Baby steps in evaluating the capacities of large language models

Michael C. Frank[1†]
[1]Department of Psychology, Stanford University, Stanford, California, USA
[†]email: mcfrank@stanford.edu

Large language models show remarkable capacities, but it is unclear what abstractions support their behavior. Methods from developmental psychology can help researchers understand the representations used by these models, complementing standard computational approaches—and perhaps leading to insights about the nature of mind.

Imagine first contact with an alien intelligence. A scientist might ask, do the aliens have the same concepts as humans? Do they understand other minds? Can they reason about cause and effect? Such scenarios are common in science fiction—and in the past few years, in interactions with large language models (LLMs). Yet developmental psychologists have been asking such questions for years about another kind of alien intelligence: human children. Methods from this research can help researchers probe the capacities of LLMs.

Language modeling has been used for decades as a technique for predicting the next word in a sequence. Such models were widely used as simple baselines for human sequence learning, but it was taken for granted that they would not display other abilities. Yet in the past few years, predictive models trained on massive datasets have begun to show a host of interesting behaviors in complex task contexts [1]. LLMs make predictions that are grammatical and semantically coherent, and they produce satisfactory—and sometimes even delightful—results when asked, for example, to make analogies, summarize text, or compose poetry.

These results have led both lay users and researchers to speculate about what underlies LLMs' seemingly intelligent behaviors – whether they possess human-like cognitive abstractions or whether these behaviors result from simple word prediction (albeit at massive scale). For example, LLMs can correctly answer questions about the beliefs of a character in a story, even when those beliefs are false. This finding could be taken as evidence that LLMs have acquired the set of abstractions about human internal states that together are referred to as "theory of mind". Alternatively, it could be that LLMs are showing highly fluent responses to superficial linguistic cues [2].

Such debates are consequential because abstract representations, which enable flexible and adaptive behavior across a wide range of contexts, are a key feature of the cognition of mature humans. Their presence in an LLM provides a proof of concept that such abstractions can be learned from data rather than being innately specified [3]. Yet creating behavioral tests that conclusively demonstrate the presence of a particular representation is challenging, whether the test is being given to an LLM or to a child. The trouble is that many test questions can be answered via multiple strategies.

In these cases, comparative and developmental psychologists are generally guided by a principle known as Morgan's Canon [4]: don't jump to the conclusion that a system has a high-level abstraction when a lower-level capacity would suffice to explain its behavior. Yet beyond this general principle, developmental researchers have converged around a set of empirical strategies for making claims about the presence of abstract representations. Many of these can be applied directly to LLMs, although LLMs also allow approaches that are impossible (or unethical) to implement with human learners.

The developmentalist's toolkit

First and foremost, generalization to novel situations is critical for making claims about abstraction. If a child has seen a particular stimulus item before, they might produce a learned response reflecting prior experience, rather than reasoning based on an underlying abstract representation. Consequently, many developmental studies rely on teaching novel words like "dax" or showing children novel objects like "blicket detectors." Using words or objects that children could not have encountered outside the laboratory removes the possibility that they are relying on a previously learned stimulus-response mapping to complete the task. Because the largest LLMs have been trained on hundreds or thousands of scientific papers containing examples of evaluations in both machine learning and psychology, standard experimental prompts are probably useless in evaluating such models. Model responses could reflect experience with the stimuli—for example, providing sample answers from a research paper—rather than generalization. When designing a new study for preschoolers, researchers often break out the markers and glue to create novel stimuli; scientists will need to be equally creative in designing new tasks for LLM evaluation.

Developmentalists also often avoid using rich, naturalistic stimuli. This practice might seem surprising: stimuli that resemble infants' day-to-day experience should be easier for them to process than more schematic stimuli. However, for babies—and for LLMs—the richer a stimulus is, the more routes there are to a lower-level solution [5]. For example, to test whether babies know that greater effort indicates a more valuable goal, researchers chose to show simple geometric figures "jumping" over barriers of different heights. Creating videos with human actors would have been more ecologically valid, but it would also have confounded the effort involved in an action with other visual and social cues such as facial expression and body posture [6]. Simplified stimuli are a challenge for LLMs and other models because they are typically outside of the model's training distribution—but for that reason they provide a strong test of the model's underlying abstractions.

