initially drew attention, and LLMs currently show little promise of solving this mystery. Second, what can the artificial reveal about the natural? Put simply, the implications of LLMs for our understanding of the cognitive structures and mechanisms underlying language and its acquisition are like the implications of airplanes for understanding how birds fly. Third, LLMs cannot constitute scientific theories of language for several reasons, not least of which is that scientific theories must provide interpretable explanations, not just predictions. This leads to our final point: to even determine whether the linguistic and cognitive capabilities of LLMs rival those of humans requires explicating what humans' capacities actually are. In other words, it requires a *separate* theory of language and cognition; generative linguistics provides precisely such a theory. As such, we conclude that generative linguistics as a scientific discipline will remain indispensable throughout the 21st century and beyond.

2 "Unconstrained" Learning from Big Data is Not Human

OpenAI's newest product at the time of writing, GPT-4,[1] performs very well on a wide range of standardized tests (though see Martínez 2023). For example, on the LSAT, OpenAI reports that GPT-4 scored better than roughly 88% of human test takers seeking admission to American law schools, a significant boost from last year's OpenAI product, GPT-3.5, which only outperformed about 40% of human test takers. Both the performance and pace of improvement are remarkable achievements. Nonetheless, the fact remains that young adults who pass the LSAT are doing so without having read the trillions of sentences and structured internet data that these large language models are trained on. The difference in training regimes is stark and highlights the fundamental question: how do humans come to pass the LSAT, and other standardized tests, on comparatively so *little* data?

---

[1] https://web.archive.org/web/20230314174836/https://openai.com/research/gpt-4. GPT stands for "Generative Pre-trained Transformer."

Parallel questions arise in the domain of language acquisition by children, which is characterized by a relative paucity of linguistic experience. Children are exposed to at most about ten million tokens per year (Hart and Risley, 1992; Gilkerson et al., 2017), and most children have vocabularies of under one thousand words around age three, regardless of the language being learned (Bornstein et al., 2004; Fenson et al., 1994). Yet these same children produce sentences that largely obey the grammatical rules of their communities' languages (Berko, 1958; Brown, 1973; Montrul, 2004; Yang, 2006; Phillips, 2010; Slobin, 2022). Thus, a fundamental question of linguistic study is how children become fluent in their native language(s) at a young age from so little data and experience.

The disconnect between the linguistic experience (input) and the linguistic capacity (output) is what gives rise to the *The Poverty of the Stimulus* argument for the hypothesis that many aspects of language learning and representation are innate (Chomsky, 1959, 1980; Nowak et al., 2001; Yang, 2013). Under this hypothesis, children generalize from their limited input in specific ways, navigating a constrained space of possible natural language grammars. Consequently, they do not consider all logically possible generalizations that are consistent with their linguistic experience. Rather, the particular structure of the hypothesis space facilitates the rapid development of their linguistic capabilities. The Poverty of the Stimulus is not tied to a specific theory of language, such as Minimalism or particular variants thereof, but rather follows from the basic problem of making generalizations from experience, as we describe in the next section.

The contrast between the input to the child and the input to LLMs is striking, but Piantadosi is not concerned by this discrepancy. He makes two claims. The first is that LLMs refute so-called nativist theories of language because "modern language models succeed despite the fact that their underlying architecture for learning is relatively unconstrained" Piantadosi (2023, 18). In other words, he argues that broadly "blank slate" approaches to language learning are in fact viable, with LLMs serving as a proof of concept. The second claim is that "our methods for training [LLMs] on very small datasets will inevitably improve" (Piantadosi, 2023, 14). From these two claims, Piantadosi concludes that LLMs show that unconstrained learning from small data is possible. The remainder of this section addresses each argument in turn.

2.1 Feasible Learning Must be Constrained

Piantadosi is not the first researcher to claim that the general architecture of LLMs, or deep artificial neural networks (ANNs) more broadly, "is relatively unconstrained." As Baroni (2022, 5) points out, work situating deep networks "within a broader theoretical context [does so] invariably in terms of nature-or-nurture arguments resting on a view of deep nets as blank slates." For example, Warstadt et al. (2019, 637) take the position that that "if linguistically uninformed neural network models achieve human-level performance on specific phenomena...this would be clear evidence limiting the scope of phenomena for which the [argument of the Poverty of the Stimulus] can hold." Pater (2019, 43) similarly claims that "with the development of the rich theories of learning represented by modern neural networks, the learnability argument for a rich [Universal Grammar] is particularly threatened." We refer the reader to Baroni (2022) for a plethora of further examples of such claims drawn from the literature.

If deep ANNs really were so unconstrained, however, why would machine learning scientists constantly tinker with the layers, the gating mechanisms, the architectures, and the tuning of the hyperparameters? The reality is that these systems are biased in ways that are not well-understood. Paraphrasing a turn of phrase from Rawski and Heinz's (2019) critique of Pater (2019), "Ignorance of bias does not imply absence of bias." Indeed, Kharitonov and Chaabouni (2020) found that when deep ANNs were trained on a small dataset which could have been generated by either a hierarchical or linear function, LSTMs with attention and Transformers apparently inferred a hierarchical function, while LSTMs without attention and CNNs inferred a linear one. Such a result indicates the presence of robust biases in the network architecture and certainly does not support a "blank-slate" view of deep networks. Though Piantadosi cites Baroni (2022) in support

of his claim that LLMs constitute linguistic theories (§4), he fails to note the second half of Baroni's claim: that deep networks "are linguistic theories, *not* blank slates" (pg. 6, emphasis ours). As Baroni (2022, 7) argues, "it is more appropriate, instead, to look at deep nets as... encoding non-trivial structural priors facilitating language acquisition and processing." Such "non-trivial" priors are fundamentally at odds with a view of LLMs as "relatively unconstrained."

But *even with such non-trivial priors*, the success of current LLMs depends at least in part on being trained on inhumanly large amounts of data (e.g., Kaplan et al., 2020). Indeed, results from the field of computational learning theory (CLT) have established that the kind of "relatively unconstrained" learning Piantadosi suggests is not possible given feasible computational resources, in terms of time and data. CLT studies what it means to learn a concept from experience from a formal mathematical perspective. The primary conclusion from this research is that there are fundamental computational laws of learning that cannot be shortcut. Informally, these laws say that there is a trade-off between the *family of concepts* that one wishes to learn, the *kinds of data* one wishes to learn those concepts from, and the *computational resources* (time and space) within which one needs to accomplish that learning. In particular, it is not possible to learn all computable concepts from arbitrary data presentations representative of them, with feasible amounts of computational resources. This central result is encountered again and again as researchers study different definitions of what "learning" means and examine which families of concepts are learnable under such definitions (Gold, 1967; Wolpert and Macready, 1997; Vapnik, 1998; Jain et al., 1999; Niyogi, 2006; De Raedt, 2008; Mohri et al., 2012; Valiant, 2013).

Figure 1 visualizes some of the parameters involved in defining a learning problem. In the figure, the y-axis represents all possible concepts, including uncomputable ones. The x-axis represents all logically possible data presentations: computable ones, uncomputable ones, ones with only positive examples, ones with only negative examples, and ones with both positive and negative examples. A learning problem is defined in part by selecting some subset C of the logically possible concepts and identifying, for each concept c in C, the data presentations D_c that a learner is expected to succeed on. Then a learning algorithm A can be said to learn the concepts C from data presentations of kind D if and only if for all c belonging to C, and for all d belonging to D_c, it is the case that $A(d) \approx c$. Learning problems may also bound the computational resources that A is allowed to use, possibly as a function of d and c. Readers are referred to (Heinz, 2016) for a survey of different formal learning paradigms.

Piantadosi cites Chater and Vitányi (2007) as an example of work that apparently defies these computational laws because these authors present a paradigm and algorithm that learns any computable language. However, there are two important qualifications to this work that Piantadosi fails to mention. First, Chater and Vitányi themselves note that their result only holds because they *reduce the instance space of data presentations*. In particular, the learner of Chater and Vitányi (2007) is only required to succeed on positive data presentations "generated by some monotone computable probability distribution" (p. 138). This result is actually similar to one obtained by Gold (1967, p. 469) who showed that the class of computable languages is learnable when the positive data presentations are limited to ones generated by a particular kind of computable function (Theorem I.7). Chater and Vitányi's result, like Gold's, follows in part because the instance space of the learning problem has been reduced to cases where the data presentations include positive examples generated by computable processes (see Figure 1). In other words, Chater and Vitányi's results exemplify the trade-off mentioned above: if one reduces the instance space of the learning problem, by limiting which *data presentations* learners have to succeed on, one can expand other aspects of the instance space, such as the *family of concepts* to be learned. This observation is not new; it is originally due to Gold (1967).

While Chater and Vitányi's reduction of the instance space is significant, however, it is not enough to yield *feasible* learning of all computable languages. Another foundation of Chater and Vitányi's 2007 theoretical result is that they allow their algorithm to make "uncomputable" (p. 136)
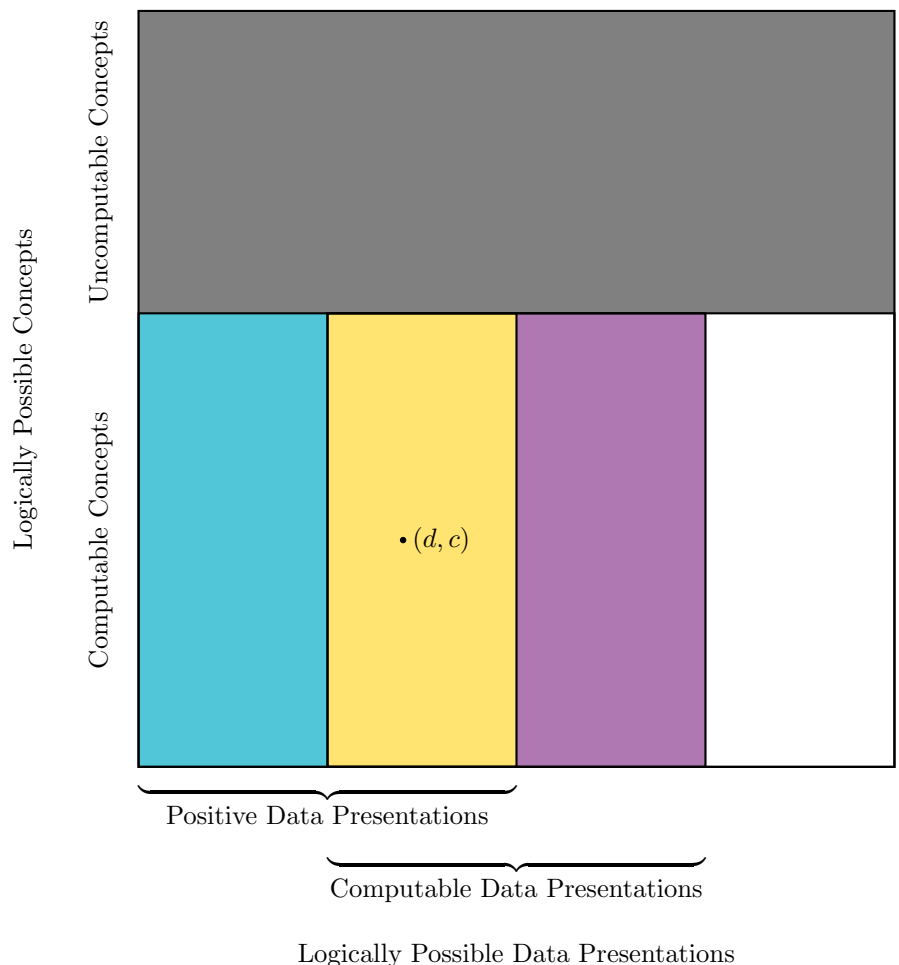
Figure 1: A visualization of how learning problems are defined. Defining a learning problem requires defining which kinds of concepts (y-axis) ought to be obtained from which kinds of data (x-axis). Another important parameter (not shown here) are computational complexity restrictions on the learning mechanism itself. The yellow rectangle exemplifies some choices in formulating a learning problem. Point (d, c) is an instance of the problem: learning concept c from data presentation d.

