Computation or Weight Adaptation? Rethinking the Role of Plasticity in Learning

Gili Lior[1]*, Yuval Shalev[1,2]*, Gabriel Stanovsky[1], Ariel Goldstein[2,3]

[1]School of Computer Science and Engineering, The Hebrew University of Jerusalem, Israel

[2]Department of Cognitive and Brain Sciences, The Hebrew University of Jerusalem, Israel

[3]Business School, The Hebrew University of Jerusalem, Israel

*Equal contributors

Abstract

The human brain is an adaptive learning system that can generalize to new tasks and unfamiliar environments. The traditional view is that such adaptive behavior requires a structural change of the learning system (e.g., via neural plasticity). In this work, we use artificial neural networks, specifically large language models (LLMs), to challenge the traditional view about the role of plasticity in learning and suggest that such an adaptive behavior can be achieved solely through computation if the learning system is sufficiently trained. We focus on statistical learning paradigms. These require identifying underlying regularities in seemingly arbitrary word sequences and are largely considered to require neural plasticity. LLMs can capture arbitrary structures without weight adaptation despite the divergence from their natural language training data. Our work provides novel insights into the role of plasticity in learning, showing that sufficiently trained learning systems are highly flexible, adapting to new tasks and environments solely through computation, much more than previously acknowledged. Furthermore, our work opens the door for future research to use deep learning models to conjure hypotheses about the brain.

Introduction

The human brain is a learning system, able to quickly adapt to new environments and tasks using previously acquired knowledge and experiences. For example, riding a bicycle is a complex skill requiring substantial practice. However, it would not be surprising if an experienced skateboard rider could excel at bicycle riding after just a few attempts, leveraging their previously acquired balance and coordination skills. This generalization ability, sometimes referred to as transfer learning (1), raises a fundamental question: to what extent can a learning system show a non-trivial application of its already-acquired knowledge? Can a complex adaptive behavior be performed solely through computations without changing the system's structure? In the brain, such changes are identified as

neural plasticity (2), while in artificial neural networks, change is equivalent to weight calibration. In this work, we use artificial neural networks to question the role of plasticity in learning.

Recent advances in deep learning literature show that trained artificial neural networks (specifically Large Language Models; LLMs) seem to learn various tasks from a few examples seen only during inference, without explicit weight calibration (Figure 1; (3)). This phenomenon, often termed in-context learning (ICL), demonstrates that pre-trained systems (e.g., LLMs) can leverage their knowledge to perform fundamental new tasks without plasticity. Specifically, we focus on statistical learning paradigms, which are widely used to test human learning capabilities and are considered to require brain plasticity (4).

