The simple reason LLMs are not scientific models (and what the alternative is for linguistics).

Joe Collins (NTNU) *joe.collins@ntnu.no*

Piantadosia (2023) presents an argument that Large Language Models (LLMs) have refuted Chomskyan linguistics. There have been multiple responses (e.g. Katzir 2023; Kodner *et al* 2023; Milway 2023; Rawsi & Baumont 2023; etc.), and several commentors have argued that there is an important sense in which LLMs are not actually scientific theories (Kodner *et al* 2023; Rawsi & Baumont) – a point that this article agrees with. Here I argue for a mathematically explicit reason why LLMs are not scientific theories, which I have not seen fully expounded in the responses to Piantadosia (2023). This argument concurs with Chomsky's criticism that LLM-style models don't answer "why" questions. Although it disagrees with Chomsky and others on some of the specifics.

The argument can be summarized as follows: The fundamental reason that LLMs cannot be scientific theories is *not* because they are probabilistic, or because they involve parameter tuning. Nor even does it have to do with their lack of human intelligibility. As Piantadosia notes, such things are common enough among mature sciences. Rather, the issue is that the representational capacities of LLMs (and their connectionist siblings) are *unbounded* in a way that makes their representations arbitrary[1].

Neural Networks (NNs) such as LLMs belong to a broad category of Universal Function Approximators (Cybenko 1989; Yun *et al* 2019). This makes them fundamentally quite different to other statistical models, such as linear regression. When a scientist performs a linear regression, they are (implicitly) testing a specific hypothesis – namely that the variables in question are *linearly* correlated. The scientist could test a different hypothesis by fitting a different function to the data.

In the case of UFAs however, there is no such hypothesis. This is because UFAs do not fit an *a priori* function, but rather infer a function from the data. And what has been shown formally is that UFAs are, in principle, capable of approximating any mathematical function (Ismailov 2023). Therefore, proving that any given function (e.g. language) can be approximated by such a model is trivial. As Peterson (2022) puts it: "*there is no aspect of the architecture of a NN that fundamentally limits its approximation power*" (p.14).

To be clear then, even if LLMs could represent human language absolutely perfectly, this would prove nothing about human language, beyond the rather uninteresting fact that language is (at least arbitrarily close to) some mathematical function.

Piantadosia pre-empts this criticism somewhat. However, his counterargument conflates the distinction between the *representational* capacity of a Neural Network (NN) and the ability of the *training algorithm* to converge. The practical limitations of NNs stem from the *training algorithms*, whereas Piantadosias conclusions about language are based on the *representational* properties of LLMs. For example, the fact that LLMs use continuous representations is taken by Piantadosia to be a repudiation of Chomsky's assumptions to the contrary. However, **one cannot draw that conclusion from the performance of LLMs.** Because this property of the representation has nothing to do with the fact that the LLM is representing language, rather **the LLM would approximate every possible mathematical function this way.**

[1] This is an extreme and explicit example of the "multiple realisability" argument made by Kodner *et al* (2023).

In summary, the criticism that LLMs fail to answer "why" questions should not be misunderstood as a speculative or philosophical point about the nature of scientific inquiry. LLMs are not theories for an explicit reason: because they are a universal approximation method that works by summing over many, more general functions. Any such approximation will therefore have the same general properties irrespective of what is being approximated. This makes them akin to tools such as (generalized) Fourier series and Taylor expansions – not scientific theories.

Can LLMs Tell us Anything?

One consequence of the Universal Approximation Theorem is that NN architectures are not actually computational architectures in the normal sense. That is, one could not place MLPs, LSTMs, transformers, etc., into a hierarchy analogous to the Chomsky hierarchy. Different NN architectures might converge at different rates (either in terms of model size or training time) for a given function, but they all possess the same computational capacity. This makes it meaningless to debate whether (e.g.) transformers are a better model of the mind than LSTMs.