calculations. Incidentally, this is in contrast to Gold's Theorem I.7, which only considers computable algorithms. While Piantadosi is optimistic about what Chater and Vitányi's "ideal" learner means in practice, Chater and Vitányi are more circumspect. Setting aside the fact that that the heuristics needed to bypass the uncomputable calculations render their theoretical result inert, they acknowledge that "real language learning must occur reliably using limited amounts of data" and therefore "a crucial set of open questions concerns how rapidly learners can converge well enough on the structure of the linguistic environment to succeed reasonably well in prediction, grammaticality judgments and language production" (Chater and Vitányi, 2007, 155). In this regard, it is worth mentioning that, to our knowledge, every learnability result that presents an algorithm which "learns" the class of all computable languages or functions requires *infeasible* amounts of time and data. Feasible learning – that is, learning with limited time and data – requires navigating a *restricted hypothesis space*.

Even the empiricist-minded Clark and Lappin (2011, chapter 7) recognize that constraining the hypothesis space is likely the best way to obtain feasible learning results. They suggest one approach is to "construct algorithms for subsets of existing representation classes, such as context-free grammars" (Clark and Lappin, 2011, 149). In other words, they advocate restricting the class of grammars targeted by learning algorithms to *subclasses* which are feasibly learnable. Such a reduction of the hypothesis space is, at its core, an innate mechanism akin to those advocated for by proponents of the Poverty of the Stimulus argument (e.g. Nowak et al., 2002).

At the end of the day, the results from CLT clearly and firmly support the Poverty of the Stimulus argument. *Even if* LLMs were achieving some sort of "unconstrained" learning as Piantadosi argues, CLT tells us that this would only be possible *because* they were trained on inhumanly large training data, based on the tradeoffs outlined above. Learning from small, feasible amounts of data and computational resources *requires* constraining the hypothesis space, meaning that *even if LMs eventually succeed* at learning from small data, it will be because they encode "non-trivial structural priors facilitating language acquisition and processing" as Baroni (2022, 7) suggests. At some level, Piantadosi must understand this, because he himself suggests (pg. 14) that we might improve training of LMs on small datasets by "build[ing] in certain other kinds of architectural biases and principles" or "consider[ing] learning models that have some of the cognitive limitations of human learners." But what are these biases, principles, and limitations if not some form of the Universal Grammar that Piantadosi (2023, 19) claims LLMs prove "to be wrong"?

2.2 Small Language Models are Anything but Inevitable

Piantadosi (2023, 14) is optimistic that "our methods for training [LLMs] on very small datasets will inevitably improve." This optimism, however, is not well-motivated. Firstly, both model size and training size have seen exponential increases in recent years, and model performance has increased proportionally (Kaplan et al., 2020). Creating smaller models trained on plausible data has never been a central goal of natural language processing (NLP) – the field from which LLMs emerge – because this field seeks primarily to optimize performance for engineering tasks in a world of increasingly available training data and computing power. Secondly, work claiming to successfully train LLMs to achieve human-like performance on human-sized inputs suffers from persistent flaws. Their evaluation methods are weak, often not accurately testing the linguistic phenomena that they purport to or adequately controlling for the presence of side-channel information that might be exploited. Such test sets are particularly susceptible to "shortcutting" by neural models, which have a notorious propensity for exploiting unintentional statistical side-channel information across machine learning domains (Narla et al., 2018; Chao et al., 2018; Sun et al., 2019; Hassani, 2021; Wang et al., 2022). Hence, it is reasonable to be skeptical that models' apparent success on existing evaluation metrics reflects an encoding of the grammatical principle these metrics supposedly test. We elaborate on these rationales below.

Over the last four decades, natural language processing has consistently progressed by consuming more and more data. Just considering the recent history of transformer LMs, while BERT (Devlin et al., 2019) was trained on 3.3 billion tokens, GPT-3 (Brown et al., 2020) was trained on 300 billion, two orders of magnitude increase in only a year, and thousands of times the input available to a human child. This growth in training data has been facilitated by improved computing hardware and a steady increase in model size: in the last few years alone, the definition of LLM has shifted from models with 94 million parameters (ELMo; Peters et al., 2018) to 340 million parameters (BERT-Large; Devlin et al., 2019), 11 billion (T5; Raffel et al., 2020), and 175 billion (GPT-3; Brown et al., 2020). While this pattern has often been dubbed the "Moore's Law of NLP," it actually far *exceeds* the exponential rate of growth predicted for transistors by the original Moore's Law (Figure 2; Liang 2023). Though the exact number of parameters or training data size of GPT-4 are not publicly available, these trends suggest that GPT-4 likely has at least two orders of magnitude more parameters than GPT-3, and is likely trained on as many orders of magnitude
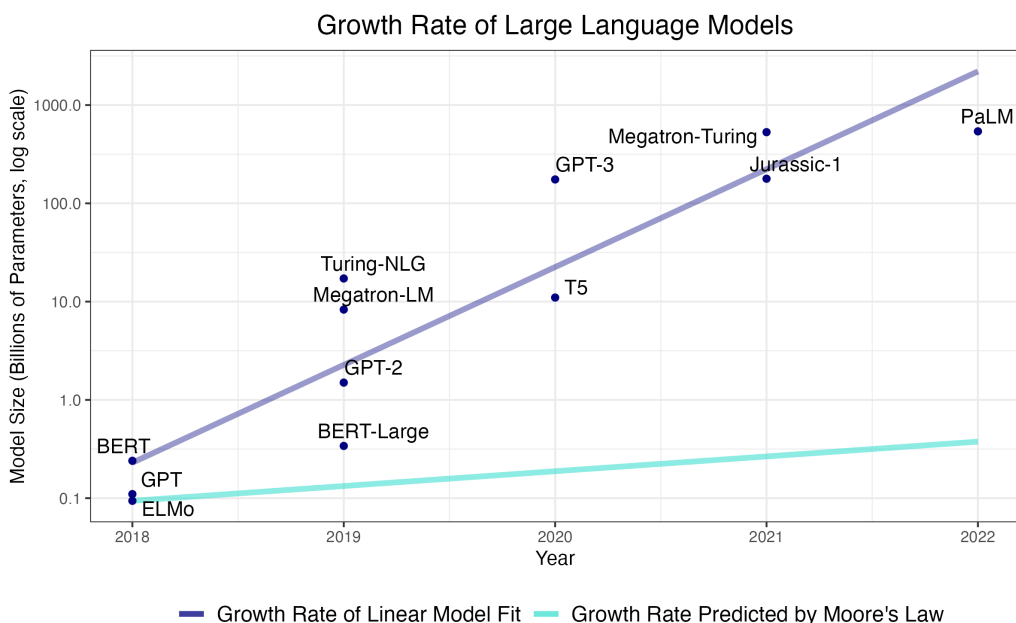
Figure 2: Growth in size of large language models compared to the predicted Moore's Law growth rate, beginning with ELMo.

more data; this would match the number of input tokens of perhaps *millions* of human learners, approaching the combined lifetime experience of all 2022 LSAT takers. What's more, as Piantadosi (2023, 14) himself acknowledges, "work examining the scaling relationship between performance and data size show that at least current versions of the models do achieve their spectacular performance only with very large network sizes and large amounts of data (Kaplan et al., 2020)." Put simply, Piantadosi's suggestion that the performance of LLMs with smaller data will greatly improve is overly optimistic, because this has *never* been the trajectory of NLP. Nothing short of a paradigm shift would be required to get researchers chasing state-of-the-art performance to work with less training data.

We have seen this play out in the field in recent years. While there have been many pushes towards smaller data and more efficient training approaches, such efforts have, as of yet, failed to substantially alter the mainstream course of the field. These pushes include the 2022 Annual Meeting of the Association for Computational Linguistics (ACL) theme track on "*Language Diversity: from Low-Resource to Endangered Languages*,"[2] the upcoming BabyLM Challenge shared task,[3] the now-completed DARPA "*Low Resource Languages for Emergent Incidents*" (LORELEI) project,[4] and the leaked document from Google discussing the success of open source order-10 billion ('small' in the current era) parameter LLaMA-variants.[5] However, for every small data workshop, there are large data workshops, like the 2021 Workshop on Enormous Language Models,[6] whose call argued that "naïve extrapolation of these trends suggests that a model with an additional 3-5 orders of magnitude of parameters would saturate performance on most current [2021] benchmarks." And of course, the continued development and uptake of GPT-family and their LLM competitors, the most

---

[2] https://www.2022.aclweb.org/post/acl-2022-theme-track-language-diversity-from-low-resource-to-endangered-languages
[3] https://babylm.github.io/
[4] https://www.darpa.mil/program/low-resource-languages-for-emergent-incidents
[5] https://www.semianalysis.com/p/google-we-have-no-moat-and-neither
[6] https://welmworkshop.github.io/

data-hungry class of NLP models ever built, is undeniable. Large, well-funded research teams are not yet investing in learning from small data to anywhere near the extent that they are investing in learning from enormous data. We hope this changes too, but we cannot count on it.

