CellPress

Trends in Cognitive Sciences

Forum

Bridging the data gap between children and large language models

Michael C. Frank[1,*]

Large language models (LLMs) show intriguing emergent behaviors, yet they receive around four or five orders of magnitude more language data than human children. What accounts for this vast difference in sample efficiency? Candidate explanations include children's pre-existing conceptual knowledge, their use of multimodal grounding, and the interactive, social nature of their input.

How much learning is needed for the emergence of intelligence? Some rule-based systems were designed to act intelligently in the absence of any adjustments based on training data. By contrast, modern LLMs exemplify the opposite strategy: they are fed with massive, internet-scale text datasets, and their performance typically grows in proportion to the available data and computation [1]. The resulting models are surprisingly competent at a wide range of tasks, although they still show systematic flaws in reasoning and information retrieval.

For many observers, the most interesting feature of LLMs is their ability to reason flexibly about new tasks based on a verbal query, synthesizing information in a text 'prompt' to generate, for example, an explanation, a poem, a piece of computer code, or a tabular dataset. This behavior, sometimes termed 'few shot' or 'in-context learning', appears to emerge only at very large scales of training data. Yet another type of intelligence performs in-context learning with far less data than even the smallest of LLMs: humans. From an early age, children can reason flexibly about novel tasks, and by middle childhood they can quickly master new games, devices, and environments. What can we learn about human and machine intelligence by comparing their data efficiency?

Measuring the gap

The scale of training data for current LLMs is unprecedented. Training datasets are typically measured in tokens, a metric that includes words but also punctuation and morphological subparts of words. GPT-3 was trained on 5 x 10^{11} tokens [2] and Chinchilla was trained on 10^{12} tokens [1]. Many companies keep training-set sizes secret, but a recent leak suggested that one industry model was trained on 3.6 x 10^{12} tokens. How do these numbers compare with human language experience? Comprehensive word counts are difficult to collect, but sampling and extrapolation can provide reasonable upper and lower bounds for language input (Figure 1).

A soft upper bound on a child's linguistic input – language produced by the people around them – is around 10^6 words per month [3,4]. For a 5-year-old, that would be 6 x 10^7 words; for a 20-year-old it would be 2 x 10^8 words. We also might assume that a 20-year-old has been reading for 10–15 years, and for much of this time they are reading two or three books (10^5 words each) per week for an extra ~10^7 words per year. Our rough upper bound for a literate 20-year-old could be as high as 4 x 10^8 words (or even higher if they read constantly).

By contrast, children growing up in environments with limited language are estimated to experience around 1 x 10^5 words per month, up to an order of magnitude fewer than children in richer linguistic environments [5]. Without the boost from literacy, a lower bound on language experience would be around 6 x 10^6 words by the age of 5 and 3 x 10^7 by age 20.