This is because, in practice, the limitations of NNs derive from the interaction of the architecture with the training algorithm and the data sets, as well as the physical limitations of time/compute/etc. Thus, if we could learn anything at all about language from LLMs, it would be from the behaviour of the training algorithms, not the representational properties of the network architecture themselves. However, it is fairly widely acknowledged, even by machine learning giants like Geoffrey Hinton, that backprop training algorithms are wildly different from human learning in all manner of fundamental ways (Hinton 2022). It is not surprising then that the behaviour of these algorithms diverges from human learning (see e.g. Evanson, Lakretz & King 2023 for discussion). As such, these algorithms can't really be considered models of human learning *per se*.

So, the fact that transformers can "learn" language better than LSTMs, for example, only really proves that transformers work better with backprop on language data. It is not an argument that transformers are a better model of human minds than LSTMs, because human minds very clearly do not rely on backprop.

At best, we could hope that the properties of these "unnatural" training algorithms can be used to detect properties of language. For example, one could imagine that the relative "learnability" of different aspects of human language might contain an implication about the relative smoothness or dimensionality of language. Such an insight could even prove to be critical for our understanding (who knows?). Crucially however, this would not be an example of LLMs functioning as a scientific model of language or the human mind. Rather this would be an example of backprop/gradient decent algorithms being treated as an *object of study*, the understanding of which enabled a new inference in some other domain. Thus, if the practical limitations of LLMs can tell us anything about language, it will be because we have first learned something interesting about the properties of their training algorithms.

What Should This Mean for Linguistics?

Despite my fundamental disagreement about LLMs, I should acknowledge that I do agree with several of Piantadosia's points on the failings of generative grammar. I agree that many of its dogmas were ultimately unhelpful, and I might even be inclined to make some of these arguments more forcefully. The rejection of quantitative methods, and information-theoretic methods in particular, left generative linguists with the tools to generate endless structures, but with no meta-language for comparing and quantifying those structures. The predictable result is decades worth of quasi-philosophical discussion about whose approach to trees/features/derivation/whatever is more

"elegant" than the next person's. The field is well overdue an admission that this approach was never going to lead anywhere interesting.

However, I would also argue that the tide has already begun turn on these issues for reasons that have nothing to do with LLMs. Various approaches that fuse formal linguistic models with information theoretic methodologies have been proposed in recent years (Levy 2008; Goldsmith & Riggle 2012; Rasin *et al* 2021; etc.). And see Kush & Dillon (2021) for general discussion of how quantitative approaches to sentence processing depend on the Chomskyan program.

Another point where I agree with Piantadosia is that the study of complexity and emergence other fields can serve as inspiration for the study of language. However, I would argue that training LLMs very much *not* like the study of complexity and emergence. In fields like statistical physics, for example, genuine *understanding* of a system comes in the form of compact, analytical descriptions, e.g. partition functions.

In fact, I would argue that adopting the methods of statistical physics would be a far smaller adjustment than even many generativists would suspect. This is because the methodology of statistical physics is in many ways comparable to "rationalist" linguistics: the researcher begins with a well-defined mathematical system before proceeding to methodically study the properties of that system. Where generative linguistics differs is that its rejection of quantitative methods restricts the discussion to hard boundaries: grammatical vs ungrammatical, finite vs infinite, etc. The statistical physicist however, is free to study the contours inside the boundaries: to them, the space is Cartesian, not Aristotelian.

There are good reasons the sorts of methods employed in statistical physics could be fruitfully employed in the study language. Many examples of this can be seen from recent years (e.g. de Giuli 2019; Collins 2020; Longobardi & Treves 2023). The latter is a particularly neat example of syntactic parameter configurations treated as a consequence of spin-glass dynamics. And the history of spin-glasses echoes much of Chomsky's "Galilean" method – they are an example of research that was pursued for decades before making any useful "predictions". As physicist Marc Mézard put it[2]: *"Spin glasses are useless. Even the most imaginative physicists, submitted to grant pressure, could not find applications for these materials."* (Mézard 2022)

Yet, as Mézard notes, decades of purely intellectual study of these "useless" models ultimately resulted in a Nobel prize for Giorgio Parisi, when it transpired that such models could give insight into complexity problems in a wide variety of fields. The history of spin-glasses in statistical physics then, bears much similarity with Chomsky's "Galilean" ideals, and is quite at odds with the purely predictive, data-driven methodology of machine learning.