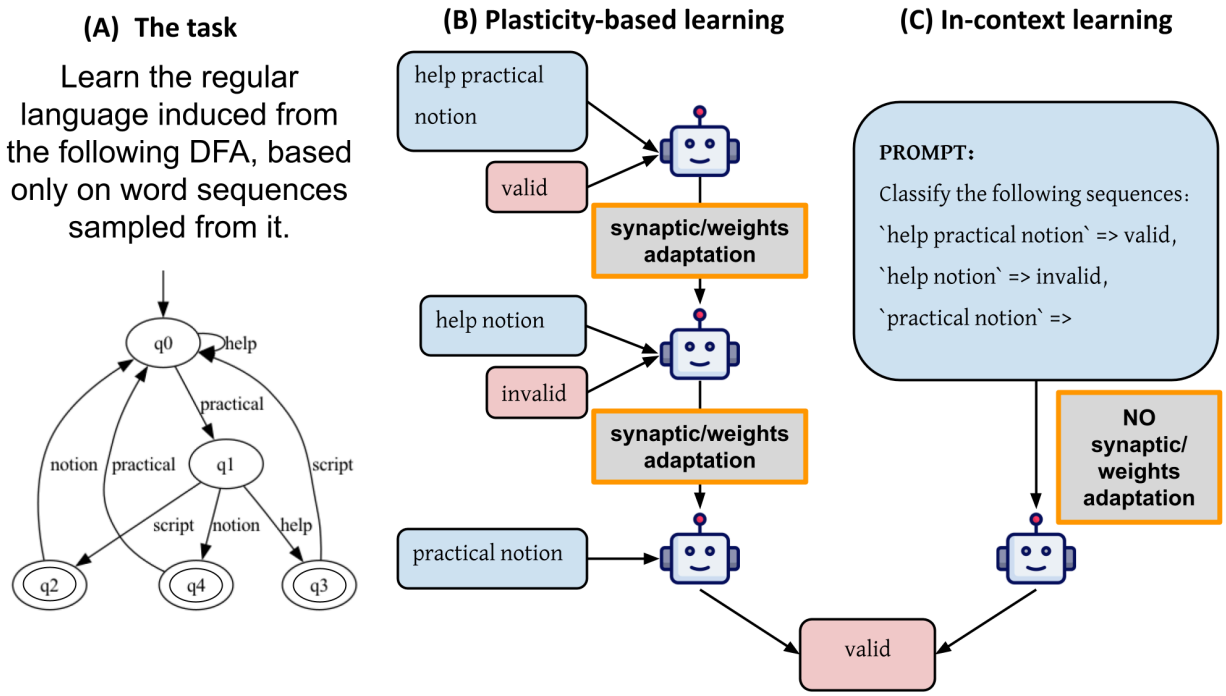

Figure 1: Demonstrating plasticity-based vs in-context learning. (A) Example for statistical learning task to identify a regular artificial language induced from a DFA. It can be learned via (B) plasticity-based learning, i.e., fine-tuning, or (C) In-context learning (ICL) paradigms. Fine-tuning involves plasticity, as the model updates its parameters following feedback over its predictions (gradient update), while ICL only involves prompting with a few demonstrations of the task.

The statistical learning paradigm measures participants' ability to observe samples from a new environment following some arbitrary regularity (e.g., dictated by a context-free grammar) and predict outcomes consistent with its underlying structure (5, 6). Learning in this task occurs as the cognitive system's behavior is adapted when exposed to new samples and structures (7, 8). We simulate two learning tasks, inspired by two different

statistical learning paradigms: Artificial Grammar Learning (AGL; (9)) and the Serial Reaction Time Task (SRTT; (10)). SRTT is a motor learning task that measures reaction time to stimuli. Participants in this task show a decrease in their response times to stimuli drawn from a repeated sequence. This paradigm indicates that people can adapt their expectations to a sequence structure without knowing there is an underlying regularity in the sequence they observe (11, 12). The AGL paradigm is an implicit learning task where participants are presented with sequences of words that follow the rules of a regular language. Afterwards, they are asked to classify the validity of new sequences (i.e., whether they belong to that language or not). Participants in this task are not instructed to search for any regularity in the stimuli, but their ability to determine the validity of unseen sequences indicates that they have managed to learn the underlying regularities of the regular language. These paradigms are well established in the field of statistical learning and demonstrate the ability of humans to adapt their behavior to unrecognized statistical structures. Interestingly, these two tasks diverge from the natural language used for training LLMs. Yet, we show that LLMs can adapt to statistical patterns similar to those underlying the stimuli in these tasks without plasticity.

Even though statistical learning in humans is known to involve neural plasticity (4) , several recent studies in the neuroscience literature suggest the involvement of other mechanisms in learning, and not solely depending on neural plasticity. Kumaran & McClelland (13) and Schapiro et al. (14) showed models of the hippocampus in which synaptic plasticity is used to learn the training dataset. Still, the underlying regularities are not encoded in the model weights. Then, generalization in this network occurs through recurrent connections without additional training. These models demonstrate how the brain may use computations to solve statistical learning tasks by integrating knowledge to find regularities that were not explicitly encoded in weights. Sorscher et al. (15) showed a biologically plausible model for the brain in which extensive training yields representations that require very little plasticity to adapt to new labels. This model demonstrates how rich representations can be used for out-of-distribution generalization, which could compensate for the lack of data samples in learning tasks.

Inspired by this line of work, we showcase LLMs to derive new insights that may apply to the human brain. While the mentioned neuroscience literature builds dedicated models, curates special training sets, and tests on similar instances to those in their training set, in our work, we use off-the-shelf LLMs, trained on massive amounts of natural language. We show that LLMs can generalize to complex synthetic structures without any form of explicit plasticity.

Our findings suggest a broader spectrum of potential learning mechanisms that could coexist or complement synaptic plasticity's role in human cognition. Furthermore, this work showcases how advancements in artificial intelligence, particularly deep learning, provide

novel insights into foundational concepts like learning and plasticity. It presents a new scientific perspective to the broader cognitive science community.

