The limitations of large language models for understanding human language and cognition

Author 1[*1], Author 2[1], Author 3[*2]

[1] Affil 1
[2] Affil 2

Abstract:
Several influential researchers have recently argued that the capabilities of Large Language Models (LLMs) – whether it be their impressive performance, or specific limitations - can provide new insights into longstanding debates about the role of learning and/or innateness in human language. Here, we argue on two grounds that LLMs alone tell us very little about human language and cognition in terms of acquisition and evolution. First, any similarities between human language and the output of LLMs are purely functional. In other words, *what* LLMs do is superficially similar, but *how* they do it is not: the input LLMs learn from is fundamentally different from human input. In contrast to the rich - but quantitatively limited - multimodal data humans leverage in language learning, LLMs rely on vastly greater quantities of unimodal text data, and even recent multimodal efforts only entail mappings between images and text. Second, we turn to functional similarities between human language and LLM output, and show that these are also more limited than some authors claim, in particular because the forms LLMs can produce are underpinned by text. LLMs were designed to imitate the very specific behavior of human *writing,* but human language is a much broader phenomenon. In sum, due in large part to its focus on text, we argue that LLM performance cannot provide any direct or immediate insights into mechanistic questions about human language and cognition, and shares limited functional similarity with human language.

1. Introduction

Text generated by large language models (LLMs) is now often indistinguishable from human written text, creating acute risks ranging from widespread threats to job security (Eloundou, Manning, Mishkin & Rock, 2023), to undetectable, rapidly spreading disinformation (Pan et al., 2023). For many cognitive scientists, the apparent ability of this technology to pass a text-based Turing test is striking. Recent work has argued that the impressive ability of LLMs to generate text (or its specific mistakes or limitations) can provide insights into how human linguistic

cognition works. For example, Contreras Kallens, Kristensen-McLachlan and Christiansen (2023) suggest that LLMs' striking success shows that "grammatical language can be acquired without the need for a built-in grammar". They, and others (e.g., Piantadosi, 2023), make this argument in specific contrast to theories which propose that domain-specific, often "innate"[1] capacities are essential to human language acquisition. At the same time, proponents of domain-specific, "innate" approaches focus on some specific *shortcomings* in LLM performance, arguing that these provide support for domain-specific accounts (Chomsky, Roberts & Watumull, 2023), or at least leaves these accounts with more explanatory power (Katzir, 2023). In short, both sides of a complex, heated discussion with roots in philosophical discussions about the roles of nature and nurture in language learning have declared LLMs relevant to this debate (see Pleyer & Hartmann, 2019 for a detailed discussion). Researchers on each side of the issue have declared that LLMs provide support for their particular theory of how human language and cognition work.

In the interest of transparency, the authors have mixed views on this issue. Both Author 1 and Author 3 work in more usage-based traditions which emphasize the role of learning, and consider domain general learning an important force in language, with arguments for domain-specific claims requiring very strong evidence. Author 2's work frames investigations into language acquisition in the generativist tradition, and is sympathetic to the idea that language-specific learning processes or biases are necessary for children to make sense of their language input, particularly in the timeframe and trajectory they do. In all, while we occupy different parts of the spectrum of this debate, none of the authors takes an especially strong stance. We consider it likely that complex interactions between input (including social learning,

[1] We introduce scare quotes for the term "innate" because of the generally ill-defined nature of what exactly this would mean, particularly for capacities relating to complex behavioral traits like language (see Mameli & Bateson, 2011, for a discussion). Also note that domain specificity and innateness, while often co-morbid theoretical commitments (innate and domain specific for nativists, and learned and domain general for usage-based theories), need not be entwined (e.g., a domain specific language capacity could be entirely learned).

cultural transmission, and interaction) and robustly developing cognitive capacities (potentially including some specialization for particular features of language like recursive computation) contribute to language acquisition. In short, our arguments here are not in service of using LLMs to advance this debate in one direction or the other. Rather, our key argument is that the *performance* of LLMs alone cannot meaningfully advance this debate *at all*.

We situate our arguments within a Tinbergian framework for understanding complex evolved traits (Tinbergen, 1963; see also Bateson and Laland, 2013), including complex behavioral traits like human language (Spike, 2017; Scott-Phillips, Dickins & West, 2011). This framework considers traits in terms of ultimate (why) and proximate (how) questions. In ultimate terms, we can investigate the evolution of a trait (its *phylogeny*), and why it evolved (its *adaptive function*). Proximately, we can look at how a trait works: how it develops over the lifespan of an organism (its *ontogeny*), and how it works within the organism (the *mechanism*, e.g. how something works in the brain). First, we argue that the phylogeny and ontogeny of language in LLMs differ fundamentally from human language, limiting their explanatory power in terms of how humans learn language, and evolved the capacity to do so. Second, we argue that even similarities in the function of LLMs and human language are severely limited, further constricting their relevance for understanding fundamental questions about human language learning and evolution. Overall, we conclude by arguing that key advances are needed in understanding the exact nature of human language input before LLMs can become a useful tool to further understand human language and cognition.

