Use of symbolic inductive biases for few-shot learning in humans and machines

Vijay Marupudi (vijaymarupudi@berkeley.edu) & Steven T. Piantadosi (stp@berkeley.edu)
Department of Psychology, UC Berkeley

Abstract

People are capable of learning with very little data. We argue that this is because people possess an symbolic inductive bias, consistent with the language of thought (LoT) hypothesis (Fodor & Pylyshyn, 1988). To evaluate this claim, participants were asked to learn list functions by predicting how to transform an input list of numbers to an output list. We used participants' predictions on trials with no feedback to search for a congruent representation using a LoT model. The model was able to predict participants' heldout responses at above chance levels, even for participants who were unable to learn the function. Furthermore, LLMs and a neural network trained with a LoT inductive bias displayed similar patterns. These findings suggest that people and ML models display signatures of symbolic representation use to accomplish few-shot learning.

Keywords: neural networks; symbolic representations; concepts; few-shot learning; rule learning

Introduction

Understanding the nature of mental representations is a central goal of cognitive science. Decades of research has centered on whether our mental representations are more consistent with either symbolic rules or the representations learned by connectionist networks (McClelland & Patterson, 2002; Pinker & Prince, 1988; Pinker & Ullman, 2002; Rumelhart & McClelland, 1987). The recent success of deep neural networks at a wide array of tasks has renewed criticisms of symbolic models of cognition being inadequate and potentially unnecessary (Binz et al., 2025; Griffiths et al., 2025). While this debate was helpful for determining the behavioral signatures and advantages of the representational formats, we believe it is not productive to treat these representations as being exclusive. It is likely that symbolic theories are useful abstractions about non-symbolic representations and that connectionist networks learn symbolic representations (Fodor & Pylyshyn, 1988).

Determining the reasons why people adopt a representation format can help us understand their mental representations. Symbolic representations, in particular, present numerous advantages. Fodor (1975) argued that people's mental representations followed a symbolic language-like structure. In this representational system, called the *language of thought* (LoT), people represent concepts as a composition of primitive symbols. This flexible representation system is capable of systematicity, productivity, and compositionality, all signature characteristics of human behavior (Fodor & Pylyshyn, 1988). Goodman et al. (2008) formalized this notion into a Bayesian inference paradigm to learn compositional program-like representations from data, given a set of primitive symbols and operations. This probabilistic language of thought (pLOT) paradigm was successful at modeling how people learn Boolean concepts of varying complexity (Feldman, 2000). Critical to this approach was a prior that penalized complexity, echoing a similar bias found among people (Chater & Vitányi, 2003; Pothos & Chater, 2002). Work following a similar paradigm has successfully modeled human learning of categories (Piantadosi et al., 2016), number concepts (Piantadosi, 2023; Piantadosi et al., 2012), geometric shapes and patterns (Amalric et al., 2017; Sablé-Meyer et al., 2022) and probabilistic inference (Gerstenberg & Goodman, 2012).

We argue that people's ability to reason with symbolic LoT representations explains their few-shot learning ability, i.e. the ability to learn concepts quickly with very little data. Conventional deep neural networks notoriously require large amounts of data to perform similarly to humans. In contrast, human children, who have experienced far less data, are impressively competent at a multitude of tasks. For example, Carey and Bartlett (1978) found that children can learn a word with only a single example of its use. This may be due to an inductive bias towards learning LoT representations. These representations can be effective because they encompass a large conceptual space, encourage reuse, and are easily generalized (Rule et al., 2020). They allow people to draw upon their prior knowledge of learned operations and primitives over the course of their lifetime and therefore can reduce the amount of data necessary to learn certain concepts.