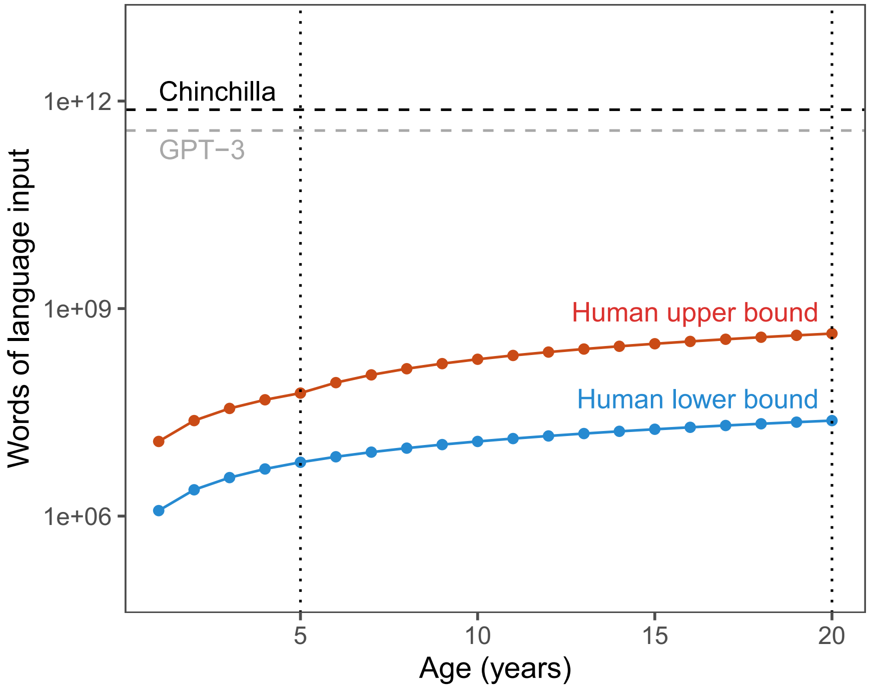
Importantly, even a child receiving much less language will still be able to reason about novel tasks: for example, learning the rules of a new board game at school. By contrast, language models trained on human-like amounts of data can at best provide incoherent 'autocomplete'-like behaviors, with no in-context learning. Thus, there is a difference of up to five orders of magnitude in language input between LLMs and human children, and at least three between LLMs and even the most literate adults. What factors explain the far greater efficiency of human learners?

Minding the gap

Let us consider three potential – not mutually exclusive – explanations for human sample efficiency.

The first explanation is that an immense evolutionary history has shaped human minds and brains prior to their initial contacts with data. Some researchers posit that infants have innately specified 'core knowledge' of objects, agents, and events, comprising the foundations of a conceptual model of the world [6]. Other developmental theories posit architectural constraints on learning that lead to the quick emergence of fundamental knowledge structures [7]. In either case, initial constraints – perhaps expressed via specific patterns of brain connectivity – could provide a major speedup in how much experience an agent needs to bootstrap further reasoning.

The second explanation is the richness of the grounded, sensory experience available to human learners. Children's experiences contain a profusion of auditory, visual, haptic, gustatory, olfactory, and somatosensory data. In constructivist



Trends in Cognitive Sciences

Figure 1. Graph of language input to humans versus large language models (LLMs). A gap of around three to five orders of magnitude exists between estimated human language inputs (red and blue lines) and the inputs to LLMs (dashed lines).

proposals, these data allow children to create and refine theoretical models of the world [8]. Multimodal data also 'ground' language, providing concrete extensional meanings for many words. In contrast, LLMs must induce world knowledge from a single stream of information that primarily contains language – sometimes alongside a complex mélange of computer code and other types of information – rather than connecting linguistic information to external experiences.

The final potential explanation is the type of language input that humans receive, which for children is often generated through structured social interactions in which the child plays a part [9]. Some of this input is simplified by adults, with limited vocabulary and lower sentence complexity. Such interactional input differs dramatically from the training data provided to LLMs, which make predictions about vast amounts of text from decontextualized sources and with no opportunity to interact or intervene. One observation supporting this hypothesis is that newer LLMs are trained via reinforcement learning using human feedback; it is likely that this 'interaction training' is responsible for the success of products like ChatGPT in responding appropriately in conversation.

Crossing the gap

Beyond these three substantive factors, we should consider the possibility that much of the apparent difference between LLMs and human learners is due to differences in evaluation. LLMs are often evaluated on complex reasoning tasks, while tasks for children are typically simple and highly scaffolded. Understanding the gap between LLMs and children will require synchronized evaluation (as well as unambiguous positive evidence that models are truly passing the evaluations, rather than memorizing test data). Tasks like MEWL, a battery of word learning tasks [10], or the Baby Intuitions Benchmark, a set of social cognition tasks [11], attempt to create apples-to-apples comparisons. However, most developmental experiments are not conducted in a unimodal, language-only format, so comparison between children and LLMs can be challenging.