I would suggest the following conclusion: The fundamental difference between fields like statistical physics and generative linguistics is the breadth of mathematical tools employed, not some disagreement about the nature of scientific inquiry. So, I agree with Piantadosia when he suggests that students of linguistics should learn a much wider range of mathematical tools than is currently the case (the tools of generative linguistics alone have proven too meagre and brittle to withstand the full complexity of real human language). However, the correct conclusion is not that Chomsky's method has been "refuted", nor that linguists should simply abandon their methodology entirely. Rather, the conclusion is that generative linguistics needs a bigger toolbox. And fields like statistical physics (among others) are exactly where we should be looking for new tools – not machine learning.

[2] Credit to Kawng-Il Ryom for bringing this to the authors attention.

Bibliography

Collins, J. S. B. (2020). The Phonological Latching Network. *Biolinguistics*, *14*, 102-129.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, *2*(4), 303-314.

DeGiuli, E. (2019). Random language model. Physical Review Letters, 122(12), 128301.

Evanson, L., Lakretz, Y., & King, J. R. (2023). Language acquisition: do children and language models follow similar learning stages?. arXiv preprint arXiv:2306.03586.

Goldsmith, J., & Riggle, J. (2012). Information theoretic approaches to phonological structure: the case of Finnish vowel harmony. Natural Language & Linguistic Theory, 30, 859-896.

Hinton, G. (2022). The forward-forward algorithm: Some preliminary investigations. *arXiv preprint arXiv:2212.13345*.

Ismailov, V. E. (2023). A three layer neural network can represent any multivariate function. *Journal of Mathematical Analysis and Applications*, *523*(1), 127096.

Katzir, R. (2023). Why large language models are poor theories of human linguistic cognition: A reply to Piantadosi. Biolinguistics, 17, Article e13153. https://doi.org/10.5964/bioling.13153

Kodner, J., Payne, S., & Heinz, J. (2023). Why Linguistics Will Thrive in the 21st Century: A Reply to Piantadosi (2023). *arXiv preprint arXiv:2308.03228*.

Kush, D., & Dillon, B. (2021). Sentence Processing and Syntactic Theory. A Companion to Chomsky, 305–324. doi:10.1002/9781119598732.ch19

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126-1177.

Longobardi, G., & Treves, A. (2023). Grammatical Parameters from a Gene-like Code to Self-Organizing Attractors. arXiv preprint arXiv:2307.03152.

Mézard, M. (2022). Spin glasses and optimization in complex systems. *Europhysics News*, *53*(1), 15-17.

Milway, D. (2023). A response to Piantadosi (2023). *Lingbuzz Preprint. url: https://lingbuzz. net/lingbuzz/007264*.

Petersen, P. C. (2020). Neural network theory. *University of Vienna*, *535*.

Piantadosi, S. (2023). Modern language models refute Chomsky's approach to language. *Lingbuzz Preprint, lingbuzz*, *7180*.

Rasin, E., Berger, I., Lan, N., Shefi, I., & Katzir, R. (2021). Approaching explanatory adequacy in phonology using Minimum Description Length. Journal of Language Modelling, 9(1), 17–66. https://doi.org/10.15398/jlm.v9i1.266

Rawski, Jon & J Baumont. 2023. Modern Language Models Refute Nothing. https://lingbuzz.net/lingbuzz/007203.

Yun, C., Bhojanapalli, S., Rawat, A. S., Reddi, S. J., & Kumar, S. (2019). Are transformers universal approximators of sequence-to-sequence functions?. *arXiv preprint arXiv:1912.10077*.