Danny Fox and Roni Katzir*

Large Language Models and theoretical linguistics

https://doi.org/10.1515/tl-2024-2005

Abstract: Some recent publications have made the suggestion that Large Language Models are not just successful engineering tools but also good theories of human linguistic cognition. This note reviews methodological and empirical reasons to reject this suggestion out of hand.

Keywords: large language models; theoretical linguistics; scientific explanation

1 Introduction

Some recent publications have made the suggestion that large language models (LLMs) are not just successful engineering tools but also good theories of human linguistic cognition. So good, in fact, as to challenge theories coming out of generative linguistics.[1] We use the term *The LLM Theory* to refer to what we take to be the underlying proposal, namely that humans (a) are born with representations that are fundamentally like those of the massive artificial neural networks that are the basis of current LLMs, and (b) learn through the same methods used by these networks. This note is a brief reminder of why the LLM Theory cannot be taken seriously.

1 Some of the work prefers to say that LLMs are good *models* rather than good *theories*. We think that this terminology is unhelpful, since it invites an unhelpful equivocation. 'Model' can be used synonymously with 'theory', in which case it implies a commitment to how things are in the world. This is the sense that is relevant to science and to the present discussion. (We will say more about theory below.) But 'model' can also be used to imply matching observations and making predictions without any ontological commitments. This can include crude approximations, often with many free parameters. It can be useful for engineering purposes but is obviously irrelevant as science. This terminological blur can be used to equivocate, with suggestions of theoretical progress (sense 1), though entirely based on data-fitting with many free parameters (sense 2). We therefore stick to 'theory' in what follows.

We thank Moshe E. Bar-Lev, Milica Denić, Yosef Grodzinsky, Julia Horvath, Nur Lan, Idan Landau, and Ezer Rasin.

***Corresponding author: Roni Katzir**, Tel Aviv University, Tel Aviv, Israel, E-mail: rkatzir@tauex.tau.ac.il
Danny Fox, Massachusetts Institute of Technology, Cambridge, MA, USA, E-mail: fox@mit.edu

First, a brief reminder of what ought to be the terms of the discussion. Like all empirical sciences, the study of linguistic cognition proceeds through inference to the best explanation.[2] The goal is to explain, not to approximate surface frequencies. Just as with laws of physics or of any other scientific domain, laws of cognition – some of which we allude to below – are isolatable parts of best explanations. As such, they are very different from surface-true regularities. A law might be operative and yet obscured on the surface by a variety of other factors. Conversely, a surface-true generalization might be the result of unenlightening accidents that have little or nothing to do with anything interesting. Data matter very much, of course, but only in light of inference to the best explanation. We might be impressed by an LLM generating a Shakespearean sonnet or by LLM activity correlating with data from brain imaging, but unless these observations bear on theory selection, they are not going to tell us much about underlying machinery.

2 A methodological point

Generative linguistics claims to have uncovered non-trivial laws under important idealizations. The LLM Theory, on the other hand, masks the putative discoveries through massive parameterization. It is therefore obvious that the LLM Theory should be rejected on simple methodological grounds: using an opaque, massively parameterized network as our working hypothesis ensures that, if there are systematic aspects of human linguistic cognition, they will remain well-hidden. Understanding the world, regardless of scientific discipline, always starts from the assumption that the world can be understood. This assumption, which is sometimes attributed to Galileo, is implicit in all scientific work we know of. To be sure, the assumption is just a working assumption; we may sometimes have to reject it and conclude that certain phenomena are not understandable. But if we start from the assumption that the world is unintelligible, we are not likely to discover anything.

3 The empirical failure

This methodological point helps explain the astounding empirical failure of the LLM Theory, to which we now turn. We structure our discussion around three key aspects of human linguistic cognition: competence versus performance; correctness versus probability; and representations. Chomsky (1957, 1959, 1965) used these three matters

2 Sometimes also referred to as abduction, among other names. See Harman (1965). Inference to the best explanation can also be cast in Bayesian terms.

to dismiss earlier inadequate theories of human linguistic cognition, and they are equally relevant in the context of the current debate.[3]

4 Competence versus performance

Our linguistic competence gives rise to hierarchical representations, among them also so-called center embedding structures. (Note that this could be argued for even if speakers could never perform center embedding in practice.) Our performance, on the other hand, is limited by memory resources, something that manifests itself in difficulties that increase with each level of center embedding. The effects of performance limitations on the surface realization of our linguistic competence leads to many situations in which linguistic laws are obscured on the surface by other factors.

The distinction between competence and performance explains why manipulations of working memory (e.g., parallel tasks, sleep, noise) affect our performance on various tasks. Bad explanations avoid the distinction altogether. Similarly, a bad explanation of a desktop computer would treat it as a finite-state automaton rather than a general-purpose (Turing-complete) computing device that happens to have a limited tape. The competence-performance distinction was discussed in detail as early as Yngve (1960), Miller and Chomsky (1963), and Chomsky (1965) and is a defining property of generative theories since then. One might have expected the LLM Theory to try to derive this fundamental aspect of human linguistic cognition. The challenge is not easy: current LLMs mimic the surface manifestations of human competence and performance within a single system; when they make human-like mistakes it is because those mistakes are what they have learned, not because their working memory has failed them. And yet meeting this challenge – of appropriately distinguishing competence and performance – is the entry requirement for any debate on the nature of language.