Even when we have evaluated models and humans more comparably, it is likely that a gap in input will persist. After all, human adults – who pass most LLM evaluations [2] – still have not experienced anywhere near the amount of language that the models have. Testing hypotheses about this gap will require more work on sample efficiency, which has not been a priority given that the artificial intelligence community has embraced scale as the route to better performance [1].

Still, several recent efforts are promising. The BabyLM challenge (https://babylm.github.io) asks entrants to train models on curated 10^7- and 10^8-word datasets, providing a framework for comparing different LLM architectures on human-scale datasets. Holding data constant is an effective research design for understanding how learning architectures affect outcomes, though one worry is that without high-quality training data – for example, coherent, interactive dialogue about the here-and-now – even the best architecture might fail.

Unfortunately, our best resource for transcripts of grounded, interactive language use, the Child Language Data Exchange (CHILDES), is too small to train an LLM. Extant multimodal datasets are even smaller, and still only include visual and linguistic data. One creative workaround is to generate data automatically: the TinyStories corpus is an LLM-generated corpus of child-appropriate stories with a restricted vocabulary [12], and small-scale LLMs trained on this corpus show surprising competence. Applying this approach to the generation of multimodal data could be a promising direction for creating more developmentally appropriate training corpora.

As LLMs grow ever larger, there is a real risk that developers will run out of high-quality training data. An alternative path is to figure out how to make better use of the available data, increasing efficiency by pursuing learning strategies that are better aligned with human learning. Doing so will require a better understanding of the gaps between human learners and current models, however. Bridging these gaps may prove rewarding for our understanding of human as well as machine intelligence.

Acknowledgments

Thanks to Noah D. Goodman, Christopher Potts, and Alvin Tan for helpful discussion.

Declaration of interests

No interests are declared.

[1]Department of Psychology, Stanford University, 450 Jane Stanford Way, Stanford, CA 94305, USA

*Correspondence:
mcfrank@stanford.edu (M.C. Frank).

https://doi.org/10.1016/j.tics.2023.08.007

References

1. Hoffmann, J. *et al.* (2022) Training compute-optimal large language models. *arXiv* Published online March 29, 2022. http://doi.org/10.48550/arXiv.2203.15556
2. Brown, T.B. *et al.* (2020) Language models are few-shot learners. *arXiv* Published online July 22, 2020. http://doi.org/10.48550/arXiv.2005.14165
3. Roy, B.C. *et al.* (2015) Predicting the birth of a spoken word. *Proc. Natl. Acad. Sci.* 112, 12663–12668
4. Dupoux, E. (2018) Cognitive science in the era of artificial intelligence: a roadmap for reverse-engineering the infant language-learner. *Cognition* 173, 43–59
5. Bergelson, E. *et al.* (2019) What do North American babies hear? A large-scale cross-corpus analysis. *Dev. Sci.* 22, e12724
6. Spelke, E.S. and Kinzler, K.D. (2007) Core knowledge. *Dev. Sci.* 10, 89–96
7. Tenenbaum, J.B. *et al.* (2011) How to grow a mind: statistics, structure, and abstraction. *Science* 331, 1279–1285
8. Gopnik, A. and Wellman, H.M. (2012) Reconstructing constructivism: causal models, Bayesian learning mechanisms, and the theory theory. *Psychol. Bull.* 138, 1085–1108
9. Clark, E.V. (2016) *First Language Acquisition* (3rd edn), Cambridge University Press
10. Jiang, G. *et al.* (2023) MEWL: few-shot multimodal word learning with referential uncertainty. *arXiv* Published online June 1, 2023. http://doi.org/10.48550/arXiv.2306.00503
11. Stojnić, G. *et al.* (2023) Commonsense psychology in human infants and machines. *Cognition* 235, 105406
12. Eldan, R. and Li, Y. (2023) TinyStories: how small can language models be and still speak coherent English? *arXiv* Published online May 24, 2023. http://doi.org/10.48550/arXiv.2305.07759