April 26, 2023

Why large language models are poor theories of human linguistic cognition. A reply to Piantadosi (2023).

Roni Katzir, Tel Aviv University

In a recent manuscript entitled "Modern language models refute Chomsky's approach to language", Steven Piantadosi proposes that large language models (LLMs) such as GPT-3 can serve as serious theories of human linguistic cognition. In fact, he maintains that these models are significantly better linguistic theories than proposals emerging from within generative linguistics. He takes this to amount to a refutation of the generative approach.

Piantadosi's proposal is remarkable. Not because he proposes to examine LLMs using tools from cognitive science: others have done so fruitfully before (see, e.g., Gulordava et al. 2018, Warstadt, Singh, & Bowman 2019, Lakretz et al. 2021, Baroni 2022, and Wilcox, Futrell, & Levy 2022). Rather, what makes Piantadosi's paper so surprising is his suggestion that LLMs are good theories of (actual) human cognition. Since LLMs were designed to be useful engineering tools, discovering that they teach us about how humans work would be startling indeed, akin to discovering that a newly designed drone accidentally solves an open problem in avian flight. Still, this is Piantadosi's claim, and the present note shows why it is wrong.[1]

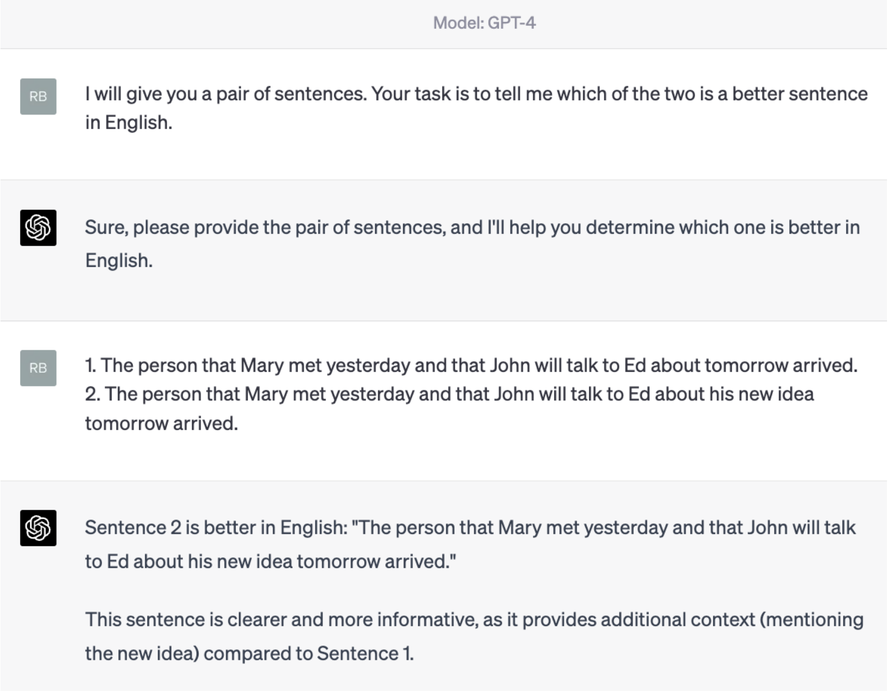
Linguistic biases and representations. If LLMs are even remotely like humans, their linguistic behavior should be similar to that of humans after training on the same amount of linguistic data. This is just a necessary condition of adequacy, and it is a weak one: a model could be very different from us and still make the same distinctions that we make after similar training. (In the sections below I mention multiple ways in which a model can differ fundamentally from us even if its linguistic preferences after training are like ours.) But current LLMs fail to clear even this low bar, showing that using them to model human linguistic cognition is misguided.