If people use LoT representations to learn, one would expect people to be able to learn complex symbolic concepts quickly with little data, as the structure of these concepts would match a symbolic inductive bias. Rule et al. (2024) demonstrated this ability by asking people to determine how an input list of numbers is transformed into an output list. Participants' knowledge of these list functions was tested by asking them to predict the output of new input lists. For example, participants would be shown pairs like `[1, 3, 2] -> [6]` and `[2, 2, 3] -> 7`. Then, they would be asked to predict the output of a new input list `[5, 4]`. Despite being provided no information about what operations to expect, more than 54% of the functions were learned by more than 50% of the participants. This is remarkable given the difficulty of the task, where the probability of answering correctly by chance is $\approx 10^{-30}$. List functions could include any possible computation and thus

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

encompass an infinite hypothesis space. Nevertheless, many people were capable of finding the right transformation with only 8 examples, suggestive of a symbolic inductive bias.

These findings, however, do not exclude the possibility that people use alternative non-symbolic representations to learn these concepts. This possibility is challenging to rule out due to the difficulty of accessing people's mental states during the learning process without resorting to self-reports, which may prompt post-hoc rationalizing or influence the reasoning process (Nisbett & Wilson, 1977). In order to understand the nature of participants' internal representations during the learning process, we asked participants to complete a modified list function learning task. Instead of presenting them with input-output list pairs in sequence as in Rule et al. (2024), we interspersed these trials with 11 test input lists on which *no* feedback was provided, meaning that we recorded generalization after each increment of labeled data. Participants' responses to these test lists were then used to infer a symbolic rule using the LoT library Fleet (Yang & Piantadosi, 2022). The success of this approach in predicting people's heldout responses was then used as an index of how biased towards symbolic representations participants' were. We find that people's mental representations, even during early learning, were more consistent with LoT representations than expected by chance. These results are consistent with the hypothesis that LOTs guide inductive inferences over the course of learning.

Furthermore, we show that this bias is not unique to humans. We evaluated large language models (LLMs) and a meta-trained neural network with an inductive bias towards LoT representations and found that they too show evidence of a symbolic inductive bias during learning. These findings are aimed at restructuring the debate on symbolic vs connectionist representations into a more productive and testable question — Do symbolic representations serve as a useful description of a particular behavior?

Methods

Participants

This study was approved by the institutional IRB. 555 participants were recruited using the online platform Prolific (https://prolific.com). They were required to reside in the United States and speak fluent English. The mean age of the participants was 41.102 (σ = 12.57). 256 participants identified as women, 242 as men, and 2 did not provide gender information. 330 participants identified was White, 78 as Black, 42 as Asian, 34 as Mixed, 14 as Other, and 2 did not respond. Data from 58 participants were excluded from the analyses for behavior that indicated inattention and a lack of effort such as responding with a copy of the input list, responding with a constant value, or not updating predictions in response to incoming data. Participants took an average of 20.75 minutes to complete the study (σ = 12.76). Participants were paid 5 USD for their participation and were provided up to 3.2 USD as a bonus payment based on the number of trials they responded to correctly.

Design and procedure

Participants were asked to predict how an input list of numbers was transformed to an output list. They first completed a practice set of 5 trials that followed the identity list function to acquaint themselves with the interface. Then, it informed them that the rule was changed and that the experimental trials will begin. Participants were assigned to one of 10 possible functions as a between-subjects factor. At no point was the definition of the function presented to them.

Experimental trials began with 5 blocks, each consisting of one trial with feedback and 11 intermediate trials without. 11 intermediate trials were chosen to provide Fleet, the LoT model, enough data to infer a symbolic rule, while holding out a trial for cross validation. For all types of trials, participants were shown an input list of numbers and were tasked with typing the transformed output list of numbers. On trials with feedback, the correct output list followed by green arrow or a red cross was displayed based on the accuracy of the participants' response. On trials without feedback, they were only shown a question mark. Participants could view all previous trials with feedback as they completed the task. The trials without feedback were presented to obtain a measure of participants' mental representations during learning. After all blocks were presented, participants finished the study by completing 4 additional trials on which feedback was provided. These trials were presented to determine whether participants were ultimately able to learn the list function. This resulted in a total of 64 experimental trials.