2. Barking up the wrong hot air balloon

In ultimate terms, traits can be *homologous* (sharing phylogeny and adaptive function, like the wings of a robin and a blackbird) or *analogous* (sharing adaptive function, but evolving independently, like the wings of a robin and a bat). The evolutionary ancestry of human language and LLMs may be more intertwined than a robin's wing and a bat's wing, but they are

nonetheless *analogous*: human language and LLMs share similar adaptive function, but arose from completely distinct phylogenetic processes. While human language is the emergent product of natural and cultural selection, LLMs are the product of intentional human design. Studying LLMs to learn about human language is not like studying a robin wing to learn about flight in birds, or even like studying bat wings to learn about flight in birds (or in general terms). It is like studying a hot air balloon to learn about flight in naturally evolving organisms.

The proximate dynamics of human language development - how it develops over the lifespan, and by what mechanisms - differ fundamentally from LLMs. The ontogeny of human language differs markedly from how LLMs "learn," because the nature of LLM input is radically different from human input. While many researchers have pointed out that the notion of "meaning" in LLMs is either different (Piantadosi & Hill, 2022), vague (Mitchell & Krakauer, 2023), or non-existent (Bender & Koller, 2020) – we extend this to emphasize that the *forms* LLMs learn from are also fundamentally different. In both usage-based and nativist accounts, input plays a crucial role: usage-based accounts suggest that input, while limited, is sufficiently rich in itself to explain the development of competence via domain general mechanisms. Nativist accounts, on the other hand, argue that input is impoverished in certain aspects, such that "innate" language-specific cognitive capacities must play some role. Regardless of a researcher's theoretical commitments in this debate, the nature of language *input* in humans plays a key role. The fact that LLM input is fundamentally different from human input precludes any useful insights they can provide in terms of human language development.

LLMs are trained on vast quantities of text alongside "fine-tuning" and further *reinforcement learning from human feedback* (RLHF). Child language learners are famously insensitive to the rare instances of explicit feedback they receive in input (Braine, 1971; Brown & Hanlon, 1970; Marcus, 1993). In contrast, the closest analogue for LLMs, RHLF, involves using explicit human rankings or annotations to feed back into model training, and is essential particularly for the impressive performance of more recent LLMs (Lambert et al., 2022).

Moreover, children become competent language users as part of normal development several years before they learn to successfully interact with text (if they ever do; see 3.3). Children acquire a breadth of linguistic competence (discussed in further detail in section 3.1) with access to only a fraction of the input LLMs require for syntactically acceptable performance[2] - indeed, the 3-year child's range of linguistic competence outstrips that of many "competent" LLMs (e.g. interpreting negation; Kalouli et al 2022; temporal and physical reasoning; Borji, 2023; see also Katzir, 2023), despite receiving a quantity of input that is four to five orders of magnitude smaller than what LLMs require (Frank, 2023).

Ongoing research is addressing these fundamental mismatches in input, for example, efforts such as BabyLM (Warstadt et al., 2023). and BabyBERTa (Huebner et al., 2021) attempt to train LLMs on more developmentally plausible corpora[3], and so may provide more direct insights into how humans learn language. Huebner et al (2021) exposed an LLM based on RoBERTa (Liu et al, 2019), nicknamed BabyBERTa, to a more ecologically valid input dataset (ten iterations of the same 5 million words of child directed speech, which they claim is roughly equivalent to the language experience of an English-acquiring 6-year-old). While BabyBERTa performed favorably relative to RoBERTa baselines on aspects of morphosyntax such as agreement and argument structure, its performance was greatly reduced on NPI licensing, island effects and superlative quantifiers (Huebner et al 2021: 629). Both BabyBERTa and RoBERTa also perform generally less accurately on question structures than most 2-year-olds (Newport, Gleitman and Gleitman 1977; e.g. subject-auxiliary inversion). While efforts to use

[2] We refrain from making specific token estimates of child language input here for two reasons. First, while we can have a reasonably accurate token count for the training set used for an LLM, specific token estimates of child language input rely on scaling up counts from limited temporal and demographic samples, and are unlikely to meaningfully reflect reality. Second, studies which attempt to estimate token counts in child language often have an explicit aim of making comparisons across socio-economic strata, perpetuating the harmful "word-gap narrative" (Figueroa, 2022). Nonetheless, the point stands that children much younger than 10 are proficient users of complex language, and have encountered far fewer tokens at an earlier age than most LLMs receive; in general, the fact that LLM training sets are much larger in terms of their number of word tokens than natural child language input is not disputed.

[3] Developmentally plausible both in terms of size and in terms of being derived from databases of transcribed child directed speech like those in CHILDES (MacWhinney, 2000).

more ecologically valid input address some of the fundamental developmental dissimilarities between humans and LLMs, *large* language models such as GPT-4, Bard, and BERT still require vastly larger quantities of ecologically implausible, text-based input. Irrespective of training set size, it is unclear whether their impressive performance is possible without the use of developmentally implausible RLHF.