---

[1] Obviously there is nothing wrong with considering unusual theories. While scientific work typically focuses on just a handful of theories at a time — those that seem the most promising at the moment — it can be a productive exercise to occasionally examine alternative hypotheses that don't seem as promising. But as engineering products developed without cognitive plausibility as a goal, LLMs are *very* unpromising *a priori* as models of human cognition, as mentioned above, and with on the order of 10^{11} parameters in current models, they are not very likely to be particularly insightful, either. (At this level of complexity one might simply use a living child as a model.) For the purposes of the present discussion, though, I will set aside such considerations and focus on standard empirical tests of adequacy, showing that LLMs fail on all of them.

April 26, 2023

Here is a simple illustration. In the following exchange, sentence (1) is an acceptable English while sentence (2) is not, but ChatGPT thinks that the opposite is true:[2]



Model: GPT-4

RB

I will give you a pair of sentences. Your task is to tell me which of the two is a better sentence in English.

Sure, please provide the pair of sentences, and I'll help you determine which one is better in English.

RB

1. The person that Mary met yesterday and that John will talk to Ed about tomorrow arrived.
2. The person that Mary met yesterday and that John will talk to Ed about his new idea tomorrow arrived.

Sentence 2 is better in English: "The person that Mary met yesterday and that John will talk to Ed about his new idea tomorrow arrived."

This sentence is clearer and more informative, as it provides additional context (mentioning the new idea) compared to Sentence 1.

The above is not a random anecdote. It illustrates a robust, systematic generalization, originally noted by Ross (1967): if you leave a gap in one conjunct ("... that Mary met __ yesterday") you need to leave a gap in the other conjunct ("that John will talk to Ed about __ tomorrow" and not "John will talk to Ed about **his new idea** tomorrow") as well. Every English-speaking child has a solid grasp of this generalization, but as we just saw, ChatGPT does not. Other current LLMs seem to do no better: in Lan, Chemla, & Katzir (2022) we checked how four different LLMs handle Ross's generalization, and all four failed. Presumably there are too few examples of this kind in the data to support the learning of Ross's generalization from a human-sized corpus (or even one several orders of magnitude bigger) by a learner that is not suitably biased (see Pearl & Sprouse 2013 for discussion of the frequency of relevant data in corpora).[3] Children do arrive at the relevant constraint, which shows that they *are* suitably biased and that LLMs are poor theories of human linguistic cognition.[4]

---

[2] Unless otherwise specified, all illustrations use exchanges with GPT-4 from 2023-04-19.

[3] To be sure, human learners are exposed to much extra-linguistic information that current LLMs are not trained on and that could in principle give humans an unfair advantage in certain tasks (in particular, in tasks involving world knowledge). However, for Ross's generalization I know of no reason to think that such extra-linguistic information would be relevant in any way.

[4] This is an instance of the argument from the poverty of the stimulus, a central form of argument in linguistics ever since Chomsky (1971, 1975). Piantadosi claims that LLMs eliminate the argument from the poverty of the stimulus (p. 19), but as shown by the failure of LLMs to acquire constraints such as the one just mentioned, this argument is still here.

A similar point could be made with any number of textbook linguistic phenomena. See, for example, Yedetore et al. (2023) for an illustration of the same point using the phenomenon of the reversal of order between the subject and the auxiliary verb in languages like English. In fact, much of the research in theoretical linguistics concerns systematic aspects of human linguistic cognition that could provide alternative illustrations to the same point: when exposed to corpora that are similar in size to what children receive (or even much bigger corpora), LLMs fail to exhibit knowledge of fundamental, systematic aspects of language. They are too different from us to pass even this elementary test of adequacy. And obviously nothing will change in this conclusion if the next GPT, with even more parameters and trained on even more data, will be able to acquire a given generalization. The point is not that the relevant knowledge cannot be learned in principle from data. (Given suitable biases and representations, the grammatical knowledge that linguists propose can be acquired from a sufficiently large corpus. See Hsu, Chater, and Vitányi 2013 for discussion and references. Though whether LLMs can do the same is unclear.) Rather, the point is that what LLMs succeed or fail to learn from a given corpus is entirely different from what humans succeed or fail to learn from a comparable amount of information. This, in turn, shows that their representations and inductive biases are so different from ours as to make their use as a model of human linguistic cognition nonsensical.[5]