Apparently bucking the trend for exponential growth, a series of recent papers have claimed to show that LLMs can already perform well with training data that more closely resembles a human learner's input (e.g., Huebner et al., 2021; Zhang et al., 2021; Warstadt and Bowman, 2022; Hosseini et al., 2022).[7] The authors support their conclusions by training models on smaller, often domain-relevant data sets, such as a pre-processed version (Huebner and Willits, 2021) of the English subset of the CHILDES collection of child-directed speech corpora (MacWhinney, 2000), and testing their behavior on grammar test suites of the kind cited by Piantadosi (e.g., Warstadt et al., 2020; Gauthier et al., 2020; Huebner et al., 2021). The reasoning behind such test suites is as follows: models are presented with grammatical-ungrammatical sentence pairs, designed to test the model's ability to discriminate between them according to some carefully chosen syntactic (or in practice, also semantic or lexical) phenomenon. Such phenomena include, for example, coordinate structure islands, long-distance subject-verb agreement, or an appropriate choice of negative polarity items. In gradient versions of the tasks, the model assigns a probability to both sentences in the pair, and if the grammatical sentence is awarded the higher probability, the model succeeds at the test. Alternatively, in binary versions of the task (e.g., Warstadt et al., 2019), a classifier is trained on top of the LLM, and the resulting model's task is to classify sentences as grammatical or ungrammatical.

If the goal of these test suites is to show that a model is encoding knowledge of the grammar, however, further assumptions are required regarding the interpretation of the model's outputs. Firstly, one must assume that the values output by the model are, in general, a good reflection of human acceptability judgments. However, it is not immediately obvious that this is the case: Lau et al. (2017), for example, report mixed results when correlating model predictions with human acceptability judgments. Similarly, though Piantadosi cites Warstadt et al. (2019) as an example of an LSTM matching well with human judgments, he neglects to address the authors' own conclusions that the model "perform[s] far below human level on a wide range of grammatical constructions" (Warstadt et al., 2019, 625). Indeed, the mid-70% performance Piantadosi refers to is accuracy aggregated across the entire test set. When the MCC, a special case of Pearson's r for Boolean variables, is measured instead, the model achieves only about 0.3, compared to humans' 0.65-0.8 (Warstadt et al., 2019, 630). Here, it appears that good performance in terms of simple accuracy does not necessarily reflect human-likeness. Furthermore, when evaluated on controlled test sets targeting specific grammatical principles, performance is extremely mixed. The best-performing model achieves nearly 1.0 MCC on a basic SVO word-order task, but only 0.15 on the reflexive-antecedent agreement task. By contrast, Warstadt et al. (2019, 635) argue that "most humans could reach perfect accuracy," or an MCC of 1, on the same task.

But even if one were to set these quantitative concerns aside – numbers are likely to go up over time – there is a second, more fundamental underlying assumption: that a model will succeed at these tasks *if and only if* it somehow encodes something equivalent to the grammar, or at least the relevant portion of the grammar. If there are other possible explanations for the success of the model on a given test, then the tests alone can only tell us about a model's predictive abilities, not any grammar it may encode. Unfortunately, this assumption is immediately undermined. Creators of these test sets often fail to control for side-channel information that a model could exploit in order to "succeed" at the task. Since neural models are well-known to make use of such side-

---

[7]Authors differ dramatically in how much input they consider human-like. Huebner et al. (2021) train on five million words of American English child-directed speech in their smallest experiments, while Warstadt and Bowman (2022) and the BabyLM Challenge, which also focus on American English, consider 100 million words, corresponding roughly to a ten-year-old's input, to be appropriate. However, English learners express inflectional morphology, agreement, and many major syntactic phenomena within three to four years, with only a third as much input (e.g., Brown, 1973; Slobin, 2022). Many of these phenomena are evaluated in popular test suites (e.g., Warstadt et al., 2020; Huebner et al., 2021).

channel information, skepticism about apparent successes on these tests is warranted. Put simply, the tests do not convincingly show that only a model that has achieved a human-like understanding of language can succeed at these tasks.

Consider, for example, the subject-verb agreement test sentences in BLiMP (Warstadt et al., 2020), a large minimal pairs grammaticality test set. The subject-verb agreement test pairs are intended to test for long-distance agreement dependencies, with the implication that a model which succeeds should employ underlying hierarchical rather than linear representations. In our inspection of the test suite, we find that, for a full two-thirds of these sentences, the subject and verb are string adjacent (e.g., "*Most **legislatures haven't** disliked children.*") These sentences do not require a model to encode long-distance agreement. For the remaining third, there is an intervening distractor noun (e.g., "*A **niece** of most senators **hasn't** descended most slopes.*"). However, whether or not there is a distractor, it is always the first/leftmost noun that triggers agreement. Thus, a model employing "agree with the leftmost noun" would achieve perfect accuracy on the the tests even though it does not leverage anything like hierarchical structural knowledge. Indeed, theoretical linguistics provides us with the tools to more thoroughly test models, a point which we will return to in §5.2.

Despite numerous shortcomings of this type, BLiMP is a widely used test set, and will form a portion of the test data in the upcoming BabyLM Challenge. It was also used by Zhang et al. (2021), who Piantadosi cites as evidence that LLMs can learn syntax on relatively small data (10-100 million words). Due to presence of potential shortcuts in this test set, however, there is reason to be skeptical of Zhang et al.'s conclusions, and thus of Piantadosi's subsequent optimism for the prospects of small LMs. This critique is not meant to disparage the practical utility of modern LLMs or any move towards smaller LMs, but it does draw into question this kind of evidence and the wide-reaching conclusions that some researchers draw from it.[8]

The presence of potential shortcuts such as "agree with the leftmost noun" is a particular problem for evaluating ANNs in general, since these models are infamous for deftly exploiting statistical side-channel information. This is not limited to NLP alone. For example, initially promising results from CNNs trained to detect and classify skin cancer from images were overturned when it was shown that the models were actually classifying according to the presence of rulers or surgical skin ink markings (Narla et al., 2018; Winkler et al., 2019) in the positive images. The models found an inadvertent easier correlate with the task objective and focused on that instead of detecting skin cancer, creating a potentially life-threatening situation. Closer to our field, neural models have been shown to exploit the *a priori* likelihood of answers in multiple-choice visual question answering (VQA) tasks (Chao et al., 2018). A totally random baseline is expected to achieve 25% accuracy on a four-option multiple-choice test, yet the models that were tested achieved 52.9% accuracy when only exposed to the answers with no paired image or question. Since they achieved 65.7% when the task was run normally, the bulk of their performance has to be attributed to unintended statistical regularities in the distribution of multiple-choice answers rather than an understanding of the VQA task itself. This clear-cut case may also be relevant to GPT-4's performance on the multiple-choice LSAT. There are many other examples in NLP as well, including linear shortcuts in probes of linguistic structural knowledge such as patterns like "agree with the leftmost noun" and *n*-gram probabilities (McCoy et al., 2019; Kodner and Gupta, 2020) and the unintended exploitation of explicit and implicit social stereotypes in training data (Sun et al., 2019; Thompson et al., 2021), among others (Wang et al., 2022). Social biases induced by biased training data are so omnipresent that mitigation efforts have becomes a subfield unto themselves, for example spawning a series of workshops at NLP venues.[9] Given the litany of unexpected shortcuts that LLMs readily discover, it is a misjudgment to assume that they will not

---

[8]Similar points have been made regarding conclusions that can be drawn from artificial language learning experiments (Rogers and Hauser, 2009; Rogers and Pullum, 2011; Jäger and Rogers, 2012).

[9]https://aclanthology.org/venues/gebnlp/

find and take such linguistic shortcuts on the test sets of the kind cited by Piantadosi, unless it is robustly demonstrated otherwise.

Of course, the existence of these shortcuts does not mean that the LLMs subjected to these tests do not encode human-like linguistic knowledge or do not use such knowledge to solve the tests. But, it also means that success on these tests does not tell us that the models do encode linguistic knowledge either. Additional careful investigation needs to be done past just showing good performance on evaluation sets. For example, by carefully controlling test sets as in Chao et al. (2018), we can mitigate – though not necessarily remove – the opportunity for shortcuts, and by employing probes of the internal state of the models, we gain some understanding of the representations that they employ (Belinkov and Glass, 2019; Tenney et al., 2019b; Futrell et al., 2019; Liu et al., 2019; Manning et al., 2020; Linzen and Baroni, 2021; Rogers et al., 2021; Pavlick, 2022; Wilcox et al., 2022). Nonetheless, neither approach is a silver bullet, as we discuss in §4.

2.3   Section summary

Piantadosi bases his attack on nativist approaches to language science on the arguments that LLMs represent "relatively unconstrained" learners, and that successfully training such unconstrained learners on small data is not only possible, but inevitable. To the contrary, LLMs are in fact *not unconstrained*, and unconstrained learning is not possible from plausible human-sized data with feasible computational resources. These conclusions follow from fundamental computational laws that are no more violable than the conservation of energy laws in physics. Put simply, there is as much chance for feasible unconstrained learning in artificial intelligence as there is for a a perpetual motion machine in physics. Further, the prospect that engineers building LLMs will fully embrace small data is unlikely, and even if they do, current methods for evaluating such models provide little confidence that they actually encode the grammar in question, rather than exploiting statistical shortcuts to "pass" the tests.

3   Simulation is not Duplication

What can artificial intelligence tell us about natural intelligence? On the one hand, models offer existence proofs about procedures that may underlie some cognitive function. On the other hand, there is Searle's (1980, 422) critique of Turing's (1950) imitation game: "Such intentionality as computers appear to have is solely in the minds of those who program them and those who use them, those who send in the input and those who interpret the output." The human tendency to anthropomorphize may cloud our scientific judgments when it comes to inferring underlying mechanisms of complex systems. But even if one rejects Searle's position, what does imitation mean? As Chomsky (2004, 318) writes, "Imitation of some range of phenomena may contribute to this end [providing insight], or may be beside the point, as in any other domain."

3.1   Multiple Realizability

The mere fact that distinct systems exhibit the same behavior does not mean that they employ the same internal mechanisms. Guest and Martin (2023), who apply this reasoning specifically to the question of ANNs as models of cognition, present an example of two clocks, which appear identical on the outside, but are different on the inside: clock A is digital, but clock B is analog. If only the internal mechanism of clock A is known, would it be a mistake to conclude that clock B uses the same mechanism, and thus is also a digital clock? This is an incorrect conclusion despite their identical appearance and behavior. Similarly, both planes and birds can propel themselves through the air. Should we conclude that birds are powered by jet fuel because we know how to build jets but not birds?

Analogies of this sort abound. In a public discussion that Piantadosi participated in following the publicizing of his paper, Rosa Cao of Stanford asked Piantadosi about whether identical performance indicates an identical mechanism with this example:[10] two students pass an exam, but student A cheated, whereas student B genuinely understood the material. Clearly, identical performance on a test does not mean the same processes have been invoked.