Materials

In contrast to Rule et al. (2024), the numbers in the input and output lists presented to participants in this study ranged from 0-9 with a max length of 10 digits to reduce the amount of typing required. Participants were informed of the output range. The functions presented to participants were chosen from Rule et al. (2024) to include a variety of arithmetic and list operations while being difficult enough to require at least four trials to learn. Two functions were modified to accomodate the reduced number range. The list functions that the trials in this study followed were:

- Replace the fourth element with 3
- If 3 or 9 are in the list, add them to the end
- Repeat the first number 3 times
- The second to last element
- The second largest number
- Add adjacent numbers
- Divide each number by 3
- Multiples of the smallest number length times
- Remove all of the first element
- Number of values equal to the length of the list

The 64 training input lists for each of these functions were manually curated to ensure the transformations were not trivial and that feedback provided useful information for participants to learn from.

Results

Human performance at list function learning

Trials were considered accurate when participants' predictions exactly matched the outputs of the list functions. Average accuracy across all trials was high, around 37%. Accuracy was above chance levels for all functions ($p \geq 10^{-12}$). This is particularly impressive considering that majority of trials were completed after receiving very little (≤ 5 trials) feedback, indicative of an inductive bias towards symbolic concepts. The functions varied in difficulty, with the most accurate being ADD ADJACENT NUMBERS (58%) and the least accurate being MULTIPLES OF THE SMALLEST NUMBER LENGTH TIMES (19%).

Critical to our analysis is determining how the performance of participants who were able to learn the function differed from those who did not. Partial competence and the presence of LoT representations in participants who were unable to learn the list function can indicate the presence of an inductive bias towards symbolic representations that is used to learn certain tasks. Participants who answered the last 3 of the 64 trials correct were considered to have learned the list function. There was considerable variability among the list functions in the number of participants who were able to learn them. 79% of participants were able to eventually learn ADD ADJACENT NUMBERS, but no participant was able to learn NUMBER OF VALUES EQUAL TO THE LENGTH. However, note that participants were able to respond accurately on many trials even though they did not learn the function. This may be due to the discovery of an alternate rule that either partially or fully explains the data previously observed, but is not the correct function. This explains how accuracy at NUMBER OF VALUES EQUAL TO THE LENGTH was relatively high despite no participant having learned the function.

We fit a hierarchical logistic model to characterize participants' accuracy and learning at the task. We predicted the accuracy of a trial using an intercept term and the number of trials with feedback a participant had seen. Hierarchical intercepts and number of trials with feedback terms were fit for each function, participant, whether they learned the function, and the interaction between the function and whether they eventually learned the function. As expected, participants who were able to learn the function were more accurate. Participants who learned the function had 20.09 times the odds of responding accurately compared to those who did not (Contrast median $\beta = 3.00$, 95% median credible interval [MCI] = $[1.73, 4.20]$). This trend persisted across the various functions.

Participants who learned the function differed in their rate of learning over the course of the experiment. On the first trial, before any feedback was provided, both groups were similarly unlikely to answer accurately (Median contrast = -0.001, 95% MCI = $[-0.054, 0.052]$). For the rest of the trials, participants who learned the functions were more likely to respond accurately (Median contrast >= 0.103). We only found weak evidence of learning in participants who were unable to learn the function (Median $\beta = 0.38$, 95% MCI = $[-0.136, 0.917]$, $p(\beta > 0) = 0.934$). Those who eventually learned the function however show robust signs of learning (Median $\beta = 1.998$, 95% MCI = $[1.433, 2.467]$). These aggregate statistics may obscure the fact that learning at this task was not necessarily smooth. Even participants who were unable to learn the function answered many earlier trials accurately. ADD ADJACENT NUMBERS in particular symbolizes the bumpy road towards learning for those who did not learn the function. For example, more than 80% of participants answered feedback trial 5 for NUMBER OF VALUES EQUAL TO THE LENGTH accurately, despite none learning the function. These patterns persisted even after excluding all participants' trials after the point they had learned the function.