Typology. Innate biases express themselves not just in the knowledge attained by children but also in typological tendencies (that is, in cross-linguistic patterns). For example, many languages exhibit the same constraint as English does with respect to gaps in conjunction, and there are no known languages that exhibit the opposite constraint (allowing for a gap in a single conjunct but not in both). Many other universals have been studied by linguists. For example, no known language allows for an adjunct question ("how" or "why") to ask about something that has taken place inside a relative clause. So, for example, "How did you know the man who broke the window?" cannot be answered with "with a hammer," and "Why did you know the man who broke the window?" cannot be answered with "because he was angry". The same holds for all known languages in which this has been tested. A universal of a different kind is that phonological processes are always regular: they can be captured with a finite-state device that has no access to working memory. So, for example, while there are many unbounded phonological patterns, none are palindrome-like (see Heinz & Idsardi 2013). In syntax, meanwhile, roughly the

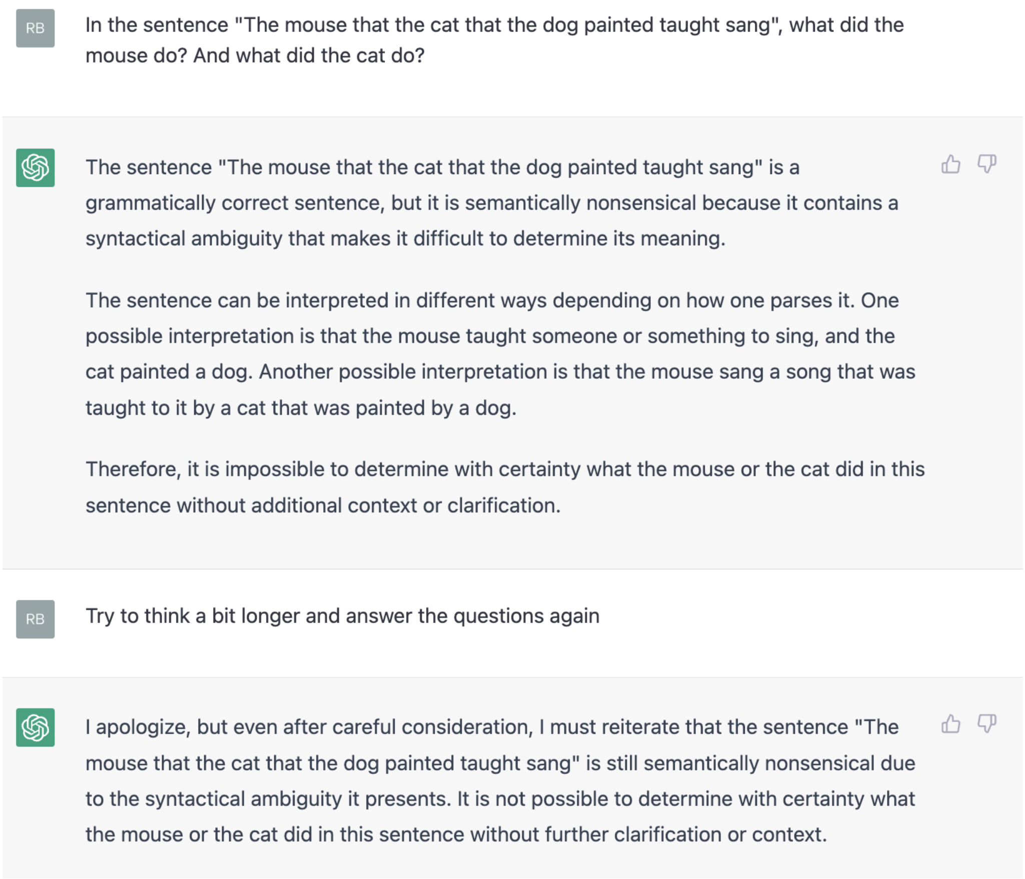
---

[5] Of course, if one is so inclined one can try to make LLMs closer to humans by designing them in advance with representations and biases that arise from research in theoretical linguistics. But doing so would amount to reversing Piantadosi's idea: rather than using LLMs as a linguistic theory one would be using linguistic theory to design LLMs.

