Large languages, impossible languages and human brains

Andrea Moro[1], Matteo Greco[1], Stefano F.Cappa[1,2]

[1]Scuola Universitaria Superiore IUSS, Pavia, Italy
[2]IRCCS Mondino Foundation, Pavia, Italy

To appear in *Cortex*.

Abstract: We aim at offering a contribution to highlight the essential differences between Large Language Models (LLM) and the human language faculty. More explicitly, we claim that the existence of impossible languages for humans does not have any equivalent for LLM making them unsuitable models of the human language faculty, especially for a neurobiologically point of view. The core part is preceded by two premises bearing on the distinction between machines and humans and the distinction between competence and performance, respectively.

What matters to identify the essential differences between Large Language Models (LLM) and the human language faculty? A canonical way of reasoning is to compare their capacity and see if machines' behavior cannot be differentiated from human behavior, as in the traditional Turing test. We would like to offer a novel perspective which in a sense is reversing the perspective. Before proceeding, we would like to highlight two preliminary considerations concerning human language and machines.

1. Perception, cartesian creativity and lies:

A quick inspection on the history of AI, reveals that the area in which artificial neural networks (ANN) have been most successful is visual perception, where the machines can perform core tasks with an unprecedented accuracy. In particular, performance-optimized computational models based on deep convolutional neural networks (DCNNs) can predict the actual neurophysiological responses in macaque and human brains during the performance of object recognition tasks (Yamins et al., 2014). These results have led to the contention that ANN can be applied to other cognitive domains with the aim to "reverse engineer" the responsible brain mechanisms by relying on predictive capacities based on statistical learning and the notion of "surprisal" (Schrimpf et al., 2021; Vaswani et al., 2017). This raises at least two different and independent kind of problems.

First, it has been shown that the for the measure of surprisal to be relevant for human language it must be the case that some notion of syntactic structure, beside the basic identification of the parts of speech, must be incorporated. The bare probability for a

word to follow another one in a given corpus is not sufficient to capture even basic aspects of human language ((Greco et al., 2023)and references therein).

Second, LLM do not seem to always provide a correct analysis of linguistic structures in a comprehensive way. A crucial case study is the one provided in (Lorusso et al., 2019). The work reports the analysis of a very simple string, namely "noun phrase verb noun phrase", where the verb is the copula. Modern linguistics recognized two completely different types of copular sentences of this type, exemplified by two sentences like *a picture of the wall was the cause of the riot* vs. *the cause of the riot was a picture of the wall* (cf. (Everaert and Van Riemsdijk, 2008). These two apparently identical syntactic structures, actually involving the same lexical items, are in fact instances of two opposite symmetrical structures: one where the subject (a picture of the wall) precedes the predicate (the cause of the riot), and the other where the predicate precedes the subject, respectively. These two structures have many very different properties. A prototypical simple contrast is the following: *which riot do you think that a picture of the wall was the cause of?* vs. **which wall do you think that the cause of the riot was a picture of?* Pretrained parser as well as "Google translator" are shown to miss this basic distinction completely. Humans do it much better.

A separate issue pertains to the output of LLM. Ever since Descartes, it is commonly claimed that Human language production is "creative" in a technical sense, i.e. linguistic expressions can be generated as stimulus independent activity whether or not they are ultimately uttered or remain inside the mind as endophasic activity or "inner speech" (Magrassi et al., 2015). LLM on the other hand are obviously lacking creativity in this sense. Interestingly, Descartes, in fact, used this very notion of language creativity to distinguish between humans vs. animals which he considered as machines essentially: "There are no men so dull and stupid, not even idiots, as to be incapable of joining together different words, and thereby constructing a declaration by which to make their thoughts understood; and on the other hand, there is no other animal, however perfect or happily circumstanced which can do the like." (Descartes, 2006, Part V).

Another crucial issue is the machines' lack of awareness of truthfulness of their texts and opinions. This can be synthesized by claiming that a machine like ChatGPT cannot lie. Indeed, it can be programmed to say the opposite of what is statistically more frequent, commonly acknowledged or contingently measured – for example, a machine can say that it's raining in Manhattan if it is informed that it is not - but this can by no means considered to be an instance of a lie, as it would lack any willingness.

The failure to analyze basic syntactic structures and the impossibility to lie are surely two major reasons not to adopt LLM as models for human language faculty. Nevertheless, we would like to provide a different independent novel reason which we consider as crucial. Let us first consider a second preliminary issue.

2. Competence vs. Performance in machines and patients.

Another issue which should be addressed when comparing machines and humans is the fundamental distinction between the general knowledge of a grammar (competence, or potential knowledge in the Aristotelian terms) and the actual exploitation of this knowledge in understanding and producing linguistic expressions (performance) in overt and inner speech as well as in any additional modality (reading, writing, signing), roughly corresponding to the Aristotelian distinction between potentiality and actual realization

(Chomsky, 1966). Obviously, competence – being a potentiality - is not directly accessible through observation by definition. Competence can be reconstructed only via assembling the explanations of the data obtained by testing single performance acts and integrating them in a global model: this can happen in several independent ways, for example via grammaticality judgments, behavioral tests (reaction times, eye tracking, etc.), neuroimaging (PET, fMRI), neurophysiology (EEG, ERP, MEG, TMS) or invasive techniques (SEE, IcEEG).