5 Correctness versus likelihood

Humans recognize certain things as correct – a cover term used here, to include 'appropriate', 'right', 'good' and related terms – but highly unlikely and other things

3 See Katzir (2023) and references therein for further discussion and illustration of some of the points that follow. While some of the empirical content here overlaps with the discussion in that paper, our current focus is different: we wish to emphasize the fact that the LLM Theory focuses on (flawed) approximation and does not even make a passing attempt at explanation.

as incorrect but likely. The two notions are distinct. Correct but unlikely is the hallmark of creativity and of scientific discovery. Often also taking a moral stance. As pointed out by Chomsky (1957), linguistic competence also involves a notion of correctness that is distinct from likelihood. Like the competence-performance distinction, the distinction between correctness and likelihood is a fundamental fact of human linguistic cognition. Any contender for the best theory should derive this fact, and all proposals within the generative tradition have. The LLM Theory, on the other hand, seems content with a mechanism that can only output probabilities and where nothing even remotely similar to a distinct notion of correctness has ever been identified.

6 Representations

This is the bread and butter of linguistic theory: what representations are possible as far as linguistic competence is concerned? We touch here on three fundamental aspects of linguistic competence: modularity, constituency, and entailment. In a nutshell, our meaning computations make reference to syntactic structure, but syntax is blind to almost all aspects of meaning; syntax makes crucial reference to a notion of constituency; and meaning makes reference to a notion of entailment. These are basic facts of human linguistic cognition and one might expect any theoretical contender to attempt to derive them (and not just to approximate their surface effects). And here too, the LLM Theory fails. Reasons of space prevent us from saying much more within the present note. Still, we hope to illustrate briefly using a couple of examples so as to give a sense of some of what any theory of linguistic cognition should derive (and not just approximate) with respect to constituency and interpretation.

Constituency means that linguistic representations involve tree-like structures. Many linguistic dependencies reference this internal structure, with displacement being one.[4] *Talk to Kim, John will* and *Kim, John will talk to* are good, matching the bracketing *[John will [talk [to [Kim]]]]*. By contrast, similar displacements that do not respect the brackets are bad; for example, **Talk to, John will Kim* (where the displaced material is not a constituent, i.e., does not appear within brackets). Semantics references this internal structure: while *Guess who saw which of the boys with a telescope?* is ambiguous, *Guess which of the boys with a telescope Mary saw?* is not; the same bracketing that is referenced in displacement is also referenced in the

4 Also referred to as movement, filler-gap dependency, and other terms. The terminological choice does not concern us here.

computation of meaning, and many other syntactic and semantic phenomena (Binding Theory, Laws of Ellipsis, Laws of Coordination, and more).

The licensing of negative-polarity items such as *any* and *ever* further illustrates the same point, as well as the fact that semantics makes use of entailment: such elements are bad unless they appear in constituents that form contexts that reverse entailment relations (in a sense that has been studied by Ladusaw 1979 and a large body of subsequent work). We therefore have *Kim spoke to every student who ever smoked* but not **Kim spoke to some student who ever smoked*.[5] Finally, modularity: while semantics makes reference to syntactic notions such as constituency, there is only very little semantic information that syntax is sensitive to, and this semantic information is highly formal and entirely isolated from world knowledge (see Fox 2000, esp. Section 2.5).[6]

Laws involving modularity, constituency, and entailment are among the most fundamental discoveries in linguistics; stating and explaining them has been a goal of all generative theories over the past 70 years. They provide a particularly striking illustration of the inadequacy of the LLM Theory. Current LLMs struggle to even approximate constituency and entailment, and it is unclear whether the learning method of these models allow them to acquire anything like constituency. And if the theory cannot even begin to approximate these notions, it cannot hope to derive them and explain why they are such fundamental aspects of linguistics. As to modularity, current LLM architectures are inherently nonmodular, which of course prevents them from deriving the essential modularity of linguistic competence.

7 Conclusion

The distinction between competence and performance and between correctness and likelihood are parts of all the best theories of human linguistic cognition, as are the aspects of linguistic representation that we briefly reviewed (modularity, constituency, and entailment). Many more could be listed – the topic of the standard multi-year curriculum in linguistics. As we briefly reviewed, the LLM Theory does not even come close to approximating the relevant observations. Obviously it cannot derive these properties of human linguistic cognition and without doing so it cannot be considered a scientific theory at all. But given the methodological point we started with, this is hardly surprising.

5 This is so because *Every A B* reverses entailment with respect to *A* and *Some A B* does not: *A* entails *A′* ensures that *Every A′ B* entails *Every A B* but does not ensure that *Some A′ B* entails *Some A B*.
6 As should be clear, modularity concerns competence. Whether semantic or pragmatic considerations play a role in the processing of linguistic inputs is a different matter.

References

Chomsky, Noam. 1957. *Syntactic structures*. The Hague: Mouton.
Chomsky, Noam. 1959. Review of B. F. Skinner. *Language* 35. 26–58.
Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.
Fox, Danny. 2000. *Economy and semantic interpretation*. Cambridge, MA: MIT Press.
Harman, Gilbert H. 1965. The inference to the best explanation. *The Philosophical Review* 74. 88–95.
Katzir, Roni. 2023. Why Large Language Models are poor theories of human linguistic cognition: A reply to Piantadosi. *Biolinguistics* 17. e13153.
Ladusaw, William A. 1979. *Polarity sensitivity as inherent scope relations*. Amherst, MA: University of Massachusetts, Amherst PhD thesis.
Miller, George & Noam Chomsky. 1963. Finitary models of language users. In R. Duncan Luce, Robert R. Bush & Eugene Galanter (eds.), *Handbook of mathematical psychology*, vol. 2, 419–491. New York, NY: Wiley.
Yngve, Victor H. 1960. A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society* 104. 444–466.