opposite holds: dependencies are overwhelmingly hierarchical, while linear processes are rare or nonexistent. Another universal, in yet another domain, is that nominal quantifiers (“every”, “some”, “most”, etc.) are generally conservative: when they take two arguments, they only care about individuals that satisfy their first argument (Barwise & Cooper 1981, Keenan & Stavi 1986). For example, no language has a quantifier “gleeb” such that “Gleeb boys smoke” is true exactly when there are more boys than smokers (a meaning that would care about smokers that are not boys). Note that there is nothing inherently strange about such a meaning. The verb “outnumbers” expresses exactly this notion. It’s just that quantifiers seem to never have such meanings.

Universals of this kind are central reasons for linguists to argue for the innate components that they propose and that look nothing like LLMs. Piantadosi dismisses universals, referring to Evans & Levinson (2009). But Evans & Levinson do not address the universals just mentioned or those like them, and neither do the many other papers that Piantadosi lists (p. 25). LLMs are not biased in a way that would lead to these universals (nor is there any reason to think that just making future LLMs bigger and more powerful than current ones should change this), and in the absence of other explanations for why these universals should arise from an unbiased learner, LLMs remain a deeply implausible model for human linguistic cognition.

Competence vs. performance. Generative linguistics has long maintained that there is a meaningful distinction between linguistic competence and performance (Yngve 1960, Miller & Chomsky 1963, Chomsky 1965). For example, limitations of working memory are taken to be responsible for the rapidly increasing difficulty of processing sentences with center embedding (“The mouse the cat the dog chased bit died”). LLMs make no such distinction, which is generally unproblematic given their use as an engineering tool. Piantadosi, however, seems to find the lack of a distinction between competence and performance unproblematic even in their use as scientific models of human cognition. This is surprising. When humans encounter a sentence with center embedding they often make mistakes, but this is not because they don’t know how center embedding works. Rather, it is because of their struggle to harness real-time processing resources to handle the sentence under consideration. This struggle, which varies with available resources and conditions (noise, sleep, etc.) is itself evidence for the distinction between competence and performance. And given more time, human speakers often succeed where at first they failed. This is part of what a theory of linguistic cognition needs to account for, and it is what the distinction between competence and performance provides. But LLMs have no such distinction: their behavior directly reflects their competence, and when they fail it is their competence that is at fault. For example, the competence of an LLM might support the correct parse of a center-embedded construction, as seems to be the case quite generally for GPT-4. Alternatively, its competence might support an incorrect parse, as seems to be the case for GPT-3.5 in many examples. But either way, it is never the case that an LLM's competence supports the correct parse and it is their working-memory limitations that make it fail (and possibly recover given more time or memory resources). To illustrate this we need to look at failures and ask whether they stem from competence or from performance. GPT-4 is therefore unhelpful for such an illustration in the case of center embedding, and I will illustrate the point with GPT-3.5 instead:



In the sentence "The mouse that the cat that the dog painted taught sang", what did the mouse do? And what did the cat do?

The sentence "The mouse that the cat that the dog painted taught sang" is a grammatically correct sentence, but it is semantically nonsensical because it contains a syntactical ambiguity that makes it difficult to determine its meaning.

The sentence can be interpreted in different ways depending on how one parses it. One possible interpretation is that the mouse taught someone or something to sing, and the cat painted a dog. Another possible interpretation is that the mouse sang a song that was taught to it by a cat that was painted by a dog.