Even in terms of proximate mechanisms, years of research show that capacity limitations in child memory and attention, as compared to adult learners (and certainly as compared to LLMs), are likely what *enable* remarkably successful language learning in children. Narrower windows of attention and more limited working memory capacity, for example, mean that learners benefit from "starting small", which in turn has been shown to play a key role in the processes of pattern detection and generalization (e.g., Arnon, 2021; Elman, 1993; Newport, 1990).

3. The limits of functional similarity

In the previous section, we established how LLMs and human language are fundamentally different in terms of their phylogeny, development and mechanisms. Consequently, LLMs have little potential in terms of pushing forward our understanding of how human language works in the brain, how children learn language, or how language evolved. Here, we address additional limits of functional similarity between human language and LLMs. In short, LLMs cannot do much of what human language users can do. Many experts have noted functional limitations of LLMs in terms of meaning (sometimes framed in terms of understanding or knowledge; Mitchell, 2019; Bender & Koller, 2020; Mitchell & Krakauer, 2023). Here, we take a slightly different focus, and argue that the linguistic *forms* LLMs require – written text – place fundamental limits on their general functional similarities with human language.

3.1 Children do more with less

Infants are remarkably sophisticated statistical learners (see e.g., the groundbreaking work of Saffran et al., 1996; also Schuler et al., 2021) and children's language learning is demonstrably shaped by their particular input (e.g., Ambridge et al., 2015). At around 3 and 4 years of age, they can track long-distance dependencies to maintain both pronominal reference (Karmiloff-Smith, 1985; Serratrice, 2005) and wh-reference (Hyams & Sigurjónsdóttir 1990; Thornton, 1990) - an aspect of Turing-type tasks that some LLMs have only recently achieved with uneven performance. Moreover, children also respect hierarchical structure (Crain & Nakayama, 1986), use and comprehend sentential negation (Pea, 1978; Dimroth, 2010), and represent natural-language interpretations of quantifiers such as "some" and "all", even if they differ in quality from adult usage (e.g., Chierchia et al., 2001). The two latter features of child - indeed human - language, are beyond the capabilities of some encoder-only LLMs (e.g. BERT, ALBERT, RoBERTa, Kalouli et al., 2022), with work still to be done on decoder-only LLMs (e.g. GPT).

The language capacities - including phonological, morphological, syntactic, semantic, and pragmatic capacities - of even a 3-year-old child are remarkably mature (Ambridge & Lieven, 2011; Lust, 2006; Rowland, 2014; Valian, 1986). Models trained on child-directed speech increasingly demonstrate a core role for pragmatics in developing adult-like grammars, indicating that children cross-check their pragmatic and syntactic hypotheses (Yang, 2022). The speech or sign signal received by human language learners and users itself contains rich information that is not straightforwardly encoded in text, including prosody and gesture (e.g., Crystal, 1973; Speer & Ito, 2009). Information in these richer signals plays an active and early role in how children differentiate morphosyntactic structures (Geffen & Mintz, 2015; Casillas & Frank, 2017). What children – and thus, humans - do with language has very limited functional similarity to what LLMs do. In the task of learning language, children are not just acquiring the

ability to generate grammatical strings, but also a wide array of functional capacities that LLMs lack, and that are not explicitly under development given that the functional remit of these models is confined to generating plausibly human *writing*.

3.2 Beyond broadcast transmission

Language is not just the production of morpho syntactically well-formed and semantically sensible utterances for broadcast transmission. The context of natural language learning is much richer and more complex than "language" in the context of LLMs. There are key elements of linguistic interaction, language learning, and cognition that are not available to or emergent in LLMs. For example, turn taking (e.g., Casillas et al., 2016; Levinson, 2016; Stivers et al., 2009), co-speech gesture and multimodality (e.g., Goldin-Meadow & Brentari, 2017; Kita et al., 2007; Rasenberg et al., 2022), repair (e.g., Dingemanse et al., 2014, 2015; Hayashi et al., 2013), and common ground (e.g., Brennan & Clark, 1996; Brown-Schmidt & Duff, 2016; Clark & Wilkes-Gibbs, 1986) are all essential parts of natural language in interaction. This also extends to other aspects of our broader socio-cognitive suite including ostensive inference, perspective taking and joint attention (e.g., Heintz & Scott-Phillips, 2023; Tomasello et al., 2005).