Results

Our experiments are designed to test the flexibility of the large language models without any weight adaptation. In particular, we focus on the recent Mistral model (16) in learning arbitrary structures without changing their inner state, i.e., without updating their parameters (weights). We specifically choose the Mistral LLM as it is currently one of the best-performing open models (16), and in the supporting information (SI) we show that our findings hold for other off-the-shelf LLMs.

We formulate two statistical learning tasks to test the ability of LLMs to capture the underlying regularities of a statistical structure, which can be represented as a directed graph with transition probabilities between states (Fig. 2.A). First, inspired by SRTT, we test the abilities of an LLM to capture deterministic regularities, i.e., each state is connected to a single consecutive state. In SRTT, a reduction in participants' reaction time to stimuli reflects that the stimuli become more expected, which means that participants learn the underlying regularities of the sequence, similarly to LLMs learning the underlying regularities of a deterministic graph.

Second, inspired by the AGL task, we test LLMs' ability to learn non-deterministic regularities. In AGL, participants are presented with multiple sequences sampled from a regular language. The validity classification task as formulated in AGL can be simulated using a deterministic finite automaton (DFA; Fig. 2A). To test LLMs, we formulate the task as follows: given a prefix of a sequence, we test the probabilities that the model assigns to all possible tokens in some language's vocabulary. In this paradigm, learning is demonstrated when the LLM assigns significantly higher probabilities to valid next tokens than to invalid next tokens in some sequence of the corresponding regular language.

We do not perform model parameter updates in our experimental setup; we only use an incremental input setup. In each iteration, we feed the model with a single additional word that is concatenated to all previous sequences of words (Fig. 2B). This new word is a valid continuation of a sequence, corresponding to each of the tasks. After each concatenation of a word, we extract the probability that the model assigns to the next word (Fig. 2C). We will show model flexibility in learning the underlying probabilities without changing its inner state, but only by feeding it with a longer input congruent with its regularity.

To generate multiple sequences congruent with the AGL-like experiment, we perform a random walk on a graph that is similar to the DFA that forms the regular language, which is further explained in section A in SI, and demonstrated in Fig. S1.