Therefore, it is impossible to determine with certainty what the mouse or the cat did in this sentence without additional context or clarification.

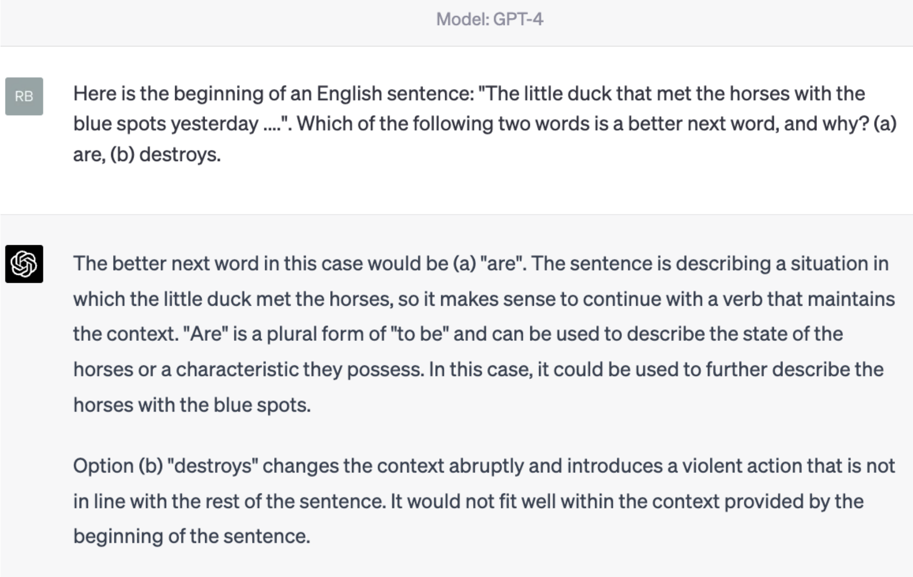
Try to think a bit longer and answer the questions again

I apologize, but even after careful consideration, I must reiterate that the sentence "The mouse that the cat that the dog painted taught sang" is still semantically nonsensical due to the syntactical ambiguity it presents. It is not possible to determine with certainty what the mouse or the cat did in this sentence without further clarification or context.

ChatGPT is clearly having difficulties with the center-embedding sentence above. As mentioned above, humans do, too. But differently from humans, and as just discussed, further time and resources are of no use to ChatGPT. When it misparses a sentence, this is because its knowledge is flawed, not because processing problems got in the way.

The competence-performance distinction helps put into context various apparent similarities between humans and LLMs that have been reported in the literature. For example, Gulordava et al. (2018) show that LLMs can learn non-local agreement and that they make agreement errors that are in some ways similar to those of human speakers. The competence-performance distinction helps illustrate that these errors are not, in fact, human-like: when humans make an agreement error and are given a chance to reread their sentence, they will often correct it; for the LLM, the error is a matter of competence and will not be corrected. This is not a problem for Gulordava et al., who avoid making claims about LLMs as linguistic theories, but it is a problem for Piantadosi's view.

Likely vs. grammatical. LLMs also lack a distinction between likelihood and grammaticality. The two notions often overlap, but they are conceptually distinct: some things are unlikely but correct, and others are likely but incorrect. Human speakers can tease these notions apart (as Chomsky 1957 pointed out and as discussed in further detail by Berwick 2018, Sprouse et al. 2018, and others). LLMs cannot: any attempt they might make to judge goodness is based on likelihood.[6] This means that they will generally prefer an ungrammatical but somewhat likely continuation over a grammatical but unlikely one:



Here is the beginning of an English sentence: "The little duck that met the horses with the blue spots yesterday ....". Which of the following two words is a better next word, and why? (a) are, (b) destroys.

The better next word in this case would be (a) "are". The sentence is describing a situation in which the little duck met the horses, so it makes sense to continue with a verb that maintains the context. "Are" is a plural form of "to be" and can be used to describe the state of the horses or a characteristic they possess. In this case, it could be used to further describe the horses with the blue spots.