Given these premises, LLM cannot add any contribution to our understanding of human competence of language: any performance test on a machine is by definition unreliable since reaction times measured on machines strictly depends on hardware and technological factors which cannot be compared to brain reaction. Needless to say, any direct question about competence posed to a machine would be as doomed as any other metatheoretical similar question, like overt comment on grammatical structures, to a machine, since even humans do not have direct access to it and this is obviously strictly dependend on the model adopted.

A different issue pertaining to LLM and human faculty of language regards the relevance of clinical studies. Patient studies are based on pattern of preserved and impaired language performance by subjects affected by brain damage. Imaging and neurophysiological experiments are based on the collection of pieces of evidence about brain activity while subjects are engaged in language processing task, most typically sentence comprehension, i.e. a language performance task. The recent findings that the most powerful ANN models are able to predict almost the totality of the explainable variance in neural responses to sentences collected with two different modalities (functional MRI and electrocorticography) (Schrimpf et al., 2021) is a signature of the excellent performance of the "new artificial intelligence" in natural language processing, but by no means does this fact necessarily imply their isomorphism with brain computations. While there is no doubt that statistical factors, such as surprisal, play a role in human information processing – as we noted before – but that there is robust evidence from neuroscience studies that distinct neural mechanisms are involved in sequential vs. hierarchical processing in the primate brain (Chao et al., 2018). Interestingly, as for the case of sentence processing, a recent study indicated that ANN models are dependent on both mechanisms, while the reliance of human performers is dominated by structure-based computations (Nelson et al., 2017; Zacharopoulos et al., 2022). The central role of hierarchical computation in language processing is also indirectly supported by the fact that abstract multi-word representations are actually emerging, without explicit supervision, in models trained exclusively for sequential word prediction(Lakretz et al., 2021). Of course it remains to prove that the two hierarchical mechanisms, the natural one and the artificial one, are isomorphic, let alone essentially the same one.

All in all, the direct comparison of actual examples of linguistic interaction with machines would not allow us to any conclusive remark on whether LLM are suitable models for human language (Jonas and Kording, 2017), insofar as they cannot be really compared to models of competence (Katzir, 2023; Lampinen, 2022). Nevertheless, there is a third issue we can take into consideration upon which we can address the fundamental question proposed here as to what matters to identify the essential difference between LLM and the human language faculty.

3. Impossible languages and the brain.

Neuroimaging techniques, has allowed scientists to cast a bridge between theoretical linguistics, in particular theoretical syntax, and brain activity (Embick and Poeppel, 2006; Cappa, 2012)and references therein). More specifically, a robust correlation between linguistic theory and neurobiology has been established which we can capitalize on when considering LLM and human brains.

Evidence based on comparative analysis of different languages across the world proved that only a subset of possible grammars is actually realized, namely those grammars based on hierarchical syntactic structures, generated by recursive rules (Berwick and Chomsky, 2016). More precisely, based on purely comparative data, this is supported by the fact that the opposite type of rules based on linear order ("flat rules") are never found in any language of the world nor in children's spontaneous production. A simple prototypical case study is offered by subject verbal predicate agreement in a language like English. A noun like *Mary* would trigger agreement on a verb like *sing* yielding: *Mary sings*. Suppose now *Mary* is embedded in a hierarchically larger constituent, say *the friends of Mary*: if this larger constituent is syntactically connected with the same verb the correct grammatical output would be *the friends of Mary sing*, not **the friends of Mary sings*, although *Mary* is adjacent to the verb *sing* exactly as in the previous example. Simply, the syntax of human language ignores the physical realization of a string of words, i.e. its linear order, while it computes hierarchical (recursive) structures, only. The adaptive reason as to why this restriction holds is arguably to simplify computation by infants and let them converge on their grammar in a reasonable amount of time, given the severe restrictions imposed by evolution on brain plasticity (Berwick and Chomsky, 2016; Friederici et al., 2017) reducing in fact spontaneous language acquisition to a selective process within the realm of possible grammars as proposed in (Mehler and Dupoux, 2002)

All in all, the distinction between possible vs. impossible grammar turned out not to be "a cultural or arbitrary convention" to use Eric Lenneberg's own seminal words (see Lenneberg, 1967). The empirical proof is that when human brains compute impossible languages, the canonical networks selectively associated to language computation, either with real words or pseudowords, are progressively inhibited (Tettamanti et al., 2002; Musso et al., 2003; Moro, 2016) for a general presentation). In other words, the distinction between possible vs. impossible languages constituting the "boundaries of Babel" is crucially an embodied one.