In terms of modality, some success has been shown with "multimodal" models that combine images and text (S. Huang et al., 2023; Bubeck et al., 2023), but their input is still narrow and individualized, and their output is a broadcast transmission rather than a product of interaction. Overall, this sense of "multimodality" is impoverished compared to human experience, and at most merely polysemous with the kind of multimodality researchers of human language consider increasingly essential for language (Kita et al., 2007; Rasenberg et al., 2022)[4]. Crucially, many of these key interactional aspects of linguistic form (and meaning)

[4] Bisk et al., (2020) provide a concrete way of thinking about increasing multimodal complexity in models, however, while they introduce multimodality at level WS3 (Perception), the applications of this are largely confined to integrations of images or other visual models with text. Researchers in language and

are emergent from interactions *between* language users, not necessarily confined properties of individual learners (Dingemanse et al., 2023).

In contrast with LLMs, human learners have a rich swathe of data available to them when tackling the task of learning and using language that we have barely even begun to identify, let alone quantify. Given this, it is perhaps unsurprising that, when considered in terms of tokens alone, the scale of training data LLMs require dwarfs the number of tokens human learners encounter. In other words, it may be that the total input human learners receive across their much richer experience is equivalent, in information theoretic terms, to the vast training sets of text required by LLMs (see Frank, 2023 for additional discussion of some ways human language input is richer than that of LLMs). However, the starting point to testing such a hypothesis is understanding and emphasizing the *differences* between humans' input and LLM input: identifying and quantifying the range of rich input humans receive, pushing for radical transparency regarding the nature of training data and representations in LLMs (Gebru et al., 2023; Bender et al., 2021), *and then* making **structured** comparisons between input sets. In short, arguing for any equivalence between humans and LLMs is premature without first making key strides in better understanding how human language learning works *and* achieving full transparency surrounding the training sets and inner-workings of often-proprietary LLMs.

3.3 Beyond written forms

LLMs reliance on text input excludes any language without a written form. In the first instance, this excludes an entire modality of natural human language: sign languages. Since the mid-20th century, the study of sign languages has profoundly deepened our understanding of human language by disentangling our complex communication system (and its cognitive underpinnings) from speech. However, no sign language has a widely used written form. Deaf communities

cognition more broadly are likely to identify WS4 (Embodiment and Action) and even elements of WS5 (The Social World) as being part of “multimodality” in language.

around the world fight for recognition and access to signed languages every day, in particular to prevent the dire (and common) consequences of childhood language deprivation (Hall et al., 2019). Encouraging an even narrower concept of language than the historical focus on spoken languages, the designers of LLMs equate language with text, fostering this misconception in the lay public, and further endangering the language access rights of deaf signers. In using LLMs as if they are revelatory models of how human language works (rather than technology designed to imitate the specific function of human writing), cognitive scientists risk also encouraging and perpetuating this misconception.

Natural language without a written form is not unique to sign languages: language in face-to-face interaction emerged, at minimum, hundreds of thousands of years before writing systems (Lock & Gers, 2012). This point also stands for contemporary languages: of the 7,168 living languages listed on Ethnologue, only a little over half (4,178) use a writing system (Eberhard et al., 2023), which in many cases was borrowed or adapted following colonization rather than being designed for the language in question (e.g., the use of the Roman alphabet for Swahili). Even for languages and cultures with bespoke, established writing systems going back hundreds or thousands of years, widespread literacy is a phenomenon that emerged in most populations only in the last century (Roser & Ortiz-Ospina, 2016). Natural languages, spoken or signed, emerge spontaneously in communities of users; in contrast, writing systems must be intentionally invented (or adapted), taught, and learned. In short, reading and writing (and thus, text) are themselves language technologies: writing is a (sometimes lossy) model of much more complex linguistic behavior (Lock & Gers, 2012). While writing has been around much longer than LLMs, it is nonetheless only recently in widespread use in the longer context of human history.

Of the 4,178 spoken languages with a writing system, LLMs still only consider a fraction of these, with the largest, BLOOM, covering 46 languages (Scao et al., 2022). Even these multilingual models exhibit uneven performance across languages: performance seems to scale

with the size and quality of a training set, giving these models considerably superior performance in English relative to other languages (see H. Huang et al., 2023, for a brief review). Over-estimating the relevance of LLMs to understanding human language and cognition risks further amplifying harmful existing biases towards English in cognitive science (Blasi et al., 2022). That some LLMs use other languages is only a patina of diversity; even where these show more or less equivalent performance to large models of English, we are still placing particular focus on only the written form of only (some of) the languages that happen to be written. In short, substantial equity and diversity issues arise if we narrowly define language as only those languages that are written, even if this is merely implicit in the argument that LLMs provide fundamental insights about human language more broadly. These issues not only affect the ethics of our scientific practice (a fundamental problem for AI that extends well beyond this; Bender et al., 2021; Birhane & van Dijk, 2020; Erscoi et al., 2023; Rillig et al., 2023), but risk leading us to only attempting to understand a confined subset of human linguistic cognition.