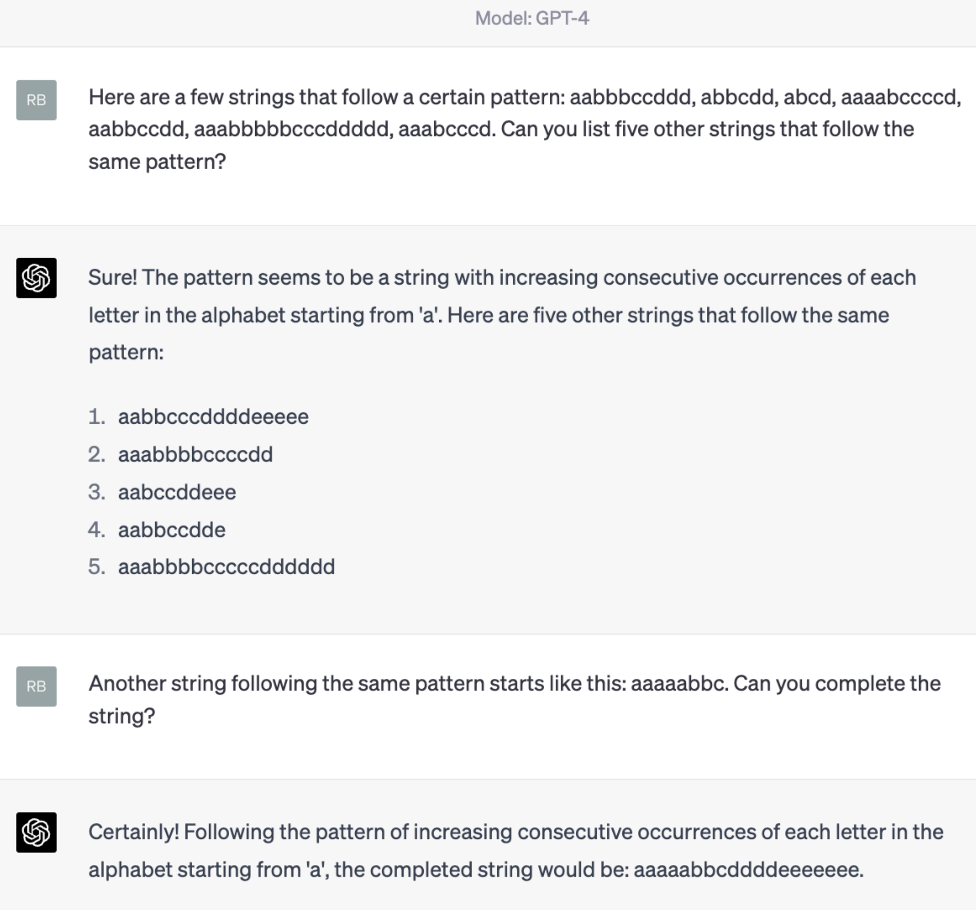
Option (b) "destroys" changes the context abruptly and introduces a violent action that is not in line with the rest of the sentence. It would not fit well within the context provided by the beginning of the sentence.

---

[6] Piantadosi suggests that probabilities are a good thing in a linguistic model because of compression (pp. 16-7). But he confuses two possible roles for probabilities: as part of linguistic knowledge and as part of the learning model. The generative approach, starting with the remarks that Piantadosi brings from Chomsky 1957, has rejected a role for probabilities within the grammar. (This rejection is based on empirical considerations rather than on any hostility to probabilities: if probabilities can be part of the grammar, one would expect to occasionally find grammatical processes that are sensitive to probabilities; such processes have not yet been found, at least in syntax and semantics, so it makes methodological sense to prevent them from being stated within the grammar in the first place.) But non-probabilistic grammars can still be learned in a probabilistic, or compression-based framework. In fact some recent generative work argues that this is exactly the correct approach to learning (Katzir 2014, Rasin & Katzir 2020).