This very distinction turns out to be very useful since it provides us with a different and novel point of view to distinguish between LLM and the human language faculty. In fact, since the distinction between possible vs. impossible languages cannot be formulated by definition for LLM, neither formally nor empirically, we can conclude that there can be no equivalent of "impossible language state" for any machine programmed by these models. Synthesizing, machines appear to be able to compute all sorts of impossible languages, including those based on "flat", i.e. non-hierarchical rules (Moro, 2023). Indeed, LLMs and also other types of transformer models learn impossible grammars just as well as human grammars (Chomsky and Moro, 2022)and references therein). Ultimately, we can conclude that the reason why LLM are not good models for the human language faculty is not that they just can't reach our competence. The reason is rather quite the opposite: they do outperform us, showing that the real difference between machines and humans is that the former do not have our limits.

4. Concluding remarks: we are our limits.

LLMs and the machines which are programmed and trained by relying on them such as ChatGPT cannot be considered as suitable models for human languages for at least three independent reasons, each with a different force: (i) the lack of cartesian creativity and awareness; (ii) the lack to simulate human's competence in dealing with for some basic elementary structures; (iii) ultimately, the fact that there is no comparable state for the machine to the "Impossible language state" characterizing human brains. In other words, LLM do not have intrinsic limits nor any similar hardware correspondence. In synthesis: machines lack any embodied syntax which is in fact the fingerprint of human language.

All in all, LLMs such as ChatGPT, despite their (potential) utility for language tasks, can by no means be considered as isomorphic to human language faculty as resulting from brain activity and as such they can at best offer data reflecting third factor properties in the sense of Chomsky, namely "principles not specific to the faculty of language" (Chomsky, 2005) . Our limits, which make language acquisition possible, cannot be even defined with respect to machines whose tasks and nature are completely different. Eventually, we are our limits.

References

Berwick RC & Chomsky N (2016). Why only us: Language and evolution, MIT press.

Cappa SF (2012). Imaging semantics and syntax. Neuroimage 61: 427-431.

Chao ZC, Takaura K, Wang L, et al. (2018). Large-scale cortical networks for hierarchical prediction and prediction error in the primate brain. Neuron 100: 1252-1266. e1253.

Chomsky N (1966). Cartesian Linguistics: a chapter in the history of rationalist thought Harper and Row. London.

Chomsky N (2005). Three factors in language design. Linguistic inquiry 36: 1-22.

Chomsky N & Moro A (2022). The secrets of words, MIT Press.

Descartes R (2006). A Discourse on the Method, OUP Oxford.

Embick D & Poeppel D (2006). Mapping syntax using imaging: problems and prospects for the study of neurolinguistic computation. Encyclopedia of language and linguistics 2: 484-486.

Everaert M & Van Riemsdijk HC (2008). The Blackwell companion to syntax, John Wiley & Sons.

Friederici AD, Chomsky N, Berwick RC, et al. (2017). Language, mind and brain. Nature human behaviour 1: 713-722.

Greco M, Cometa A, Artoni F, et al. (2023). False perspectives on human language: Why statistics needs linguistics. Frontiers in Language Sciences 2: 1178932.

Jonas E & Kording KP (2017). Could a neuroscientist understand a microprocessor? PLoS computational biology 13: e1005268.

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.

Lakretz Y, Hupkes D, Vergallito A, et al. (2021). Mechanisms for handling nested dependencies in neural-network language models and humans. Cognition 213: 104699.

Lampinen AK (2022). Can language models handle recursively nested grammatical structures? a case study on comparing models and humans. arXiv preprint arXiv:2210.15303.

Lenneberg E (1967). 1967: Biological foundations of language. New York: John Wiley.

Lorusso P, Greco MP, Cristiano C, et al. Asymmetries in extraction from nominal copular sentences: a challenging case study for nlp tools. Proceedings of the Sixth Italian Conference on Computational Linguistics CLiC-it 2019 (Bari, November 13-15, 2019), 2019. CEUR.

Magrassi L, Aromataris G, Cabrini A, et al. (2015). Sound representation in higher language areas during language generation. Proceedings of the National Academy of Sciences 112: 1868-1873.

Mehler J & Dupoux E (2002). Naître humain, Odile Jacob.

Moro A (2016). Impossible languages, MIT press.

Moro AC (2023). Embodied syntax: impossible languages and the irreducible difference between humans and machines. SISTEMI INTELLIGENTI 2.

Musso M, Moro A, Glauche V, et al. (2003). Broca's area and the language instinct. Nature neuroscience 6: 774-781.

Nelson MJ, El Karoui I, Giber K, et al. (2017). Neurophysiological dynamics of phrase-structure building during sentence processing. Proceedings of the National Academy of Sciences 114: E3669-E3678.

Schrimpf M, Blank IA, Tuckute G, et al. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. Proceedings of the National Academy of Sciences 118: e2105646118.

Tettamanti M, Alkadhi H, Moro A, et al. (2002). Neural correlates for the acquisition of natural language syntax. Neuroimage 17: 700-709.

Vaswani A, Shazeer N, Parmar N, et al. (2017). Attention is all you need. Advances in neural information processing systems 30.

Yamins DL, Hong H, Cadieu CF, et al. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proceedings of the national academy of sciences 111: 8619-8624.

Zacharopoulos C-N, Dehaene S & Lakretz Y (2022). Disentangling Hierarchical and Sequential Computations during Sentence Processing. bioRxiv: 2022.2007. 2008.499161.