In sum, participants' performance on the list function learning task reflects an intentionally difficult set of stimuli. Yet, participants performed above chance. The variability in the accuracy of their predictions and their ability to eventually learn the list functions provide a rich dataset to determine the nature of the inductive biases used during their learning process.

Fleet predictive accuracy

One contribution of the current work is our formalization of how much participants' representations agreed with LoT theories. Participants' predicted output lists on trials without feedback were used to infer the LoT expression that most likely generated their sequence of responses. This inference was conducted using the LoT library Fleet, which has been shown to be able to learn list function concepts similar to those tested in this study (Rule et al., 2024). We used Fleet to infer a program that predicted participants' output lists in response to the input lists, rather than the correct output lists. These inferred programs can be interpreted as participants' internal representations. Fleet was provided a rich set of primitives and operations consisting of booleans, integers, lists, logical operators, quantifiers, list operators, arithmetic, currying, etc. The operators provided to Fleet were interpretable and intuitive, which allowed it to infer explainable programs from participants' responses. The grammar provided to Fleet inherently included a bias towards simplicity, as the prior probability for primitive compositions of high depth is lower compared to simpler programs.

This configuration of Fleet was defined in order to embody a symbolic list function program learner, i.e., if participants were biased towards LoT representations, then Fleet can infer their mental representation using their intermediate outputs. If participants were not biased towards LoT representations for learning, then Fleet's inductive bias towards LoT representations would perform poorly and be expected to perform poorly at predicting participants' output lists. Thus, Fleet's ability to predict participants' outputs serves as a proxy for participants' inductive biases during learning.

Given a set of input-output pairs, there always exists a function that transforms all inputs to the corresponding outputs. To prevent Fleet from overfitting to participants' data, we used a leave one out cross validation (LOOCV) paradigm. First, Fleet was run on the set of all 11 input and participants' predicted

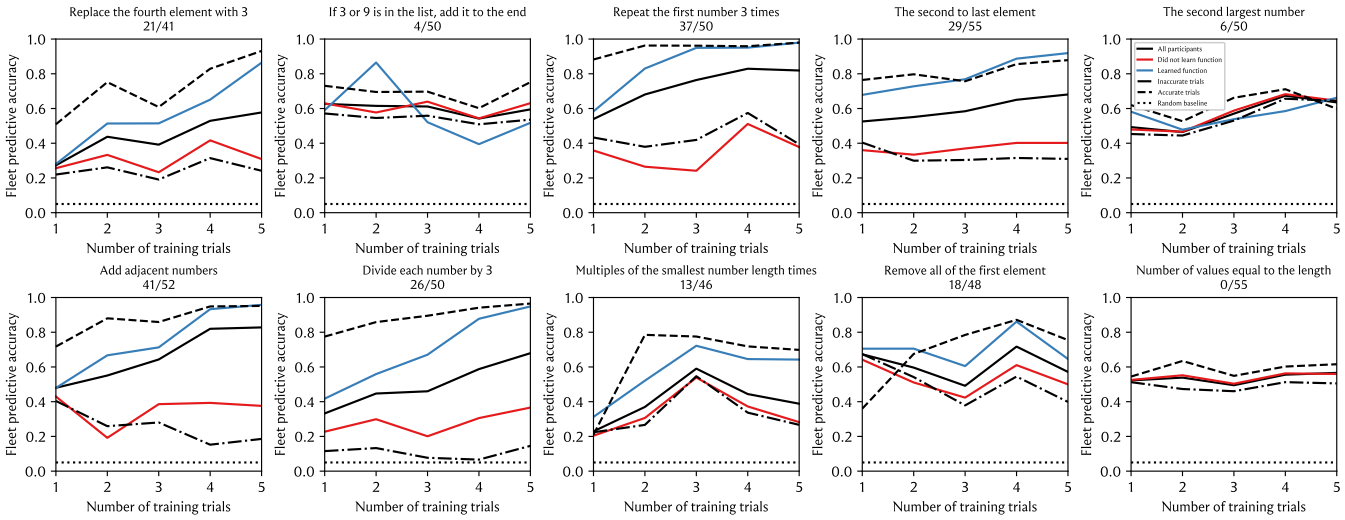

Figure 1: Fleet predictive accuracy for participants' responses to the intermediate trials for all list functions presented to participants