In this case, ChatGPT chose the ungrammatical but frequent word “are” as the continuation instead of the grammatical (but perhaps unlikely) continuation “destroys”. The distinction between likely and grammatical, which all humans have, is entirely foreign to ChatGPT and its fellow LLMs.

Generalization. Taking a step back from the linguistic behavior of LLMs, let us look at how such models generalize more broadly. Piantadosi likens LLMs to automated scientists, or linguists (p. 19). But consider the following exchange with ChatGPT:



Here are a few strings that follow a certain pattern: aabbbccddd, abbcdd, abcd, aaaabccccd, aabbccdd, aaabbbbbcccddddd, aaabcccd. Can you list five other strings that follow the same pattern?

Sure! The pattern seems to be a string with increasing consecutive occurrences of each letter in the alphabet starting from 'a'. Here are five other strings that follow the same pattern:

1. aabbcccddddeeeee
2. aaabbbbccccdd
3. aabccddeee
4. aabbccdde
5. aaabbbbcccccdddddd

Another string following the same pattern starts like this: aaaaabbc. Can you complete the string?

Certainly! Following the pattern of increasing consecutive occurrences of each letter in the alphabet starting from 'a', the completed string would be: aaaaabbcddddeeeeeee.

Every string in the input sequence involves a’s followed by b’s followed by c’s followed by d’s, as any human reader can immediately see. A slightly closer glance shows that the number of a’s and c’s is the same in each string, and similarly for the number of b’s and d’s. That is, the sequence conforms to the pattern $a^m b^n c^m d^n$. ChatGPT, however, suggests that e’s can also be added (#1, #3, and #4) and seems entirely unaware that a’s and c’s should match in number or that b’s and d’s should match in number. This suggestion would be disappointing if it came from any scientist, automated or otherwise, and it is clearly very different from what humans conclude about the same data. And ChatGPT is not alone in this. In Lan, Geyer, Chemla, & Katzir (2022) we looked at several current neural networks and found that they performed poorly on a range of patterns similar to the one in the exchange above. One central factor behind this failure is the learning method: LLMs are trained using backpropagation, which pushes the network in very un-scientist-like directions and prevents it from generalizing and reasoning about inputs in anything even remotely similar to how humans generalize. When instead of backpropagation we trained networks using Minimum Description Length (MDL) — a learning criterion that *does* correspond to rational scientific reasoning — the networks were able to find perfect solutions to patterns that remained outside the reach of networks trained with backpropagation. Could there eventually be future LLMs (MDL-based or otherwise) that *would* strike us as scientist-like? Perhaps. But if such models do arrive they will be very different from current LLMs. In the meantime, any suggestion that LLMs are automated scientists should be treated with suspicion.

Summary. Piantadosi's excitement is premature. While LLMs are successful as engineering tools, we saw that they are very poor theories of human linguistic cognition. This is hardly a critique of the efforts behind these models: to my knowledge, all current LLMs were developed with engineering goals rather than the goals of cognitive science in mind. Nor do I see any reason to doubt that more human-like AI could in principle be built in the future. But current LLMs remain the stochastic parrots that Bender et al. (2021) tell us they are.[7] And, as highlighted multiple times above, simply making LLMs bigger and training them on more data will ensure that nothing changes in this regard. Using such models to write entertaining poems and short stories is one thing. Using them to understand the human faculty of language instead of doing actual linguistics is quite another.

References

Baroni, M. (2022). On the proper role of linguistically oriented deep net analysis in linguistic theorising. In Bernardy, J.-P. and Lappin, S., editors, Algebraic Structures in Natural Language, pages 1–16. CRC Press.

Barwise, J. and Cooper, R. (1981). Generalized quantifiers and natural language. Linguistics and Philosophy, 4:159–219.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?🦜. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 610-623).