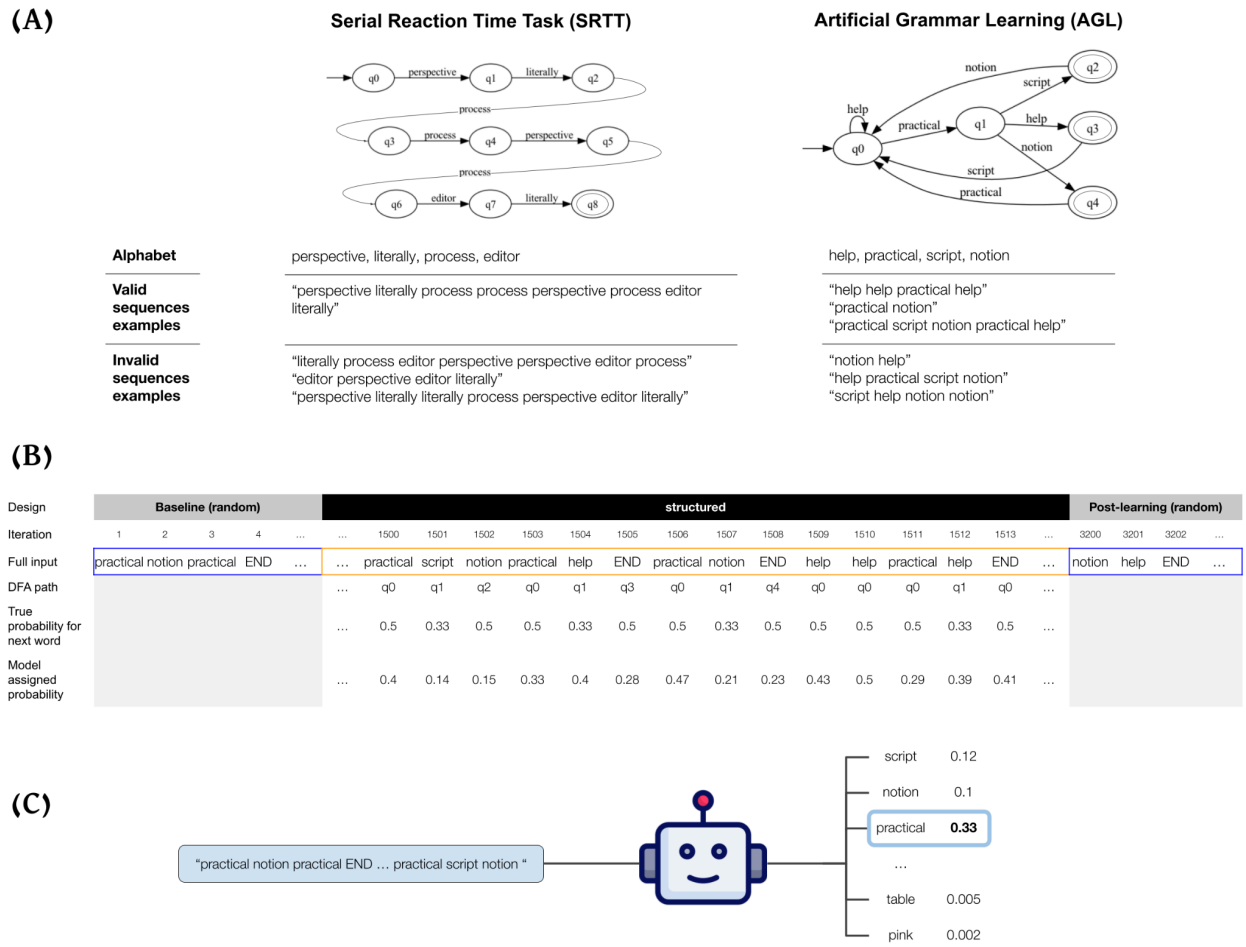

Figure 2: The experimental setup of our work. (A) The DFAs that simulate SRTT and AGL, with the graph alphabet and the induced valid and invalid sequences. The AGL is inspired by Pothos (17). (B) An input example for artificial grammar learning consists of random sequences followed by valid sequences induced from the DFA and random sequences. This sequence is induced by traversing the DFA, following its transition probabilities, which are also presented in the table. (C) Single iteration example, where we feed the model with a prefix of our input, then the model assigns a probability for each word in its vocabulary to be the next word, and we extract the probability that the model assigns to the actual next input word. Our experiment consists of len(input) iterations, where in each iteration, we add one more word to the prefix we feed the model with.

Inspired by classic SRTT experiments (11), our experiment consists of three parts. First is a baseline part of randomly generated sequences with uniform probability over a chosen alphabet (i.e., invalid sequences). This part's goal is to eliminate all biases of the pre-trained LLM towards specific sequences. Next is the structured part, which consists of sequences induced by traversing the corresponding task's DFA, i.e., valid sequences. Last is the post-learning part of randomly generated sequences again (i.e., invalid

sequences). In this paradigm, learning is demonstrated by an ascent in performance between the baseline part (first part) and the structured part (second part). Part three, which we refer to as post-learning, was designed as a control for the possibility that the LLM simply increases the probabilities of all the words from the vocabulary of the DFA.

To eliminate possible biases related to the specific vocabulary choice, our results are the mean of 100 different runs, where in each run, we choose a different vocabulary and randomly sample different sequences for both baseline and structured parts. For robustness, we provide results on more language models and different DFAs in Fig. S3-S4.