output pairs (blocks of trials without feedback) to generate a closed set of programs with high posterior probability. Then, these programs were re-evaluated using only 10 of the 11 pairs. The program with the highest posterior probability was then evaluated on the heldout pair. This was done 10 times, for all groups of 10 from 11 pairs. The average accuracy of the highest probability program to predict the heldout pair was termed the Fleet predictive accuracy. Chance-level for the Fleet predictive accuracy metric was determined using the same procedure but with trials permuted between all participants of the study. Blocks with above chance Fleet predictive accuracy indicate a symbolic inductive bias, while at or below chance level Fleet predictive accuracy suggests the usage of alternative mental representations.

Every participant completed 5 such blocks of intermediate trials without feedback, each after completing a trial with feedback. We used Fleet and the LOOCV procedure to predict their internal rule. To characterize the patterns in the Fleet predictive accuracy, we fit a hierarchical binomial model with Fleet predictive accuracy as the outcome variable. An intercept term and the number of training trials were added as predictors. The intercept and number of training trial terms were fit per participant, function, whether the participant learned the function, and the interaction between the function and whether the participant learned the function. We found that the Fleet predictive accuracy for heldout responses to all functions was above chance (≥ 0.05). This was even the case for participants who were unable to learn the list function. Given that these trials were presented after very little feedback was provided, these results provide a valuable indicator of symbolic inductive biases while learning these functions.

Participants' predictions were consistent with symbolic inductive biases even early in the learning process. The Fleet model was able to find rules that predicted participants' outputs after just one training trial was presented, even for participants who did not learn the functions (Median $p = 0.332$, 95% MCI $= [0.196, 0.489]$). We find that participants who learned the function become more rule-like over time (Median $\beta = 0.885$, 95% MCI $= [0.402, 1.277]$). Participants who were unable to learn the function, however, only showed weak evidence of becoming more rule-like with more training trials (Median $\beta = 0.185$, 95% MCI $= [-0.227, 0.599]$, $p(\beta > 0) = 0.837$). Fleet predictive accuracy for all functions over the course of the experiment are depicted in Figure 1. Particularly interesting are the trials on which participants did not respond accurately. A separate binomial model with a similar structure for these trials indicate that the Fleet predictive accuracy for inaccurate trials was above chance ($p \geq 0.05$). As can be seen in Figure1, only the Fleet predictive accuracy of inaccurate trials for DIVIDE EACH NUMBER BY 3 approaches the chance level.

The high Fleet predictive accuracy was not a direct consequence of their accuracy at the trials. While the performance of participants who were able to learn the list function was associated with higher average predictive accuracy ($r = 0.72$, $p < 0.0001$), the Fleet predictive accuracy for participants who were unable to learn the target function was more variable and was still associated with relatively high values ($r = 0.50$, $p < 0.0001$, Figure 2). We conducted additional analyses on a subset of the data without the trials that came after participants had already learned the functions and found similar results.