---

[7] Note, however, that differently from LLMs, actual parrots *are*, in fact, intelligent.

Berwick, R. C. (2018). Revolutionary new ideas appear infrequently. In Hornstein, N., Lasnik, H., Patel-Grosz, P., and Yang, C., editors, Syntactic Structures after 60 Years: The Impact of the Chomskyan Revolution in Linguistics, volume 60, pages 177–194. Walter de Gruyter.

Chomsky, N. (1957). Syntactic Structures. Mouton, The Hague.

Chomsky, N. (1965). Aspects of the Theory of Syntax. MIT Press, Cambridge, MA.

Chomsky, N. (1971). Problems of knowledge and freedom: The Russell lectures. Pantheon Books.

Chomsky, N. (1975). Current issues in linguistic theory. Mouton, The Hague.

Evans, N. and Levinson, S. (2009). The myth of language universals: Language diversity and its importance for cognitive science. Behavioral and Brain Sciences, 32:429–492.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., and Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In Proceedings of NAACL 2018, pages 1195–1205.

Heinz, J. and Idsardi, W. (2013). What complexity differences reveal about domains in language. Topics in cognitive science, 5(1):111–131.

Hsu, A. S., Chater, N., and Vitányi, P. (2013). Language learning from positive evidence, reconsidered: A simplicity-based approach. Topics in Cognitive Science, 5(1):35–55.

Katzir, R. (2014). A cognitively plausible model for grammar induction. Journal of Language Modelling, 2(2):213–248.
Keenan, E. L. and Stavi, J. (1986). A semantic characterization of natural language determiners. Linguistics and Philosophy, 9:253–326.

Lakretz, Y., Hupkes, D., Vergallito, A., Marelli, M., Baroni, M., and Dehaene, S. (2021). Mechanisms for handling nested dependencies in neural-network language models and humans. Cognition, 213:104699. Special Issue in Honour of Jacques Mehler, Cognition's founding editor.

Lan, N., Chemla, E., and Katzir, R. (2022). Large language models and the argument from the poverty of the stimulus. Under review, available at https://ling.auf.net/lingbuzz/006829.

Lan, N., Geyer, M., Chemla, E., and Katzir, R. (2022). Minimum description length recurrent neural networks. Transactions of the Association for Computational Linguistics, 10:785–799.

Miller, G. and Chomsky, N. (1963). Finitary models of language users. In Luce, R. D., Bush, R. R., and Galanter, E., editors, Handbook of Mathematical Psychology, volume 2, pages 419–491, New York, NY. Wiley.
Pearl, L. and Sprouse, J. (2013). Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. Language Acquisition, 20(1):23–68.

Piantadosi, S. T. (2023). Modern language models refute Chomsky's approach to language. Ms., available at https://ling.auf.net/lingbuzz/007180.

Rasin, E. and Katzir, R. (2020). A conditional learnability argument for constraints on underlying representations. Journal of Linguistics, 56(4):745–773.

Ross, J. R. (1967). Constraints on Variables in Syntax. PhD thesis, MIT, Cambridge, MA.

Sprouse, J., Yankama, B., Indurkhya, S., Fong, S., and Berwick, R. C. (2018). Colorless green ideas do sleep furiously: gradient acceptability and the nature of the grammar. The Linguistic Review, 35(3):575–599.

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., Futrell, R., and Levy, R. (2022). Using computational models to test syntactic learnability. Linguistic Inquiry 2022; https://doi.org/10.1162/ling_a_00491

Yedetore, A., Linzen, T., Frank, R., and McCoy, R. T. (2023). How poor is the stimulus? evaluating hierarchical generalization in neural networks trained on child-directed speech. arXiv preprint https://doi.org/10.48550/arXiv.2301.11462

April 26, 2023

Yngve, V. H. (1960). A model and an hypothesis for language structure. Proceedings of the American Philosophical Society, 104(5):444–466.