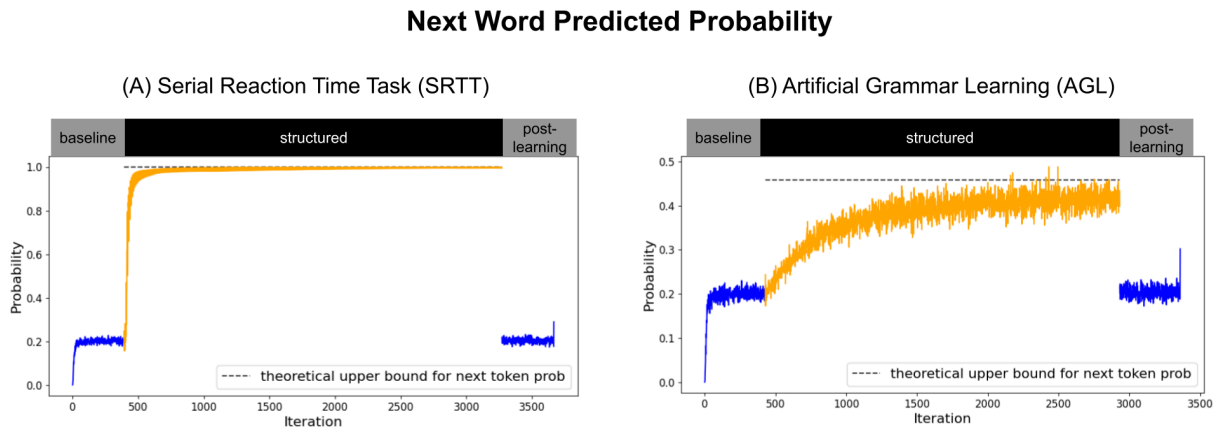

Figure 3: Next word probability for SRTT and AGL. Blue are the baseline parts, random words in random sequences, and orange is the probability of words combining sequences induced by the corresponding DFA. The black dashed line is the theoretical upper bound of the next word probability, which for SRTT sequential trials is 1, as its DFA is deterministic, and for the AGL, it is ~0.45.

In Fig. 3, we illustrate the model's predictive capabilities by presenting the probability assigned to the input's congruent next word in each iteration. At first, at the baseline part, the model quickly adjusts its predictions to the equally distributed transition probabilities between words in some vocabulary. Though we address this part as a baseline, it still requires LLM to learn some statistical structure without additional training. However, it is a much simpler structure - it can be thought of as a graph with a single state, with a single edge to itself, and all words in the vocabulary are of that edge label. These induced sequences are easier to learn in-context, as the model only needs to learn the vocabulary, and not transitions, which makes it a much easier task. Then in the structured part, for both SRTT (Fig. 3A) and AGL (Fig. 3B), we see the expected learning pattern - an ascent from chance probabilities at the baseline random part (depicted in blue) towards the theoretical transition probabilities upper bound at the structured part. The predictions are depicted in orange, and the upper bound is in a black dashed line; to explain how it is

calculated, see section B in SI. Due to the deterministic transitions of the SRTT, the theoretical upper bound for the next token probability is 1, while for the AGL, it is an average of 0.45. The model reaches 92% of the SRTT upper bound and 79% of the AGL upper bound, averaging the last 1000 iterations of the structured part. As expected, we also see a descent back to chance probabilities at the post-learning random part (in blue). In section C in SI materials and methods, we accumulate the next word probabilities into the probability of a sequence and provide a similar analysis for the next sequence probability. We further provide a statistical analysis of the results, presented in Fig. S5 In the SI.

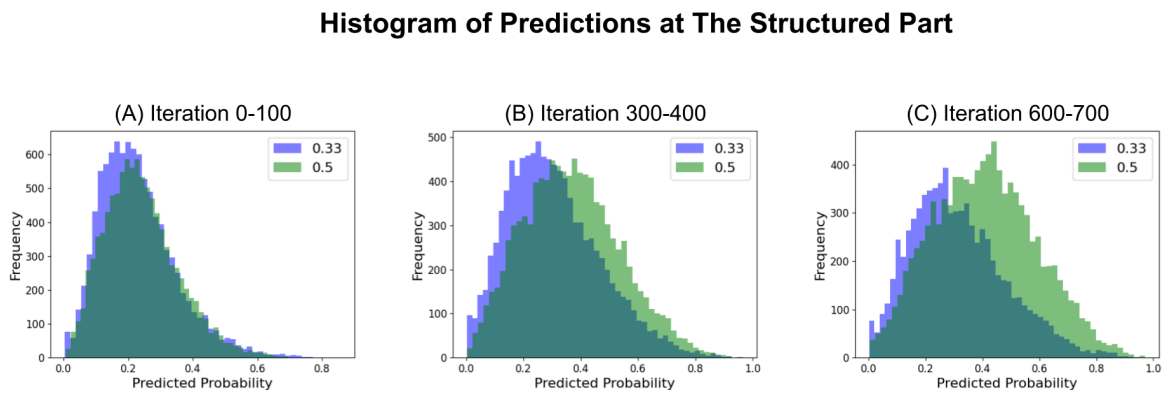

Figure 4: Evolution of Prediction Probabilities Across Experiment Lengths. Histograms of prediction probabilities for all samples, categorized by true probability (0.33 and 0.5). The three subfigures illustrate the model's learning progression as it encounters longer sequences. Initially (A), predictions for both true probability classes are dispersed, indicating random-like behavior. As sequence length increases (B and C), the distributions become more distinct and centered around their respective true probabilities, demonstrating improved model performance.