Each of these examples can be described as a case of *multiple realizability*. That is, similar outcomes can be achieved by superficially similar but underlyingly distinct mechanisms of operation, whether that is the flight of birds and jets, or the grades of honest students and cheaters. Multiple realizability is a particular problem in cases where the systems under investigation are black boxes. A clock that we cannot open to inspect is a black box, as is the instructor's view on the test preparation strategies of a student. Human cognition and the internals of massive LLMs are largely black boxes as well. A common strategy for detecting multiple realizability is to shed light on the nature of the black box with other information about its internal mechanisms. For example, if we are trying to understand how a bird propels itself through the air, we do not turn to the flight of jets, because we know from other observations that animals do not burn jet fuel. Similarly, we can identify a probable cheater if we catch them passing notes with their neighbor.

Guest and Martin (2023) formalize this reasoning for application to cognitive claims drawn from ANNs. It is an inappropriate application of modus ponens to conclude that a neural model is a cognitive model because it predicts human behavior. Rather, if the neural model is (through additional evidence) a plausible cognitive model, then we should expect it to behave in human-like ways. Thus, we should ask whether what we know about LLMs suggests any underlying commonality to humans. However, like birds and airplanes, our knowledge of the two systems suggests that the underlying mechanisms are quite distinct.

We provide three examples. First, as previously discussed, ANNs rely on many orders of magnitude more data than humans do to achieve their levels of performance. Thus, from the perspective of CLT, a human language learner or LSAT test taker are solving a different learning problem than GPT training for the same tasks. Second, the learning mechanisms employed by ANNs rely heavily on backpropagation, which neuroscientists believe is a biologically implausible way to pass and update information (Lillicrap et al., 2020; Yang and Wang, 2020).[11] Third, the success of LLMs depends in part on the large "context windows" and other built-in "hyperparameters" that they use. The context window roughly refers to the length of a sequence of words these LLMs can access when generating responses to user prompts. The size of the context windows of state-of-the-art LLMs number in the thousands. GPT-3, for example, has a context window size of about 2,000, and people speculate GPT-4 has increased this anywhere from a factor of four to a factor of 20 (nobody knows for sure since OpenAI will not release the details, see §4). There is no sense in which humans have any kind of working-memory counterpart to this, which would require a perfect memory of thousands of recently observed words.

Continuing the reasoning from Guest and Martin (2023), one may also apply modus tollens to reason through negative results. If a particular ANN does not behave in a human-like way, then that one is not a good model of cognition. Earlier ANNs consistently under-performed compared to humans, so those particular implementations could be rejected. However, no model is perfect, so the failure of a particular ANN cannot lead us to conclude that ANNs as a class should be rejected too. This unfortunately renders a simple negative existence proof for a cognitively plausible ANN untenable, just as a positive result cannot be interpreted, in itself, as a positive existence proof.

---

[10] Link to a recording of the public discussion

[11] Especially at an implementational level, even when an ANN appears to employ a similar problem-solving strategy (Zipser and Andersen, 1988; Stork, 1989). A body of literature on biologically implementable equivalents to backpropagation in ANNs exists in both the machine learning and neuroscience (e.g., Mazzoni et al., 1991; Balduzzi et al., 2015; Ahmad et al., 2020), but this is primarily focused on computational equivalents or alternatives rather than supporting the notion of standard backpropagation through gradient descent "backward in time" as a biologically plausible process. It further emphasizes the implausibility of backpropagation as the term is normally used.

One recurring source of non-human-like behavior in ANNs, however, is their inability to reproduce human-like learning errors even when they achieve high levels of performance.

A classic example of this discrepancy emerged during the Past Tense Debate, a predecessor to the modern debates on the cognitive plausibilities of ANNs which raged in the 1980s and 1990s. The debate was superficially centered on computational models for the acquisition of English past tense inflection, but the fundamental issue at stake was whether connectionist ANNs, with their supposedly *tabula rasa* nature and their distributed representations, could unseat models of inflection representation and learning drawn from prevailing linguistic theory (Rumelhart and McClelland, 1986; Pinker and Prince, 1988; Pinker and Ullman, 2002; McClelland and Patterson, 2002). What's new is old, in that sense. One major observation from the old debates was divergent behavior between human learners and the ANNs of the time in terms of *overregularization* and *over-irregularization*.

Overregularization, the application of a regular/productive/default pattern to a form that should be irregular (e.g., *feeled for *felt*) is robustly attested in observational and experimental studies on the acquisition of the English past tense as well as cross-linguistically. It makes up between 5% and 10% of productions in child German (Clahsen and Rothweiler, 1993), English (Marcus et al., 1992; Xu and Pinker, 1995; Maratsos, 2000; Yang, 2002; Maslen et al., 2004; Mayol, 2007), and Spanish verbs (Clahsen et al., 1992; Mayol, 2007). Importantly, overregularization occurs *regardless* of the frequency of the productive process (Marcus et al., 1992; Belth et al., 2021). Furthermore, overregularization often leads to a *developmental regression* or *U-shaped learning trajectory*: a dip in overall production accuracy when a child learns and begins to over-apply a productive process (e.g. Marcus et al., 1992; Ravid and Farah, 1999; Clahsen et al., 2002). Over-irregularization, on the other hand, (e.g., *wing-**wang by analogy to *sing-sang*) is far rarer in the same studies, under 1% in German participles, 0.2% in English, and 0.01% in Spanish.

However, the past forty years of debate have shown a persistent failure of ANNs to replicate these human learner patterns. In the first match of the debates, Pinker and Prince (1988) already observed frequency-dependent over-generalization in the ANN of Rumelhart and McClelland (1986) and showed that the latter's apparent developmental regression was only achieved due to severely unnatural training data presentation. While the raw accuracy of much more powerful modern ANNs dwarfs that of Rumelhart and McClelland (1986), it has been repeatedly demonstrated that these various architectures still fail to produce human-like error patterns. They are still overly frequency-dependent, fail to achieve developmental regressions where appropriate, and do not yield the expected asymmetry in overregularization and over-irregularization (e.g. Corkery et al., 2019; McCurdy et al., 2020a,b; Kodner and Khalifa, 2022; Kodner et al., 2023). It is certainly possible that an ANN architecture not yet invented will address these issues, but the persistence of these problems despite dramatic engineering advances in ANN architecture and training suggests the that issue is a more fundamental characteristic of ANNs as a class.

3.2 Failures are More Informative than Successes

Given the basic problem of multiple realizability in cognitive science, it is strange that Piantadosi endorses Warstadt and Bowman's (2022) contention that an LLM's failures are scientifically less interesting than its successes. Warstadt and Bowman's reasoning is that successes count as an existence proof that at least some member of the class of artificial neural networks can solve the task, while a failure is ambiguously attributable to either a fundamental weakness of ANNs as a class or the incidentally imperfect state of the current technology. This conclusion is wrong for two reasons. The first is practical: it requires us to accept that the task that the researcher has adopted to test some property of the ANN is itself able to discriminate between a success and a failure. As we have discussed at length, however (§2), current approaches are not convincing. Warstadt and Bowman's own grammaticality evaluation test suite, BLiMP (Warstadt et al., 2020), for example, contains many weaknesses (§2, 5). Since neural models of all shapes and sizes will exploit unintended shortcuts in their input in order to take the path of least resistance (§2, e.g.,

Narla et al., 2018; Winkler et al., 2019; Chao et al., 2018; McCoy et al., 2019; Wang et al., 2022) it is reasonable to conclude that LLMs are likely "cheating" on these tests until proven otherwise. In such cases, their successes do not constitute an existence proof, but rather a pointer to areas which require further investigation.

The second reason is again the logical problem of multiple realizability. Piantadosi as well as Warstadt and Bowman (2022) draw exactly the wrong conclusion here. A positive result lends support to an approach that we have external reasons to believe is a plausible model of cognition. It cannot itself be the justification for that assertion. A negative result, on the other hand, is proof positive that this particular model, in this particular learning setting, is not an appropriate model of cognition. Of course, a related model on a related learning problem may be an appropriate model. It would be ideal if positive results could serve as existence proofs and negative results could eliminate whole classes of models, but the universe need not orient itself for our scientific convenience.

3.3   Section Summary

Somers (2013) quotes Douglas Hofstadter, "Why conquer a task if there's no insight to be had from the victory? Okay, Deep Blue plays very good chess — so what? Does that tell you something about how we play chess? No." Claiming that the human language faculty is somehow LLM-like because GPT-4 outperforms most test takers on the LSAT is akin to concluding that Kasparov interprets a chessboard like Deep Blue because it beat him, or that birds burn jet fuel because jets out-speed and out-distance small animals. But worse, it would be akin to concluding that Clock B is digital *without understanding how digital clocks work*. What would we gain from such a conclusion? Certainly neither explanation nor elucidation: we would simply replace a mystery with a black box. If the goal of a scientific theory is to provide an explanation, then it is unclear how a scientific theory of language based on LLMs might provide this, a point to which we turn in §4.

4   LLMs are not a Scientific Theory

Piantadosi argues that ChatGPT's impressive capabilities mean that it constitutes a theory of language. He is not the only writer to advocate for adopting neural networks as theories. Potts (2019), for example, advocates for adopting deep learning as a theory of semantics. However, it is unclear what such a theory tells us about language. Echoing the connectionists during the Past Tense Debate, Piantadosi mentions gradient representations, the use of word prediction as a learning signal, and the lack of built-in constraints as elements of such a "theory." However, these are rather general properties, and no specifics are offered. It is easy to see why: the role of a scientific theory is to *elucidate and explain* (Popper, 1959), and LLMs largely fail to do either.

Our argument comes in two parts. First, as corporate models, LLMs violate the best practices of open science and software development, and their results are neither replicable nor reproducible. Second, even if the models *were* open-source, we are far from understanding how they work, so they cannot presently provide a scientific explanation. All ANNs do is predict, and prediction is not explanation. Theories that made relatively accurate predictions historically, like the well-worn example of Ptolemaic epicycles on epicycles, have turned out to be incorrect, so prediction is not the end-all for judging the success of a theory. While ANNs certainly have many useful applications – fitting hypotheses to data or carrying out downstream engineering tasks – they cannot constitute scientific theories for the simple reason that they currently explain very little about language.