Symbolic inductive biases in neural networks

The use of a symbolic inductive bias is not unique to humans. LLMs have been demonstrated to accomplish many symbolic tasks successfully, for example, gold medal standard performance at the International Mathematical Olympiad. Evidence

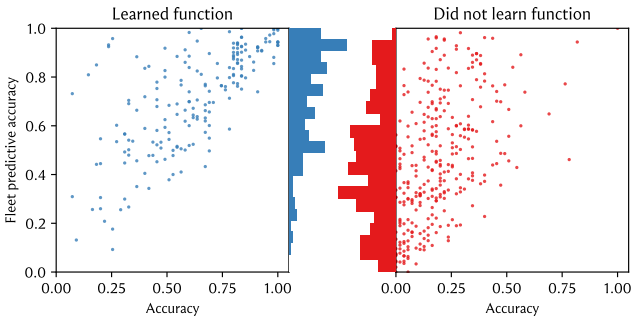

Figure 2: The association between the accuracy of participants' responses and their Fleet predictive accuracy. While the metrics are correlated, many participants who were mostly inaccurate were still strongly predicted by the Fleet model.

of high performance by models has often been presented as evidence for connectionist or alternative representations. Such an inference risks being inaccurate, as both symbolic and connectionist computational paradigms can emulate each other. We argue that it is more useful to determine whether behavior is congruent with the use of symbolic representations. That way, one can determine whether interpreting their behavior under a symbolic paradigm is a productive exercise.

To further demonstrate the validity of this approach, we used model agnostic meta learning (MAML) to induce a symbolic inductive bias into a small 4 layer encoder-decoder neural network. McCoy and Griffiths (2025) found that MAML is able to ingrain an inductive bias similar to Fleet into a neural networks which achieves comparable performance to it at a language learning task. Using a similar approach, we instead trained the model to predict the outputs of list functions, using rules that were generated by Fleet. This list function prediction model predicted 23% of list function outputs correctly compared to the 37% predicted by humans.

All models, a standard LLM (Google Gemini Flash 2.5), a reasoning LLM (Google Gemini Pro 2.5), and the MAML network performed relatively well on this difficult task. In contrast, a non-MAML trained network was unable to answer a single item accurately. We sought to test whether these models show a symbolic inductive bias similarly to humans. We used the same technique as humans to infer a congruent LoT representation. Fleet predictive accuracy for the models is depicted in Figure 3. Similarly to humans, the Fleet predictive accuracy was above chance level for all tested models. Even the MAML model, whose number of parameters and computational power pale in comparison to the standard LLMs, showed comparable levels of Fleet predictive accuracy compared to humans. While all these models use a connectionist architecture underneath, all show a bias towards LoT representations. The small MAML model was trained to use such an inductive bias, while the LLMs likely acquire it as a consequence of their natural language next token prediction objective. Therefore, all models, excluding the standard transformer, displayed evidence of a symbolic inductive bias like humans.

Discussion

This paper investigated the nature of people's mental representations while learning list functions, a task where people face a vast hypothesis space and are provided very little data. Without an inductive bias towards symbolic representations, learning such complex concepts would require a large amount of training data. Similarly to Rule et al. (2024), we find that people are capable of learning the list functions, suggesting the presence of an inductive bias towards symbolic concepts.

Throughout the early learning process, participants' responses to trials with no feedback helped us assess their inductive biases during the learning process. These responses were used to search for LoT representations that were most predictive. We found that participants' responses followed a LoT representation at above chance rates for all tested functions. Importantly, this was even the case for participants who were ultimately unable to learn the list function and on trials where participants provided an incorrect response. Together, these results stand as strong evidence towards the presence of a symbolic inductive biases in humans and hint at the use of symbolic representations during few-shot learning.

Critically, analyses of the behavior of LLMs and a MAML neural network using the same technique also revealed similar signs of symbolic inductive biases during learning. These models serve as examples of traditionally connectionist models learning to use structured representations and behaving similarly to symbolic models. Despite differences in how people and these models are architected, a symbolic description might be most apt way to characterize their behavior. Understanding the symbolic properties of these models, such as compositionality, is a topic of ongoing research that can help inform the constraints associated with symbolic inference using distributed representations (Khandelwal & Pavlick, 2025; Soulos et al., 2020). Findings from this program of research can help inform the structure of human cognition. In addition, people are subject to limited knowledge, memory, attentional, and processing constraints which may play an important role in their inference process. How this symbolic processing ability develops and functions in humans remains an open question that our formalization can help address by quantifying the likelihood of symbolic representations being used.