3.4 Language is more than text

One might argue that given the success of LLMs thus far (albeit on a small sample of written languages), they are likely to be able to deal with any form of human language. First, this argument rests on the *assumption* that a biased, English-dominant sample of languages is representative of human language (Blasi et al., 2022) - this is the very assumption we intend to challenge in arguing that LLMs are currently unsuitable for generating insights into human language and cognition. Second, there are well-documented structural (including syntactic) differences even between transcribed naturalistic speech and written text (e.g., Biber, 1998). BabyBERTa (Huebner et al, 2021) already demonstrates that a model trained exclusively on transcribed child-directed speech does not perform as well on certain syntactic structures as LLMs trained on much larger datasets of more structurally complex and explicit written language. Moreover, these smaller models focussing on structured comparisons with

ecologically valid child input have not demonstrated mastery of the kinds of open-ended text generation tasks that have so impressed many cognitive scientists (Contreras Kallens et al., 2022; Piantadosi, 2023; Frank, 2023). In part because human-like performance *given human-like input* is not the design aim of most LLMs, insights on this front are currently limited.

Finally, it is unclear how we could test an assumption that LLMs can probably deal with human language in general terms, precisely because of their fundamental reliance on text. Even advanced automatic speech recognition models (e.g., Open AI's Whisper) rely on mapping written transcripts to audio files, and consistent transcript quality has been identified as a key factor in model performance (Radford et al., 2022). Like LLMs, automatic speech-to-text transcription relies on using an existing writing system and training using large language-specific datasets. This means that the prospect of automatically generating usable text-based training data for low-resource languages is unlikely *even if they are written*, and virtually impossible if they are not. The scale of effort required to convert hundreds of thousands of hours of audio from an unwritten, low-resource language into something like the International Phonetic Alphabet (IPA) is infeasible, to say nothing of the issues inherent to collecting this much data from a low-resource language in the first place. These issues are compounded for sign languages, where no convention for consistently transcribing these languages (e.g., like the IPA for spoken languages) exists, even if we had the hundreds of thousands of hours of video data necessary as a starting point. Overall, if we were able to direct intense efforts towards collecting vast amounts of data for low-resource languages, it's not clear why these efforts would be directed toward building LLMs (instead of e.g., more deliberate and detailed efforts at language documentation, see Skirgård et al., 2023; as well as revitalisation, heritage and language justice efforts).

These substantial concerns about the representativeness of LLMs may not be especially relevant from an engineering perspective: the primary objective of much NLP research is to replicate specific functional aspects of some human languages, in order to create technology

accessible to the widest possible user base. From this perspective, the English-dominance of LLMs is a rational choice, and many may not consider the fact that LLMs cannot deal with certain kinds of languages to be a problem. Multilingual models that can perform competently in tens of languages are rightly considered an engineering feat, particularly considering the rapid pace of progress in this area. But, no matter how impressive form an engineering perspective, these models were not designed to shed light on phylogenetic, mechanistic, or developmental questions in human language and cognition. In short, it might be argued that a narrow focus on some written languages isn't a problem *for LLMs* (which is still not an uncontroversial claim, given uneven performance across languages). However, it is an acute problem for cognitive scientists attempting to use LLMs as representative models of human language or cognition.

4. Conclusions

In summary, LLMs are not designed to provide particularly strong or weak support for or against any particular theory of human linguistic cognition (nor do they incidentally provide insights in this domain). LLMs are designed to have narrow functional similarity to *written* language: they can learn to generate syntactically well-formed text in some languages. However, their reliance on (and confined functional competence in) text, the necessity of massive training sets, and need for explicit feedback in the form of RLHF, mean they are fundamentally different in all other respects. Their performance (or lack thereof) cannot contribute meaningfully to the debates about the extent to which human language learning is domain general or domain specific, or whether language involves neurological structures that are "innate" or merely develop robustly given adequate input. The forms that LLMs learn from and produce are fundamentally divorced from the vast array of behaviors, and broad base of cognition, that are tightly tied to human language. Any potential LLMs might have to push our understanding of human language and cognition forward substantially is unlikely to be realized without a more comprehensive understanding of the data humans use in language learning,

radical transparency surrounding the training sets and architecture of LLMs, and serious consideration of language diversity.

References

Ambridge, B., Kidd, E., Rowland, C. F., & Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition*. *Journal of Child Language*, *42*(2), 239–273. https://doi.org/10.1017/S030500091400049X

Ambridge, B., & Lieven, Elena. (2011). *Child language acquisition: Contrasting theoretical approaches*. Cambridge University Press.

Arnon, I. (2021). The Starting Big approach to language learning. *Journal of Child Language*, *48*(5), 937–958. https://doi.org/10.1017/S0305000921000386

Bateson, P., & Laland, K. N. (2013). Tinbergen's four questions: an appreciation and an update. *Trends in ecology & evolution*, *28*(12), 712-718. https://doi.org/10.1016/j.tree.2013.09.013

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜 *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Bender, E. M., & Koller, A. (2020, July). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185-5198). https://aclanthology.org/2020.acl-main.463/

Biber, D. (1998). *Variation across speech and writing*. Cambridge University Press, Cambridge, UK.