4.1 Corporate "Science"

The LLMs of today are a corporate product, not a scientific one. Industry dominates the creation of LLMs due to the high financial and compute costs associated with their training (Ahmed et al., 2023), and the corporations releasing these LLMs are often cagey about the details of their implementation (Liesenfeld et al., 2023). We still do not know, for example, the architecture or training data that GPT-4 uses or how often it is updated, or what kind of hand-tuning or output filters it has, because information about the product is released through public press releases rather than peer-reviewed publications as is standard in linguistics, NLP, and cognitive science. This lack of clarity means that modern LLMs are neither replicable nor reproducible, nor can the most recent LLMs be subject to the probes of internal state that earlier LLMs could be. Of course, this is likely by design: it is a savvy business strategy not to disclose the details of your model, lest a competitor beat you at your own game.[12]

Indeed, this illustrates one of the dangers of using corporate models for science: the goals of the corporations creating the models (i.e., to increase profits) are not the same as the goals of the scientists trying to probe the models (i.e., to come to a scientific understanding of language). LLMs are also constantly changing as new edge cases are found and reported, and many layers of employees are actively engaged in curating training content and guiding the outputs of the models (Perrigo, 2023; Hao and Seetharaman, 2023).[13] Again, these approaches work well for the corporations who own the models, since they want the LLMs to behave well so they can maximally profit from them. But these approaches run against the goals of science, further obfuscating the implementation details of already opaque models. The different goals of the corporations owning the models and the researchers probing them should not be taken lightly.

Consider an analogy to buying a used car: the salesperson wants to make a sale, and you want to receive information on the details and value of the car. If you wouldn't trust the salesperson to give you completely honest information about the details of the car, why would you trust corporations to do so for their LLMs? Both have the same ulterior motives: to make their product look good in order to gain maximum profit. As anyone who has bought a used car knows, these motives often lead to stretches of the truth. But there is no Carfax for LLMs: the normal process of peer review has been bypassed by industry press releases and preprint publications.

Open source LLMs, such as the newly released LLaMa 2[14] mitigate some, but not all, of these challenges but as of yet only constitute a fraction of LLM use. Additionally, LLMs billed as open source, including LLaMa 2, are also not nearly as open as their marketing would lead one to believe, suffering from poor documentation, limited or no access to training data or hyperparameters or output filtering steps, and so on (Liesenfeld et al., 2023).[15] Both the new corporate mode of publication and the importance of opensourcing models were focuses of the panel "The Future of Computational Linguistics in the LLM Age" at the recent 2023 Annual Meeting of the Association of Computational Linguistics (ACL), one of the largest gatherings of NLP researchers. This transition away from clear, open, replicable, research is clearly a concern for researchers across NLP as much as it as among linguists and cognitive scientists.

4.2 Prediction Alone is not Explanation

Even if LLMs were truly open source and all pieces were known, they would still not function as a theory of language because it is not well-understood how they work. Piantadosi (2023, 8) himself

---

[12]It is particularly surprising that Piantadosi does not express concern about this lack of openness and replicability given his past support for strong replicability requirements for scientific research (Rieth et al., 2013).

[13]Articles in Time Magazine January 18, 2023 and The Wall Street Journal July 24, 2023.

[14]Meta press release: https://about.fb.com/news/2023/07/llama-2/

[15]See https://opening-up-chatgpt.github.io/ for a growing list of open source LLM scorecards supplementing Liesenfeld et al. (2023).

acknowledges that "it [can] be hard to determine what's going on, *even though the theory is certainly in there.*" But what does it mean for a theory to be hidden in a black box?

A cottage industry has popped up over the last few years within NLP seeking to understand the inner workings and behaviors of popular ANN architectures. It goes by many names, including "explainability and interpretability" or "BERTology" (Rogers et al., 2021), though the latter term is beginning to show its age (e.g., Belinkov and Glass, 2019; Tenney et al., 2019b; Liu et al., 2019; Manning et al., 2020; Linzen and Baroni, 2021; Pavlick, 2022). This is certainly a step in the right direction, but the largest obstacle is that LLMs are, by nature, not easy to interpret. Even when all parameter values are available, which is no longer generally the case for the most powerful LLMs, it is not straightforward to map these to model behavior, and this problem is only exacerbated as model size increases. Indeed, Piantadosi himself acknowledges that "we don't deeply understand *how* the representations these models create work" (Piantadosi, 2023, 8). Understanding what goes on inside these LLMs is a bonafide research problem, but this hinders, rather than facilitates, the case for making LLMs a theory of language, as we explain in §5.

While advances in these research methodologies is progressing rapidly, the field is stuck playing catch-up with ever-evolving and increasingly opaque and corporate models. There is no equivalent to BERTology for the latest crop of LLMs because we lack the necessary access to probe them in the way we could even a few years ago. While there is obvious value in such an enterprise for the purposes of developing an understanding of state-of-the-art research tools, we ask whether it makes sense to try to explain the human mind by studying an ever-cycling menagerie of opaque human artifacts instead. We do not understand how these LLM artifacts work, they are produced by NLP researchers with entirely different goals in mind, and they will be made obsolete as soon as the next big thing is announced. If the role of scientific theory is to elucidate, then a theory of language based on LLMs does the opposite.

Even if we could pin down and perfectly probe current LLMs, these prediction machines still fall flat from the perspective of explanation. To make a classic analogy to the history of science, consider the case of Ptolemy and Copernicus: while Ptolemy was able to "fit the data" – in this case, explain the relative motion of planets – within a geocentric perspective, doing so required complicating his theory. The introduction of epicycles within epicycles within epicycles, and the fine-tuning of each planet's epicycle to best match its respective movements, eventually succeeded in fitting the observations of the day exceedingly well, and indeed *better* than Copernicus's heliocentric theory when it was introduced. But of course, we now know Copernicus's view of heliocentrism was fundamentally a step in the right direction. Later refinements of the heliocentric model turned out to not only be the correct explanation, but also a far simpler one, which not only predicts the motion of heavenly bodies but also explains them. We can see LLMs and other large statistical analyses as behaving like Ptolemy's theory: they may fit the data well, and they may even provide extremely accurate predictions. But theoretical linguistics, unlike LLMs, is able to provide a concise, explanatory account, even if – like Copernicus – this account cannot predict language use as well. Again, if the role of science is to provide elucidation and explanation, then both heliocentrism and theoretical linguistics are scientific theories, *even if* Ptolemy and LLMs win the prediction game. Piantadosi suggest that a good theory of language "*is certainly in*" the LLMs somewhere. But, if medieval astronomers had all taken that perspective on Ptolemy's model, would they have found their way to heliocentrism?

Piantadosi makes his own analogy to physics, suggesting that an ANN might be used to determine whether gravitational force falls off with distance or with distance squared. This analogy, however, confuses a tool for testing and elaborating a theory with the theory itself. In his example, the physicist already has two hypotheses (Piantadosi 2023, 7 calls them "theories"), both of which are stated with closed-form, easily interpretable, mathematical solutions ($\frac{1}{r}$ and $\frac{1}{r^2}$, respectively). In this example, the ANN – or indeed, any other means of making a maximum likelihood estimate – is only used as a tool for fitting a parameterized version of the hypotheses to the data. Crucially,

these hypotheses are generated beforehand by the underlying theory of Classical Mechanics and are not themselves a product of the maximum likelihood estimator. The role of the ANN is simply to fit a parameter to select between two possible explanations of the data; it is thus not itself the theory but rather a tool for distinguishing the predictions of two hypotheses generated by some other theory.

However, Piantadosi does not merely argue that LLMs are a useful tool for discriminating between hypotheses in theoretical linguistics. He argues that LLMs *themselves* constitute a theory that should *replace* traditional theoretical linguistics. Under such a view, the LLM would not simply be an adjudicator between hypotheses generated by existing theories, as he recognizes, "we don't explicitly 'build in' the theories under comparison" (Piantadosi, 2023, 8). However, he fails to recognize that his own analogy, which does build in and test an existing theory's hypotheses, is inappropriate because of this.

If we insist on drawing one further comparison with physics, a more fitting analogy to his argument comes from ANN applications to the three-body problem: one may observe that, despite generations of effort, theoretical physics has consistently "failed" to produce practically usable closed-form solutions to cases of this problem (because no such solution exists mathematically). However, recent approaches predicting the relative motion of three bodies statistically with ANNs (e.g., Breen et al., 2020) have shown great promise. Under Piantadosi's reasoning, the apparent superiority of these neural prediction approaches over traditional theory should render them a serious alternative theoretical basis for physics. This is, of course, absurd. ANNs and other probabilistic approximators are tools for carrying out predictions when it is too impractical or impossible to deploy a closed-form solution, not a replacement for the original underlying explanation. The three-body ANN does not tell us anything about the *theoretical bases* of the interaction of the three objects. Similarly, LLMs don't tell us anything about the theoretical bases of language merely because they make accurate predictions. The idea that a theory could be hidden in the approximator somewhere is a category error.

While we argue that Piantadosi has made an error by calling ANNs a theory of language, we agree that they have proven successful in serving as predictive models. They form the basis of increasingly useful tools for a wide range of practical applications in the sciences and elsewhere. Piantadosi (2023, 9) finds the status of LLM research "somewhat akin to the history of medicine, where people often worked out what kinds of treatments worked well (e.g., lemons treat scurvy) without yet understanding the mechanism." He also likens the field to "modeling hurricanes or pandemics" in which "the assumptions are adjusted to make the best predictions possible," but this is the same category error again. A good predictive model is not the same as a good theory. Models for predicting weather patterns and pandemics are tools in the scientific toolbox. They are not the theories themselves. The theory is our understanding of a mechanism, not merely the body of observations that spur further research. Theories in meteorology and epidemiology synthesize everything from fluid dynamics to physiology, along with direct empirical observation of real-world complex systems, and yes, computational modeling.

4.3 Section Summary

Piantadosi's endorsement of OpenAI's ChatGPT embraces corporate "science" and all the practices that it embodies: unaccessible software, data, lack of replicability, and incentives that align with the pursuit of the bottom line and not the pursuit of truth. By emphasizing prediction to the exclusion of understanding, Piantadosi promotes a disappointingly shallow interpretation of science and what it has to offer. In our view, a linguistic theory should provide explanations for linguistic capacities, not merely predict language text. This is largely concordant with the perspective of van Rooij and Baggio on the nature of theory in psychology, emphasizing explanation over prediction, an understanding of capacities over effects, and a theoretical cycle combining verbal and mathematical formalization with empirical study (van Rooij and Baggio, 2020, 2021). To call LLMs a theory

rather than a tool misses all of this entirely.