The degree of symbolic representations usage, i.e., Fleet prediction accuracy, observed in the current study should only be taken as a lower bound. Participants' may have changed their internal hypothesis for predictions over the course of the intermediate trials, as the distribution of the input lists might have provided some information for learning. The ability of Fleet to find predictive programs is also dependent on the primitives it was provided, which may differ from the primitives used by people and learned by the models. Nevertheless, we observe high Fleet predictive accuracy levels despite these limitations. We believe our formalization of the use of symbolic repre-

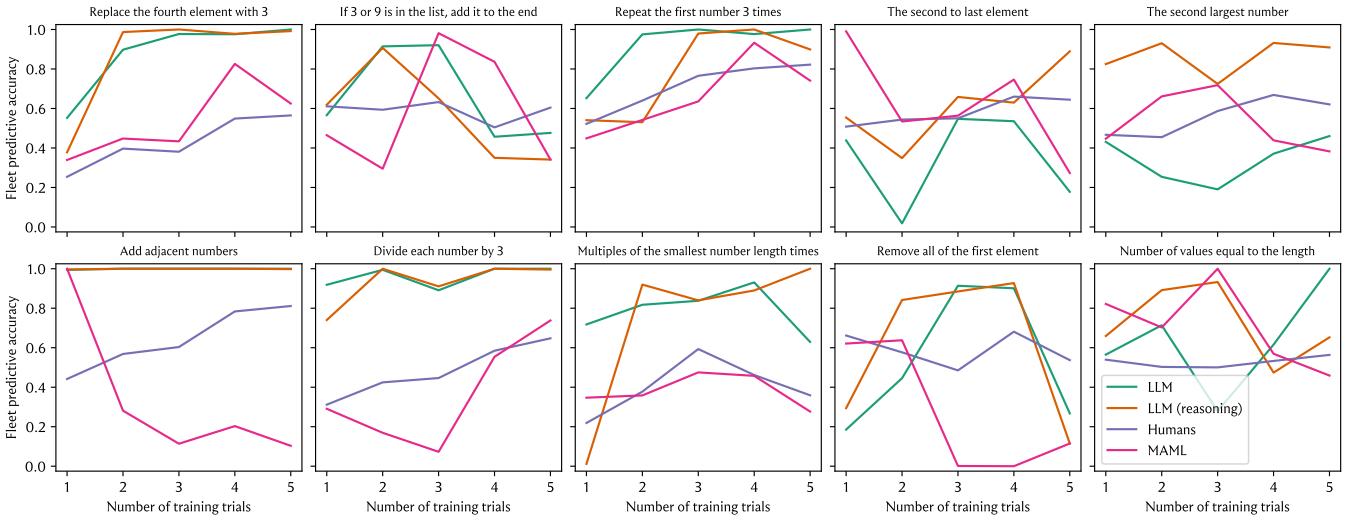

Figure 3: Fleet predictive accuracy for LLMs and the MAML neural network for all list functions. We observed above chance levels of Fleet predictive accuracy for all models.

sentations can help transform the symbols vs. unstructured representations debate into directed questions about behavioral properties and levels of analysis. Finding evidence of LoT representations in the behavior of humans, models, and even non-human animals can signpost the presence of symbolic computation and justify a symbolic level of analysis, allowing us to draw upon a rich tradition of research in cognitive science.

References

Amalric, M., Wang, L., Pica, P., Figueira, S., Sigman, M., & Dehaene, S. (2017). The language of geometry: Fast comprehension of geometrical primitives and rules in human adults and preschoolers. *PLOS Computational Biology*, *13*(1), e1005273. https://doi.org/10.1371/journal.pcbi.1005273

Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., Dayan, P., Demircan, C., Eckstein, M. K., & Éltető, N. (2025). A foundation model to predict and capture human cognition. *Nature*, 1–8.