Birhane, A., & van Dijk, J. (2020). Robot Rights? Let's Talk about Human Welfare Instead. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 207–213.

https://doi.org/10.1145/3375627.3375855

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., ... & Turian, J. (2020). Experience grounds language. *arXiv preprint arXiv:2004.10151*. https://doi.org/10.48550/arXiv.2004.10151

Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., & Majid, A. (2022). Over-reliance on English hinders cognitive science. Trends in Cognitive Sciences, 26(12), 1153–1170. https://doi.org/10.1016/j.tics.2022.09.015

Borji, A. (2023). *A Categorical Archive of ChatGPT Failures* (arXiv:2302.03494). arXiv. https://doi.org/10.48550/arXiv.2302.03494

Braine, M. (1971). On two types of models of the internalization of grammars. In D. Slobin (Ed.), *The ontogenesis of grammar: A theoretical symposium* (pp. 153–168). Academic Press.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1482–1493. https://doi.org/10.1037/0278-7393.22.6.1482

Brown, R., & Hanlon, C. (1970). Derivational complexity and order of acquisition on child speech. In J. Hayes (Ed.), *Cognition and the development of language* (pp. 11–53). Wiley.

Brown-Schmidt, S., & Duff, M. C. (2016). Memory and Common Ground Processes in Language Use. *Topics in Cognitive Science*, *8*(4), 722–736. https://doi.org/10.1111/tops.12224

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv preprint arXiv:2303.12712.

Casillas, M., Bobb, S. C., & Clark, E. V. (2016). Turn-taking, timing, and planning in early language acquisition. *Journal of Child Language*, *43*(6), 1310–1337.

https://doi.org/10.1017/S0305000915000689

Casillas, M., & Frank, M. C. (2017). The development of children's ability to track and predict turn structure in conversation. *Journal of memory and language*, *92*, 234-253. https://doi.org/10.1016/j.jml.2016.06.013

Chierchia, G., S. Crain, M. T. Guasti, A. Gualmini, and L. Meroni (2001) "The Acquisition of Disjunction: Evidence for a Grammatical View of Scalar Implicatures," in A. H.-J. Do et al., eds., BUCLD 25 Proceedings, Cascadilla Press, Somerville, Massachusetts

Chomsky, N., Roberts, I., & Watumull, J. (2023, March 8). Opinion: The false promise of ChatGPT. *New York Times*.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39. https://doi.org/10.1016/0010-0277(86)90010-7

Contreras Kallens, P., Kristensen-McLachlan, R. D., & Christiansen, M. H. (2023). Large Language Models Demonstrate the Potential of Statistical Learning in Language. *Cognitive Science*, *47*(3), e13256. https://doi.org/10.1111/cogs.13256

Crain, S., & Nakayama, M. (1987). Structure dependence in grammar formation. *Language*, 63(3), 522-543. https://doi.org/10.2307/415004

Crystal, D. (1973). Non-segmental phonology in language acquisition: A review of the issues. *Lingua*, *32*(1), 1–45. https://doi.org/10.1016/0024-3841(73)90002-8

Dimroth, C. (2010). The acquisition of negation. In L.R. Horn (Ed.), T*he Expression of Negation* (p. 39-72). De Gruyter Mouton.

Dingemanse, M., Blythe, J., & Dirksmeyer, T. (2014). Formats for other-initiation of repair across languages: An exercise in pragmatic typology. *Studies in Language. International Journal Sponsored by the Foundation "Foundations of Language"*, *38*(1), 5–43. https://doi.org/10.1075/sl.38.1.01din

Dingemanse, M., Roberts, S. G., Baranova, J., Blythe, J., Drew, P., Floyd, S., Gisladottir, R. S., Kendrick, K. H., Levinson, S. C., Manrique, E., Rossi, G., & Enfield, N. J. (2015).

Universal Principles in the Repair of Communication Problems. *PLOS ONE*, *10*(9), e0136100. https://doi.org/10.1371/journal.pone.0136100

Dingemanse, M., Liesenfeld, A., Rasenberg, M., Albert, S., Ameka, F. K., Birhane, A., ... & Wiltschko, M. (2023). Beyond Single-Mindedness: A Figure-Ground Reversal for the Cognitive Sciences. *Cognitive Science*, 47(1), e13230. **https://doi.org/10.1111/cogs.13230**

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (2023). *Ethnologue: Languages of the World* (26th ed.). SIL International. http://www.ethnologue.com

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*(1), 71–99. https://doi.org/10.1016/0010-0277(93)90058-4

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). *GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models* (arXiv:2303.10130). arXiv. https://doi.org/10.48550/arXiv.2303.10130

Erscoi, L., Kleinherenbrink, A., & Guest, O. (2023). *Pygmalion Displacement: When Humanising AI Dehumanises Women*. SocArXiv. https://doi.org/10.31235/osf.io/jqxb6

Figueroa, M. (2022). Podcasting past the paywall: How diverse media allows more equitable participation in linguistic science. *Annual Review of Applied Linguistics*, *42*, 40–46. https://doi.org/10.1017/S0267190521000118

Frank, M. (2023). Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2023.08.007

Gebru, T., Bender, M., McMillan-Major, A., Mitchell, M. (2023). Statement from the listed authors of Stochastic Parrots on the “AI pause” letter. DAIR Institute Blog, available at https://www.dair-institute.org/blog/letter-statement-March2023, accessed April 1 2023.