5 Why Linguistics Will Thrive in the 21st Century

In the previous sections, we argued that LLMs cannot constitute a scientific theory of language because they are largely proprietary and uninterpretable, and their focus is on *prediction*, not elucidation or explanation. In contrast, however, linguistic theory aims to provide an explanatory account of human languages. By making use of a set of abstract universals, linguistic theory seeks to concisely explain *why* languages are structured the way they are and make testable predictions about grammatical distinctions within and across the world's languages. We argue that *even if* LLMs appear to fit the data better than linguistic theory, only the latter succeeds as a scientific theory because only it provides an explanation of *why* the relevant patterns arise. Indeed, without linguistic theory, there would be no way to test the linguistic capabilities of ANNs. Test suites designed for phenomena such as subject-verb agreement or anaphora are designed to test whether LLMs encode the *distinctions provided by linguistic theory*. Similarly, work probing LLMs often seeks to find evidence for the abstract universals predicted by linguistic theory, for example hierarchical structure. Thus, generative linguistics broadly construed is a true scientific theory of language, one which will continue to thrive in the 21st century.

5.1 Linguistic Theories offer Explanations

Consider the simple example of subject-verb agreement, or the difference in grammaticality between "*I say she walks*" and "**I say she walk*" in most varieties of English. Linguistic theories provide interpretable mechanisms for enforcing this formal distinction in terms of *abstract universals* such as hierarchical structure, features, and locality. For example, in Minimalism, a theory of the structural basis of grammaticality, φ-features of the subject (person, number, gender) are copied to the verb, and syntactic structural locality constraints on the copying mechanism predict the difference in grammaticality of the sentences. The same theory that distinguishes the sentences above also makes cross-linguistic predictions about the typology of subject-verb agreement. Agreement relies primarily on syntactic structure, not linear order, so we do not expect to find a language in which the verb always agrees with the noun in the third linear position, for example. As new evidence regarding the typology of agreement is introduced (e.g., van Urk 2015), the theory is updated to account for this evidence; new explanations are found and new typological predictions are made and tested. This ability to explain and predict the syntactic relations of natural language contrasts sharply with the ability of LLMs (Moro et al., 2023). To the extent that LLMs show knowledge (i.e., predictive ability) of subject-verb agreement without the possibility of exploiting side-channel information (§2), they still do not provide a clear explanation as to *why* this difference exists, due to their lack of interpretability. Consequently, they cannot make the same kinds of cross-linguistic predictions that syntactic theory does.

The distinction between LLMs and linguistic theory outlined above is analogous to Chomsky's argument that Bayesian modeling and similar statistical methods "won't get the kind of understanding that the sciences have always been aimed at" but only "an approximation to what's happening," *despite* potentially fitting the data better than theoretical explanations. While we endorse Chomsky's position here, Piantadosi (2023, 26) quotes him critically, instead arguing in favor of simulating (such as with an LLM) emergent systems as an alternative to a "Gallilean" study of capacities that Chomsky, van Rooij and Baggio (2020, 2021), and others endorse.

Piantadosi counters Chomsky's point with the stock market, and example of an emergent system that he argues is "understood" through simulation. But, this is another poor analogy. The stock market, and the economy more broadly, are infamously chaotic systems, and financial institutions must continuously pour vast monetary and personnel resources into their efforts to keep predictions

up-to-date, profitable, and secret from the competition. Nobody should hope for a similar state of affairs in the sciences. Given that financial modeling makes many people a lot of money, it is telling that economic *theories*, not just massive predictive models, still form the basis of economic policy. For all the criticisms that can be levied against the United States Federal Reserve, they are still wise enough not to leave us at the mercy of some ANNs.

Piantadosi's second analogy to emergent behavior in beehives is better. However, it is problematic as well, because computational colony modeling does not rely solely on top-down predictive models. Rather, it also incorporates bottom-up explicit mathematical modeling of individual colony members (e.g., Belić et al., 1986; Bonabeau et al., 1998; Wittlinger et al., 2006). While linguistic theories based on emergence and self-organization exist (e.g., exemplar theory: Pierrehumbert, 2003; Ambridge, 2020; Gradoville, 2023), these resemble the top-down-plus-bottom-up study of insect colonies, not the current state of black box LLMs. Analogous bottom-up studies of individual neurons in LLMs or the impact of individual input tokens on LLMs is hampered by their truly massive size, computing demands, and proprietary nature.

5.2 Linguistic Theories Tell Us What to Look For

To even determine whether the linguistic capabilities of LLMs rival those of humans requires explicating what humans' capacities actually are; in other words, it requires a *separate theory of language*. Despite their flaws, evaluation suites of the likes of Gauthier et al. (2020), Warstadt et al. (2020), Huebner et al. (2021), and others exist *because* we have a linguistic theory that tells us what to look for. The same can be said for evaluation methods that probe representations in ANNs, for example by searching for the presence of hierarchical or long-distance encodings (e.g., Hewitt and Manning, 2019; Tenney et al., 2019a; Tucker et al., 2021; Papadimitriou et al., 2021). Consider, for example, the binding principles for anaphors (e.g., *himself, herself, themselves*) introduced in Chomsky (1981), of which Principle A accounts for the following differences in grammaticality:

(1) * I think she loves myself.

(2) * I love herself.

(3) I think she loves herself.

Principle A explains these differences in terms of the same abstract universals as were used for subject-verb agreement: hierarchical structure, features, and locality. In its essence, Principle A states that an anaphor must co-refer to another noun in the same sentence (explaining the ungrammaticality of (2), since *herself* has nothing to co-refer with), and that it must co-refer with the *hierarchically closest* eligible coreferent (explaining the ungrammaticality of (1), since *she* is eligible and closer to *myself* than *I* is). Coreference is implicated in Chomsky's and related theories by copying φ-features (person, number, gender) from the noun to the anaphor, for example by copying {3, SINGULAR, FEM} from *she* to *herself* in (3). Popular grammaticality test suites are designed to test for the encoding of Principle A in LMs,[16] but without the work of Chomsky

---

[16]Nevertheless, it is not clear that these tests even succeed at evaluating Principle A in the first place (cf. §2). BLiMP, for example, contains seven Principle A data sets. GPT-2 achieves 100% accuracy on the first one, `principle_A_c_command`. These, and all BLiMP sentences, were programmatically generated from templates, not extracted from real data, and these templates have introduced unintended regulariteis in the data which could be exploited as shortcuts. We observe at least one such shortcut for solving `principle_A_c_command`: Every single sentence begins with an optional determiner or quantifier followed immediately by a noun, and its anaphor is always the last word. The exact same "agree with the leftmost noun" linear rule that achieves perfect accuracy on BLiMP's subject-verb agreement test sentences would also achieve perfect accuracy here. This data set does not provide a good test of Principle A.

But least `principle_A_c_command` requires a model to recognize morphological agreement. The other data sets contain similar or worse faults. `principle_A_case_2` can be solved almost perfectly by just checking that the verb immediately following the anaphoric pronoun ends in *-ing*. *Almost* perfectly, because 13 'correct' sentences, all including the verb *skated*, are actually copies of the corresponding ungrammatical sentence, as in "`Leslie imagined`

([1981](#)) – and theoretical linguistics more broadly – we would not have this principle to test. We might intuitively know that ([1](#)) or ([2](#)) sound *bad*, and we could identify them as vanishingly rare in language corpora, but we need a theory to explain *why* they are bad. Theoretical linguistics gives us this explanation. Again, *even if* LLMs can perfectly discriminate sentences like ([2](#)) or ([3](#)) from sentences like ([1](#)), they still do not explain *why* the difference in grammaticality exists. Theories make concrete predictions about the *causes* of the difference in grammaticality between sentences, and these predictions can be empirically tested in ways that explicitly control for potential confounds.

5.3 Linguistic Theories Make Fundamental Distinctions

Consider an even more fundamental distinction. The sentence "*colorless green ideas sleep furiously*" was famously introduced by Chomsky ([1957](#)) to demonstrate the independence of *structural* information – the syntax – from information about *meaning and interpretation* – the semantics. All the bigrams in this sentence – *colorless green*, *green ideas*, *ideas sleep* and *sleep furiously* – are semantically infelicitous (i.e., they make little or no sense. Something cannot be both *green* and *colorless*, *ideas* cannot *sleep*, etc.). Despite this, the sentence is syntactically well-formed and shares an identical structure with plenty of mundane sentences like "*Fluffy orange cats sleep peacefully.*"

Piantadosi is confident that ChatGPT has uncovered this distinction on its own. He touts several sentences prompted from ChatGPT which he believes to be similar to *colorless green ideas sleep furiously* in that they are "rare but not impossible" Piantadosi ([2023](#), 16). But infelicity is not the same as rarity, and none of Piantadosi's sentences make this crucial distinction. For example, ChatGPT's "*blue glittery unicorns jump excitedly*" is not nonsensical in the same way as Chomsky's sentence: there is nothing impossible about being *blue* and *glittery*. If internet art is any indication, *glittery* may be the natural state for a *unicorn*, and it is not at all nonsensical for equine unicorns to *jump excitedly*. Indeed, only one of his examples contains even a possible infelicitous bigram: *clouds dream*. The bigram (with a few interpretations, not all relevant) returns over one hundred thousand hits on Google (contrast a Google search for "colorless yellow" or "crowded empty plywood").

Moreover, all of the examples provided by ChatGPT are templatic copies of Chomsky's sentence. Each has the form "Adjective1 Adjective2 Noun Verb Adverb," matching the sequence of the original sentence. Across all output sentences (with only slight deviations in the first), the initial adjective constitutes a color term and the second one has to do with being *glittery*, *shiny*, or a related word. It seems, then, that despite producing a canned explanation of Chomsky's sentence that was certainly presented many times over in its training data, ChatGPT has not implemented this explanation in generating new sentences. It does not seem to distinguish *rarity* from *infelicity*, or truly "understand" that the distinction made by Chomsky's sentence is the independence of syntactic structure. Rather, like the old ANNs of the Past Tense Debate, it resorts to frequency more than anything. Piantadosi ([2023](#), 16) acknowledges that ChatGPT "does not as readily generate wholly meaningless sentences ... likely because meaningless language is rare in the training data," but that was the point of Chomsky's sentence, which both ChatGPT and Piantadosi seem to have missed.

`herself skated around the hospital`."

Other test suites do not fair better. Zorro was designed specifically to test the abilities of LLMs trained on child-like data. However, all of Zorro's `binding-principle a` test sentences are similar to but even weaker than BLiMP's `principle_A_case_2`. The word following the pronoun always ends in *-ing* in the grammatical sentences and never in the ungrammatical. Additionally, the 3rd-to-last word always ends in *-ing* only in the grammatical sentences, and most simply, only the grammatical sentences even contain the substring *ing*. Any of these could contribute to an exploitable shortcut, which raises the question of what models actually do. As such, it is unclear what we can conclude from evaluation on this data, as discussed in [§2](#).