Carey, S., & Bartlett, E. (1978, August). *Acquiring a Single New Word*.
ERIC Number: ED198703.

Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7*(1), 19–22. https://doi.org/10.1016/S1364-6613(02)00005-0

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, *407*(6804), 630–633. https://doi.org/10.1038/35036586

Fodor, J. A. (1975). *The language of thought* (Vol. 5). Harvard university press.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1–2), 3–71.

Gerstenberg, T., & Goodman, N. (2012). Ping Pong in Church: Productive use of concepts in human probabilistic inference. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *34*(34).

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A Rational Analysis of Rule-Based Concept Learning. *Cognitive Science*, *32*(1), 108–154. https://doi.org/10.1080/03640210701802071

Griffiths, T. L., Lake, B. M., McCoy, R. T., Pavlick, E., & Webb, T. W. (2025). *Whither symbols in the era of advanced neural networks?*

Khandelwal, A., & Pavlick, E. (2025, October 2). *How Do Language Models Compose Functions?* arXiv: 2510.01685 [cs]. https://doi.org/10.48550/arXiv.2510.01685

McClelland, J. L., & Patterson, K. (2002). Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in Cognitive Sciences*, *6*(11), 465–472. https://doi.org/10.1016/S1364-6613(02)01993-9

McCoy, R. T., & Griffiths, T. L. (2025). Modeling rapid language learning by distilling Bayesian priors into artificial neural networks. *Nature Communications*, *16*(1), 4676. https://doi.org/10.1038/s41467-025-59957-y

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259. https://doi.org/10.1037/0033-295X.84.3.231

Piantadosi, S. T. (2023). The algorithmic origins of counting. *Child development*, *94*(6), 1472–1490.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2012). Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition*, *123*(2), 199–217.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2016). The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological Review*, *123*(4), 392–424. https://doi.org/10.1037/a0039980

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*(1–2), 73–193.

Pinker, S., & Ullman, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Sciences*, *6*(11), 456–463. https://doi.org/10.1016/S1364-6613(02)01990-3

Pothos, E. M., & Chater, N. (2002). A simplicity principle in unsupervised human categorization. *Cognitive science*, *26*(3), 303–343.

Rule, J. S., Piantadosi, S. T., Cropper, A., Ellis, K., Nye, M., & Tenenbaum, J. B. (2024). Symbolic metaprogram search improves learning efficiency and explains rule learning in humans. *Nature Communications*, *15*(1), 6847.

Rule, J. S., Tenenbaum, J. B., & Piantadosi, S. T. (2020). The Child as Hacker. *Trends in Cognitive Sciences*, *24*(11), 900–915. https://doi.org/10.1016/j.tics.2020.07.005

Rumelhart, D. E., & McClelland, J. L. (1987). Learning the past tenses of English verbs: Implicit rules or parallel distributed processing? In *Mechanisms of language aquisition* (pp. 195–248). Lawrence Erlbaum Associates, Inc.

Sablé-Meyer, M., Ellis, K., Tenenbaum, J., & Dehaene, S. (2022). A language of thought for the mental representation of geometric shapes. *Cognitive Psychology*, *139*, 101527. https://doi.org/10.1016/j.cogpsych.2022.101527

Soulos, P., McCoy, R. T., Linzen, T., & Smolensky, P. (2020). Discovering the Compositional Structure of Vector Representations with Role Learning Networks. In A. Alishahi, Y. Belinkov, G. Chrupała, D. Hupkes, Y. Pinter, & H. Sajjad (Eds.), *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP* (pp. 238–254). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.blackboxnlp-1.23

Yang, Y., & Piantadosi, S. T. (2022). One model for the learning of language. *Proceedings of the National Academy of Sciences*, *119*(5). https://doi.org/10.1073/pnas.2021865119