Geffen, S., & Mintz, T. H. (2015). Can you believe it? 12-month-olds use word order to distinguish between declaratives and polar interrogatives. *Language Learning and Development,* 11(3), 270-284.

Goldin-Meadow, S., & Brentari, D. (2017). Gesture, sign, and language: The coming of age of sign language and gesture studies. *Behavioral and Brain Sciences*, *40*, e46. https://doi.org/10.1017/S0140525X15001247

Hall, M. L., Hall, W. C., & Caselli, N. K. (2019). Deaf children need language, not (just) speech. *First Language*, *39*(4), 367–395. https://doi.org/10.1177/0142723719834102

Hayashi, M., Raymond, G., & Sidnell, J. (2013). *Conversational Repair and Human Understanding*. Cambridge University Press.

Heintz, C., & Scott-Phillips, T. (2023). Expression unleashed: The evolutionary and cognitive foundations of human communication. *Behavioral and Brain Sciences*, *46*, e1. https://doi.org/10.1017/S0140525X22000012

Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O. K., Patra, B., Liu, Q., Aggarwal, K., Chi, Z., Bjorck, J., Chaudhary, V., Som, S., Song, X., & Wei, F. (2023). Language Is Not All You Need: Aligning Perception with Language Models. arXiv preprint. https://doi.org/10.48550/ARXIV.2302.14045

Huang, H., Tang, T., Zhang, D., Zhao, W. X., Song, T., Xia, Y., & Wei, F. (2023). Not All Languages Are Created Equal in LLMs: Improving Multilingual Capability by Cross-Lingual-Thought Prompting. arXiv preprint. https://doi.org/10.48550/arXiv.2305.07004

Huebner, P. A., Sulem, E., Cynthia, F., & Roth, D. (2021, November). BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th conference on computational natural language learning* (pp. 624-646). http://dx.doi.org/10.18653/v1/2021.conll-1.49

Hyams, N., & Sigurjonsdottir, S. (1990). The development of “long-distance anaphora": A cross-linguistic comparison with special references to Icelandic. *Language Acquisition*, 1(1), 57-93.

Kalouli, A., Sevastjanova, R., Beck, C., & Romero, M. (2022). Negation, Coordination, and Quantifiers in Contextualized Language Models. In *Proceedings of the 29th International Conference on Computational Linguistics*, p. 3074–3085. https://aclanthology.org/2022.coling-1.272

Karmiloff-Smith, A. (1985). Language and cognitive processes from a developmental perspective. *Language and Cognitive Processes* 1(1), 61-85.

Katzir, R. (2023). Why large language models are poor theories of human linguistic cognition. A response to Piantadosi (2023). Lingbuzz preprint. https://lingbuzz.net/lingbuzz/007190

Kita, S., Özyürek, A., Allen, S., Brown, A., Furman, R., & Ishizuka, T. (2007). Relations between syntactic encoding and co-speech gestures: Implications for a model of speech and gesture production. *Language and Cognitive Processes*, *22*(8), 1212–1236. https://doi.org/10.1080/01690960701461426

Lambert, N., Castricato, L., Von Werra, L., & Havrilla, A. (2022). Illustrating Reinforcement Learning from Human Feedback (RLHF). *Hugging Face Blog*. https://huggingface.co/blog/rlhf

Levinson, S. C. (2016). Turn-taking in Human Communication – Origins and Implications for Language Processing. *Trends in Cognitive Sciences*, *20*(1), 6–14. https://doi.org/10.1016/j.tics.2015.10.010

Liu, Y., Ott , M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V.. 2019. RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint. https://doi.org/10.48550/arXiv.1907.11692.

Lock, A. & Gers, M. (2012). The cultural evolution of written language and its effects: A

Darwinian process from prehistory to modern day. In E. Grigorenko, E. Mambrino & D. Preiss (Eds.), *Writing: A mosaic of new perspectives* (p. 11-36). Psychology Press.

Lust, B. (2006). *Child language; Acquisition and growth*. Cambridge University Press.

MacWhinney, B. (2000). *The CHILDES project: The database* (Vol. 2). Psychology Press.

Mameli, M., & Bateson, P. (2011). An evaluation of the concept of innateness. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1563), 436–443. https://doi.org/10.1098/rstb.2010.0174

Marcus, G. F. (1993). Negative evidence in language acquisition. *Cognition*, *46*(1), 53–85.

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, *120*(13), e2215907120.