5.4 Section Summary

Put simply, without linguistic theory, we do not know what distinctions we expect LLMs to make, nor do we know how we expect them to encode those distinctions. Abstract universals such as hierarchical structure, features, and locality, give computational-level explanations for the patterns observed within and across languages: they explain the differences in grammaticality in sentences used to test LLMs, and they tell us what to look for when probing the internal state of the LLMs. Without linguistic theory, the possibility of testing LLMs is dead in the water. At the same time, however, linguistic theory goes far beyond benchmarking LLMs; it makes testable, interpretable predictions about the computational nature of cognitive linguistic representations and their relationships, explaining the variation that exists in the world's languages.

6   Conclusion

Large language models are the current pinnacle of achievement in NLP, and the hype surrounding them is not completely unwarranted. For the first time since the early days of "automatic language processing" in the 1950s and 1960s, the outputs of NLP research are of broad, accessible, even transformative, utility for the general population. But that does not mean that every claim regarding their transformative power is warranted as well. Our response to Piantadosi (2023) questions the role of LLMs in the science of language. Do they constitute a linguistic theory, as Piantadosi argues, or are they a just new and powerful tool? We have argued for the latter position for four main reasons. First, LLMs do nothing to refute the Poverty of the Stimulus argument: they are likely not as unconstrained as Piantadosi and others claim, and *even if they were*, this would only be possible because of the inhumanly massive amounts of training data to which they are exposed. In contrast to LLMs, children are fluent, competent speakers of their native language(s) after relatively little exposure; this is a central mystery of language learning that linguistics as a scientific discipline continues to explore in the 21st century. Second, it is inappropriate to conclude that because an LLM predicts human behavior in some way, it is a cognitive model: simulation is not duplication. Indeed, much of what we know about LLMs – and ANNs more broadly – suggests that they are in the same kind as relationship to humans as airplanes are to birds. Third, LLMs cannot constitute linguistic theories: they are, at the end of the day, uninterpretable, unaccessible corporate software, and they provide prediction rather than explanation. This point does not detract from the practical utility of LLMs for NLP, but being a powerful tool does not necessarily make for a powerful theory. Finally, to even determine whether the linguistic and cognitive abilities of LLMs rival those of humans requires explicating what humans' capacities actually are. Ergo, it requires a separate theory of language. We have concluded with a summary of why generative linguistics provides such a theory: it tells us not only what to look for in our models, but also makes testable, interpretable predictions about the computational nature of linguistic representations and their relationships.

Our four arguments against LLMs as a theory of language are non-exhaustive. They both complement and elaborate on points made by other authors (Katzir, 2023; Rawski and Baumont, 2023) and likely more to come. A rejection of Piantadosi's view is not a rejection of progress: the capacity of modern LLMs as NLP tools is still astounding, and their dramatic rate of growth suggests that further progress is on the immediate horizon. But when it comes to why children learn language the way they do, or why certain patterns surface and others do not cross-linguistically, LLMs have little to say. In the 21st century, these questions will continue to be asked and answered by linguistic theory.

Acknowledgments

References

Ahmad, N., van Gerven, M. A., and Ambrogioni, L. (2020). GAIT-prop: A biologically plausible learning rule derived from backpropagation of error. *Advances in Neural Information Processing Systems*, 33:10913–10923.

Ahmed, N., Wahed, M., and Thompson, N. C. (2023). The growing influence of industry in AI research. *Science*, 379(6635):884–886.

Ambridge, B. (2020). Against stored abstractions: A radical exemplar model of language acquisition. *First Language*, 40(5-6):509–559.

Balduzzi, D., Vanchinathan, H., and Buhmann, J. (2015). Kickback cuts backprop's red-tape: Biologically plausible credit assignment in neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Baroni, M. (2022). On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. *Algebraic structures in natural language*, pages 1–16.

Belić, M., Škarka, V., Deneubourg, J.-L., and Lax, M. (1986). Mathematical model of honeycomb construction. *Journal of mathematical Biology*, 24:437–449.

Belinkov, Y. and Glass, J. (2019). Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Belth, C., Payne, S., Beser, D., Kodner, J., and Yang, C. (2021). The Greedy and Recursive Search for Morphological Productivity. *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*.

Berko, J. (1958). The child's learning of English morphology. *Word*, 14(2-3):150–177.

Bonabeau, E., Theraulaz, G., Deneubourg, J.-L., Franks, N. R., Rafelsberger, O., Joly, J.-L., and Blanco, S. (1998). A model for the emergence of pillars, walls and royal chambers in termite nests. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 353(1375):1561–1576.

Bornstein, M. H., Cote, L. R., Maital, S., Painter, K., Park, S.-Y., Pascual, L., Pêcheux, M.-G., Ruel, J., Venuti, P., and Vyt, A. (2004). Cross-linguistic analysis of vocabulary in young children: Spanish, Dutch, French, Hebrew, Italian, Korean, and American English. *Child development*, 75(4):1115–1139.

Breen, P. G., Foley, C. N., Boekholt, T., and Zwart, S. P. (2020). Newton versus the machine: solving the chaotic three-body problem using deep neural networks. *Monthly Notices of the Royal Astronomical Society*, 494(2):2465–2470.

Brown, R. (1973). *A first language: The early stages.* Harvard University Press, Cambridge, MA.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Chao, W.-L., Hu, H., and Sha, F. (2018). Being Negative but Constructively: Lessons Learnt from Creating Better Visual Question Answering Datasets. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 431–441, New Orleans, Louisiana. Association for Computational Linguistics.

Chater, N. and Vitányi, P. (2007). 'Ideal learning' of natural language: Positive results about learning from positive evidence. *Journal of Mathematical Psychology*, 51(3):135–163.

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

Chomsky, N. (1959). A review of BF Skinner's Verbal Behavior. *Language*, 35(1):26–58.

Chomsky, N. (1980). On cognitive structures and their development: A reply to Piaget. In Piatelli-Palmarini, I. M., editor, *Language and Learning*, pages 35–52. MIT Press Cambridge, MA.

Chomsky, N. (1981). *Lectures on Government and Binding.* Foris, Dordrecht.

Chomsky, N. (2004). Turing on the "Imitation Game". In Shieber, S. M., editor, *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*, pages 317–321. MIT Press.

Clahsen, H., Aveledo, F., and Roca, I. (2002). The development of regular and irregular verb inflection in Spanish child language. *Journal of Child Language*, 29:591–622.

Clahsen, H. and Rothweiler, M. (1993). Inflectional rules in children's grammars: Evidence from German participles. In *Yearbook of Morphology 1992*, pages 1–34. Springer.

Clahsen, H., Rothweiler, M., Woest, A., and Marcus, G. (1992). Regular and irregular inflection in the acquisition of German noun plurals. *Cognition*, 45:225–255.

Clark, A. and Lappin, S. (2011). *Linguistic Nativism and the Poverty of the Stimulus.* Wiley-Blackwell.

Corkery, M., Matusevych, Y., and Goldwater, S. (2019). Are we there yet? Encoder-decoder neural networks as cognitive models of English past tense inflection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3868–3877, Florence, Italy. Association for Computational Linguistics.

De Raedt, L. (2008). *Logical and Relational Learning.* Springer-Verlag Berlin Heidelberg.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., Tomasello, M., Mervis, C. B., and Stiles, J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, pages i–185.

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., and Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Gauthier, J., Hu, J., Wilcox, E., Qian, P., and Levy, R. (2020). SyntaxGym: An Online Platform for Targeted Evaluation of Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online. Association for Computational Linguistics.

Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Kimbrough Oller, D., Hansen, J. H., and Paul, T. D. (2017). Mapping the early language environment using all-day recordings and automated analysis. *American Journal of Speech-Language Pathology*, 26(2):248–265.

Gold, E. (1967). Language Identification in the Limit. *Information and Control*, 10:447–474.

Gradoville, M. (2023). The Future of Exemplar Theory. *The Handbook of Usage-Based Linguistics*, pages 527–544.

Guest, O. and Martin, A. E. (2023). On logical inference over brains, behaviour, and artificial neural networks. *Computational Brain & Behavior*, pages 1–15.

Hao, K. and Seetharaman, D. (2023). Cleaning Up ChatGPT Takes Heavy Toll on Human Workers. *The Wall Street Journal*, July 24, 2023.

Hart, B. and Risley, T. R. (1992). American parenting of language-learning children: Persisting differences in family-child interactions observed in natural home environments. *Developmental psychology*, 28(6):1096.

Hassani, B. K. (2021). Societal bias reinforcement through machine learning: a credit scoring perspective. *AI and Ethics*, 1(3):239–247.

Heinz, J. (2016). Computational Theories of Learning and Developmental Psycholinguistics. In Lidz, J., Synder, W., and Pater, J., editors, *The Oxford Handbook of Developmental Linguistics*, chapter 27, pages 633–663. Oxford University Press, Oxford, UK.

Hewitt, J. and Manning, C. D. (2019). A Structural Probe for Finding Syntax in Word Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Hosseini, E. A., Schrimpf, M. A., Zhang, Y., Bowman, S., Zaslavsky, N., and Fedorenko, E. (2022). Artificial neural network language models align neurally and behaviorally with humans even after a developmentally realistic amount of training. *bioRxiv*, pages 2022–10.

Huebner, P. A., Sulem, E., Cynthia, F., and Roth, D. (2021). BabyBERTa: Learning More Grammar With Small-Scale Child-Directed Language. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646, Online. Association for Computational Linguistics.

Huebner, P. A. and Willits, J. A. (2021). Using lexical context to discover the noun category: Younger children have it easier. In *Psychology of learning and motivation*, volume 75, pages 279–331. Elsevier.

Jäger, G. and Rogers, J. (2012). Formal language theory: Refining the Chomsky Hierarchy. *Philosophical Transactions of the Royal Society B*, 367(1598):1956–1970.

Jain, S., Osherson, D., Royer, J. S., and Sharma, A. (1999). *Systems That Learn: An Introduction to Learning Theory (Learning, Development and Conceptual Change)*. The MIT Press, 2nd edition.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Katzir, R. (2023). Why large language models are poor theories of human linguistic cognition. a reply to piantadosi (2023). *Manuscript. Tel Aviv University. url: https://lingbuzz.net/lingbuzz/007190*.

Kharitonov, E. and Chaabouni, R. (2020). What they do when in doubt: a study of inductive biases in seq2seq learners. *arXiv preprint arXiv:2006.14953*.

Kodner, J. and Gupta, N. (2020). Overestimation of Syntactic Representation in Neural Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1757–1762, Online. Association for Computational Linguistics.

Kodner, J. and Khalifa, S. (2022). SIGMORPHON–UniMorph 2022 Shared Task 0: Modeling Inflection in Language Acquisition. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 157–175, Seattle, Washington. Association for Computational Linguistics.