Even for highly simplified stimuli, some superficial stimulus features are often still confounded with the manipulation of interest. Thus, the trick of a truly clever experimental design is to hold every aspect of the probe stimulus constant across conditions, while making a single key modification that changes the observer's interpretation. Classic language learning experiments demonstrate this design by using the same probe stimulus (for example, the novel word 'golatu') but creating learning environments in which the statistics of its use differ (for example, one where the syllables 'go', 'la', and 'bu' follow each other consistently vs. one where they are heard together only via the conjunction of two other words, ‘pigola’ and ‘tudaro’) [7]. This kind of design ensures that prior associations do not bias the result; without a closely matched control condition, an incidental preference for the word ‘golatu’ might lead to the appearance of success even in the absence of learning. In the case of LLMs, such matched controls are especially critical because models encode a massive set of prior associations that could bias their responses.

Finally, providing evidence for a cognitive abstraction typically requires converging evidence across multiple experimental tasks and across development. For example, children typically learn to say number words by age three. Yet when asked to give an experimenter a quantity like seven, even children who can count to ten will often provide a large uncounted pile of objects. And even after they can correctly give seven objects, they might not always understand that seven is smaller than eight. Each of these tasks (reciting the count list, enumerating a set, and comparing magnitudes) probes a different aspect of a child’s number concepts, providing clues about the underlying representation and allowing for developmental dissociations [8]. Thus, like children, LLMs need to be tested on multiple tasks and measures of the same conceptual abstraction. Ideally, these tests should be conducted multiple times over the course of model training to identify how performance changes as the model acquires more experience.

The computational scientist’s toolkit

Because LLMs are computational artifacts, scientists can also investigate their capacities using tools that are typically not available to psychologists. These methods complement carefully designed behavioral tasks, allowing insight into the mechanisms by which LLMs succeed.

First, researchers can manipulate LLM training data to investigate what inputs lead to the emergence of a target behavior, a practice known as “controlled rearing” in the animal cognition literature. As an example, experience with mental state language (verbs like ‘think’ and ‘believe’) likely plays a role in the emergence of theory of mind [9], which raises the question of whether removing such input from training corpora would change model behavior.

Second, the internal representations used by LLMs can be directly accessed, unlike in children where existing neuroscience methods provide only noisy and indirect measures of brain states. Thus, researchers can probe and intervene on the model’s internal representations to measure how specific behaviors relate to model states. Although LLMs are sometimes referred to as ‘black boxes,’ decoding and intervention methods for neural networks are advancing rapidly. One inspiring line of work used causal interventions on models’ internal representations—literally changing the values of “neurons” in the neural network—to establish equivalence between the neural network’s representations and a higher-level symbolic representation, providing insight into exactly what representations supported the model’s behavior [10].

However, both controlled rearing and probing methods require full access to LLM training and parameters. This limitation highlights an important issue for the research community: the most sophisticated and intriguing models from a cognitive science perspective—currently GPT-type models [1] —are often accessible only via limited commercial interfaces.

Synthesizing insights from human and machine learning

Careful investigation of LLMs might reveal more about their capacities, but it might also lead to insights about the nature of learning more broadly. As strong statistical learners, LLMs provide a valuable proof of concept of how abstractions can—or cannot—emerge purely from data-driven learning. Current LLM architectures start to show the ability to perform arbitrary new tasks at the level of hundreds of billions of words of training data, whereas humans need many orders of magnitude fewer. Investigating this gap in efficiency might help scientists triangulate what makes humans so efficient. The more we synchronize progress in artificial intelligence with what we know about human development, the more we will learn about each.

References

[1] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

[2] Ullman, T. (2023). Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks. arXiv preprint arXiv:2302.08399.

[3] Geiger, A., Carstensen, A., Frank, M. C., & Potts, C. (2023). Relational reasoning and generalization using nonsymbolic neural networks. Psychological Review, 130(2), 308.

[4] Sober, E. (1998). Morgan's canon. In D. D. Cummins & C. Allen (Eds.), The evolution of mind (pp. 224–242). Oxford University Press.

[5] Kominsky, J. F., Lucca, K., Thomas, A. J., Frank, M. C., & Hamlin, J. K. (2022). Simplicity and validity in infant research. Cognitive Development, 63, 101213.

[6] Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. Science, 358(6366), 1038- 1041. doi:10.1126/science.aag2132

[7] Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. Science, 274(5294), 1926-1928.

[8] Davidson, K., Eng, K., & Barner, D. (2012). Does learning to count involve a semantic induction? Cognition, 123(1), 162-173.

[9] Ruffman, T., Slade, L., & Crowe, E. (2002). The relation between children's and mothers' mental state language and theory-of-mind understanding. Child Development, 73(3), 734-751.

[10] Geiger, A., Lu, H., Icard, T., & Potts, C. (2021). Causal abstractions of neural networks. Advances in Neural Information Processing Systems, 34, 9574-9586.

Competing interests

The authors declare no competing interests.