Mitchell, M. (2019). Artificial intelligence hits the barrier of meaning. *Information*, *10*(2), 51. https://doi.org/10.3390/info10020051

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science*, *14*(1), 11–28. https://doi.org/10.1016/0364-0213(90)90024-Q

Pan, Y., Pan, L., Chen, W., Nakov, P., Kan, M.-Y., & Wang, W. Y. (2023). *On the Risk of Misinformation Pollution with Large Language Models* (arXiv:2305.13661). arXiv. https://doi.org/10.48550/arXiv.2305.13661

Pea, R. (1978). *The development of negation in early child language* (Doctoral dissertation, University of Oxford).

Piantadosi, S. & Hill, F. (2022). Meaning without reference in large language models. *arXiv preprint.* https://doi.org/10.48550/arXiv.2208.02957

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

Pleyer, M., & Hartmann, S. (2019). Constructing a Consensus on Language Evolution?

Convergences and Differences Between Biolinguistic and Usage-Based Approaches. *Frontiers in Psychology*, *10*. https://www.frontiersin.org/articles/10.3389/fpsyg.2019.02537

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. *arXiv preprint.* https://doi.org/10.48550/arXiv.2212.04356

Rasenberg, M., Pouw, W., Özyürek, A., & Dingemanse, M. (2022). The multimodal nature of communicative efficiency in social interaction. *Scientific Reports*, *12*(1), 19111.

Rillig, M. C., Ågerstrand, M., Bi, M., Gould, K. A., & Sauerland, U. (2023). Risks and Benefits of Large Language Models for the Environment. *Environmental Science & Technology*, *57*(9), 3464–3466. https://doi.org/10.1021/acs.est.3c01106

Roser, M. & Ortiz-Ospina, E. (2016). Literacy. *Our World in Data*. Accessed online (August 23, 2023) at https://ourworldindata.org/literacy

Rowland, C. (2014). *Understanding child language acquisition*. Routledge.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science*, *274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., Tow, J., Rush, A. M., Biderman, S., Webson, A., Ammanamanchi, P. S., Wang, T., Sagot, B., Muennighoff, N., del Moral, A. V., … Wolf, T. (2022). *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model* (arXiv:2211.05100). arXiv. https://doi.org/10.48550/arXiv.2211.05100

Schuler, K., Yang, C., & Newport, E. (2021). *Testing the Tolerance Principle: Children form productive rules when it is more computationally efficient*. PsyArXiv. https://doi.org/10.31234/osf.io/utgds

Scott-Phillips, T. C., Dickins, T. E., & West, S. A. (2011). Evolutionary theory and the

ultimate–proximate distinction in the human behavioral sciences. *Perspectives on Psychological Science*, *6*(1), 38-47.

Serratrice, L.(2005). The role of discourse pragmatics in the acquisition of subjects in Italian. *Applied Psycholinguistics*. 26, 437-462. https://doi.org/10.1017.S0142716405050241

Skirgård, H., Haynie, H. J., Blasi, D. E., Hammarström, H., Collins, J., Latarche, J. J., ... & Gray, R. D. (2023). Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. Science Advances, 9(16), eadg6175. https://doi.org/10.1126/sciadv.adg6175

Speer, S. R., & Ito, K. (2009). Prosody in First Language Acquisition – Acquiring Intonation as a Tool to Organize Information in Conversation. *Language and Linguistics Compass*, *3*(1), 90–110. https://doi.org/10.1111/j.1749-818X.2008.00103.x

Spike, M. (2017). The evolution of linguistic rules. *Biology & Philosophy*, *32*(6), 887-904.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., de Ruiter, J. P., Yoon, K.-E., & Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. Proceedings of the National Academy of Sciences, 106(26), 10587–10592. https://doi.org/10.1073/pnas.0903616106

Thornton, R. (1990). *Adventures in long-distance moving: The acquisition of complex wh-questions*. PhD Thesis, University of Connecticut. https://opencommons.uconn.edu/dissertations/AAI9109855

Tinbergen, N. (1963). On aims and methods of ethology. *Zeitschrift für tierpsychologie*, *20*(4), 410-433.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, *28*(5), 675–691. https://doi.org/10.1017/S0140525X05000129

Valian, V. (1986). Syntactic categories in the speech of young children. *Developmental Psychology*, *22*, 562–579. https://doi.org/10.1037/0012-1649.22.4.562

Warstadt, A., Choshen, L., Mueller, A., Williams, A., Wilcox, E., & Zhuang, C. (2023). *Call for Papers -- The BabyLM Challenge: Sample-efficient pretraining on a developmentally plausible corpus* (arXiv:2301.11796). arXiv. https://doi.org/10.48550/arXiv.2301.11796

Yang, Y. (2022). *Are you asking me or telling me? Learning clause types and speech acts in English and Mandarin*. PhD Thesis, University of Maryland. https://doi.org/10.13016/sfhx-z3h2