Kodner, J., Khalifa, S., Payne, S. R., and Liu, Z. (2023). Re-Evaluating the Evaluation of Neural Morphological Inflection Models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Lau, J. H., Clark, A., and Lappin, S. (2017). Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive science*, 41(5):1202–1241.

Liang, B.-S. (2023). AI Computing in Large-Scale Era: Pre-trillion-scale Neural Network Models and Exa-scale Supercomputing. In *2023 International VLSI Symposium on Technology, Systems and Applications (VLSI-TSA/VLSI-DAT)*, pages 1–3.

Liesenfeld, A., Lopez, A., and Dingemanse, M. (2023). Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators. In *Proceedings of the 5th International Conference on Conversational User Interfaces*, pages 1–6.

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., and Hinton, G. (2020). Backpropagation and the brain. *Nature Reviews Neuroscience*, 21(6):335–346.

Linzen, T. and Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, 7:195–212.

Liu, N. F., Gardner, M., Belinkov, Y., Peters, M. E., and Smith, N. A. (2019). Linguistic Knowledge and Transferability of Contextual Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

MacWhinney, B. (2000). *The CHILDES project: The database*, volume 2. Psychology Press, Abingdon-on-Thames.

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., and 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.

Maratsos, M. (2000). More overregularizations after all: New data and discussion on Marcus, Pinker, Ullman, Hollander, Rosen & Xu. *Journal of Child Language*, 27(1):183–212.

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., Xu, F., and Clahsen, H. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, pages i–178.

Martínez, E. (2023). Re-Evaluating GPT-4's Bar Exam Performance. *SSRN Electronic Journal*.

Maslen, R. J., Theakston, A. L., Lieven, E. V., and Tomasello, M. (2004). A dense corpus study of past tense and plural overregularization in English. *Journal of Speech, Language, and Hearing Research*, 47:1319–1333.

Mayol, L. (2007). Acquisition of irregular patterns in Spanish verbal morphology. In Nurmi, V. and Sustretov, D., editors, *Proceedings of the twelfth ESSLLI Student Session*, pages 1–11, Dublin.

Mazzoni, P., Andersen, R. A., and Jordan, M. I. (1991). A more biologically plausible learning rule than backpropagation applied to a network model of cortical area 7a. *Cerebral Cortex*, 1(4):293–307.

McClelland, J. L. and Patterson, K. (2002). Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in cognitive sciences*, 6(11):465–472.

McCoy, T., Pavlick, E., and Linzen, T. (2019). Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

McCurdy, K., Goldwater, S., and Lopez, A. (2020a). Inflecting When There's No Majority: Limitations of Encoder-Decoder Neural Networks as Cognitive Models for German Plurals. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1745–1756, Online. Association for Computational Linguistics.

McCurdy, K., Lopez, A., and Goldwater, S. (2020b). Conditioning, but on Which Distribution? Grammatical Gender in German Plural Inflection. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 59–65, Online. Association for Computational Linguistics.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). *Foundations of Machine Learning*. MIT Press.

Montrul, S. (2004). *The Acquisition of Spanish*. John Benjamins Publishing Company.

Moro, A., Greco, M., and Cappa, S. F. (2023). Large languages, impossible languages and human brains. *Cortex*, 167:82–85.

Narla, A., Kuprel, B., Sarin, K., Novoa, R., and Ko, J. (2018). Automated classification of skin lesions: From pixels to practice. *Journal of Investigative Dermatology*, 138(10):2108–2110.

Niyogi, P. (2006). *The Computational Nature of Language Learning and Evolution*. Cambridge, MA: MIT Press.

Nowak, M. A., Komarova, N. L., and Niyogi, P. (2001). Evolution of Universal Grammar. *Science*, 291(5501):114–118.

Nowak, M. A., Komarova, N. L., and Niyogi, P. (2002). Computational and evolutionary aspects of language. *Nature*, 417:611–617.

Papadimitriou, I., Chi, E. A., Futrell, R., and Mahowald, K. (2021). Deep Subjecthood: Higher-Order Grammatical Features in Multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2522–2532, Online. Association for Computational Linguistics.

Pater, J. (2019). Generative linguistics and neural networks at 60: Foundation, friction, and fusion. *Language*, 95(1):e41–e74. doi:10.1353/lan.2019.0009.

Pavlick, E. (2022). Semantic structure in deep learning. *Annual Review of Linguistics*, 8:447–471.

Perrigo, B. (2023). OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic. *Time Magazine*, January 18, 2023.

Peters, M. E., Neumann, M., Zettlemoyer, L., and Yih, W.-t. (2018). Dissecting Contextual Word Embeddings: Architecture and Representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium. Association for Computational Linguistics.

Phillips, C. (2010). Syntax at age two: Cross-linguistic differences. *Language Acquisition*, 17(1-2):70–120.

Piantadosi, S. (2023). Modern language models refute Chomsky's approach to language. *Lingbuzz Preprint. url: https://lingbuzz.net/lingbuzz/007180*.

Pierrehumbert, J. B. (2003). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and speech*, 46(2-3):115–154.

Pinker, S. and Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28(1-2):73–193.

Pinker, S. and Ullman, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Sciences*, 6(11):456–463.

Popper, K. (1959). *The Logic of Scientific Discovery*. Routledge.

Potts, C. (2019). A case for deep learning in semantics: Response to Pater. *Language*, 95(1):e115–e124.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.*, 21(1).

Ravid, D. and Farah, R. (1999). Learning about noun plurals in early Palestinian Arabic. *First Language*, 19(56):187–206.

Rawski, J. and Baumont, J. (2023). Modern Language Models Refute Nothing. *Lingbuzz Preprint. url: https://lingbuzz.net/lingbuzz/007203*.

Rawski, J. and Heinz, J. (2019). No Free Lunch in Linguistics or Machine Learning: Response to Pater. *Language*, 95(1):e125–e135.

Rieth, C. A., Piantadosi, S. T., Smith, K. A., and Vul, E. (2013). Put your money where your mouth is: Incentivizing the truth by making nonreplicability costly. *European Journal of Personality*, 27:120–144.

Rogers, A., Kovaleva, O., and Rumshisky, A. (2021). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Rogers, J. and Hauser, M. (2009). The use of formal languages in artificial language learning: a proposal for distinguishing the differences between human and nonhuman animal learners. In van der Hulst, H., editor, *Recursion and Human Language*, chapter 12, pages 213–232. De Gruyter Mouton, Berlin, Germany.

Rogers, J. and Pullum, G. (2011). Aural Pattern Recognition Experiments and The Subregular Hierarchy. *Journal of Logic, Language and Information*, 20:329–342.

Rumelhart, D. E. and McClelland, J. L. (1986). On Learning the Past Tenses of English Verbs. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*. The MIT Press.

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

Slobin, D. I. (2022). *The Crosslinguistic Study of Language Acquisition: Volume 3*, volume 3. Psychology Press.

Somers, J. (2013). The Man Who Would Teach Machines to Think. *The Atlantic*, November 2013.

Stork (1989). Is backpropagation biologically plausible? In *International 1989 Joint Conference on Neural Networks*, pages 241–246 vol.2.

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., and Wang, W. Y. (2019). Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Tenney, I., Das, D., and Pavlick, E. (2019a). BERT Rediscovers the Classical NLP Pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Van Durme, B., Bowman, S. R., Das, D., and Pavlick, E. (2019b). What do you learn from context? probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR*.

Thompson, H. M., Sharma, B., Bhalla, S., Boley, R., McCluskey, C., Dligach, D., Churpek, M. M., Karnik, N. S., and Afshar, M. (2021). Bias and fairness assessment of a natural language processing opioid misuse classifier: detection and mitigation of electronic health record data disadvantages across racial subgroups. *Journal of the American Medical Informatics Association*, 28(11):2393–2403.

Tucker, M., Qian, P., and Levy, R. (2021). What if This Modified That? Syntactic Interventions with Counterfactual Embeddings. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 862–875, Online. Association for Computational Linguistics.

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, LIX(236):433–460.

Valiant, L. (2013). *Probably Approximately Correct: Nature's Algorithms for Learning and Prospering in a Complex World*. Basic Books.

van Rooij, I. and Baggio, G. (2020). Theory Development Requires an Epistemological Sea Change. *Psychological Inquiry*, 31(4):321–325.

van Rooij, I. and Baggio, G. (2021). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *Perspectives on Psychological Science*, 16(4):682–697.

van Urk, C. (2015). *A uniform syntax for phrasal movement: A case study of Dinka Bor*. PhD thesis, Massachusetts Institute of Technology, Department of Linguistics and Philosophy.

Vapnik, V. (1998). *Statistical Learning Theory*. Wiley, New York.

Wang, T., Sridhar, R., Yang, D., and Wang, X. (2022). Identifying and Mitigating Spurious Correlations for Improving Robustness in NLP Models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1719–1729, Seattle, United States. Association for Computational Linguistics.

Warstadt, A. and Bowman, S. R. (2022). What artificial neural networks can tell us about human language acquisition. In *Algebraic Structures in Natural Language*, pages 17–60. CRC Press.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., and Bowman, S. R. (2020). BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Warstadt, A., Singh, A., and Bowman, S. R. (2019). Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Wilcox, E. G., Futrell, R., and Levy, R. (2022). Using computational models to test syntactic learnability. *Linguistic Inquiry*, pages 1–88.

Winkler, J. K., Fink, C., Toberer, F., Enk, A., Deinlein, T., Hofmann-Wellenhof, R., Thomas, L., Lallas, A., Blum, A., Stolz, W., et al. (2019). Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatology*, 155(10):1135–1141.

Wittlinger, M., Wehner, R., and Wolf, H. (2006). The Ant Odometer: Stepping on Stilts and Stumps. *Science*, 312(5782):1965–1967.

Wolpert, D. H. and Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82.

Xu, F. and Pinker, S. (1995). Weird past tense forms. *Journal of child language*, 22(3):531–556.

Yang, C. (2002). *Knowledge and Learning in Natural Language*. Oxford University Press, Oxford.

Yang, C. (2006). *The Infinite Gift: How Children Learn and Unlearn the Languages of the World*. Simon and Schuster.

Yang, C. (2013). Who's afraid of George Kingsley Zipf? Or: Do children and chimps have language? *Significance*, 10(6):29–34.

Yang, G. R. and Wang, X.-J. (2020). Artificial neural networks for neuroscientists: A primer. *Neuron*, 107(6):1048–1070.

Zhang, Y., Warstadt, A., Li, X., and Bowman, S. R. (2021). When Do You Need Billions of Words of Pretraining Data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1112–1125, Online. Association for Computational Linguistics.

Zipser, D. and Andersen, R. A. (1988). A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature*, 331(6158):679–684.