Additionally, Fig. 4 offers a granular view into the model's predictive behavior across varying experiment lengths. By examining the distribution of prediction probabilities for different true probability classes, we can directly observe the model's learning trajectory. In contrast to Figure 3, which focused on the average prediction, this figure provides insights into the spread and shape of the predictions. As the sequence length grows, the histograms in Figure 4 clearly demonstrate the model's progression from a random-like state to a more refined predictive capability, where predictions increasingly cluster around the true probability values.

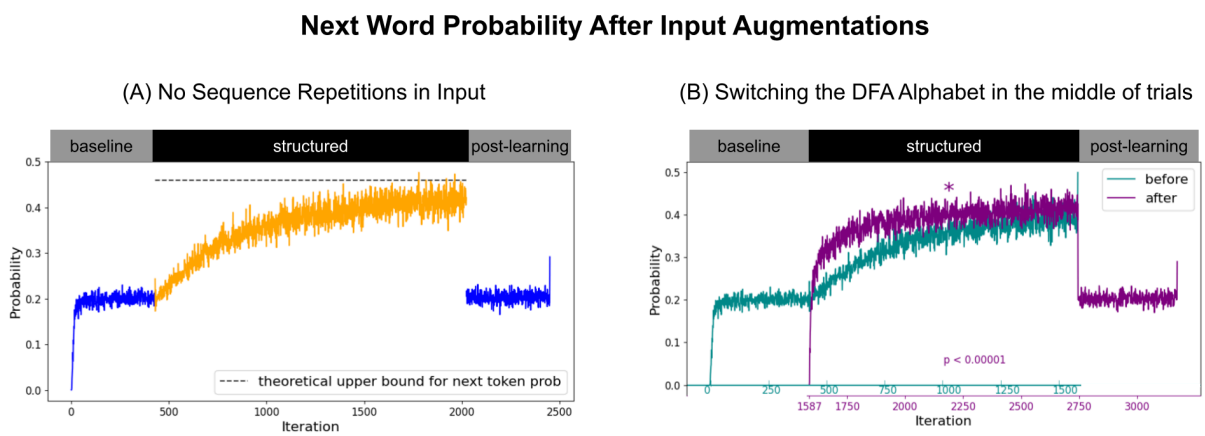

Figure 5: Next word probability for two experiments with some input augmentation. Figure A on the left presents the probability for an input with no repeating sequences, indicating that the model does not use memorization in its predictions. On the right, we present the probability of switching the DFA alphabet in the middle of the input, which results in sequences with the same structure but consisting of different words. The purple graph, which is the sequence probability after the switch, increases quicker than at the beginning of the sequential trials (in blue), indicating that the recovery from the switch is faster than the initial learning, hence evidence for the model learning the abstract structure of the graph and not just an association between the words shown in its input.

One possible hypothesis for the model assigning the correct probabilities is that the model may memorize sequences and not learn the underlying structure that induces the language. To control this alternative explanation, we run an experiment where we drop all duplicating sequences in the model input (~10% duplicate sequences). As presented in Fig. 5A, we find the learning pattern of an ascent in probabilities between the baseline and the structured part, indicating that the model does not rely on memorization but captures the underlying statistical structure.

We also suggest a setup that controls for the possible alternative explanation that the model may learn the association between specific words shown in its input and not the underlying structure. In the experiment presented in Fig. 5B, we switch the DFA alphabet in the middle of the input generation (in trial 1587), i.e., we continue to traverse the same graph, but we change the words that are the edge labels, resulting in sequences following the same structure but with a different vocabulary. In comparing the mean predicted probability before and after the switch, a paired t-test revealed a significant difference ($t(1156)=-44.67$, $p=4.62e-254$). This indicates that the model learned the structure that induces the sequences and not the specific words of the sequences. These findings align with a previous line of work regarding humans' ability to transfer learning in AGL (17). Researchers have shown that humans can transfer their learning of an underlying structure

with a certain vocabulary to a different vocabulary, quickly adapting to the new vocabulary (18, 19).

A potential confound for the increased learning rate after the vocabulary switch may stem from the initial exposure to random sequences, making it harder for the model to understand the structured rules of the regular language, necessitating an additional effort to discard previously acquired patterns from the unstructured, random phase. To explore this potential confound, we conduct a series of experiments varying the length of the random baseline phase. Our findings indicate that a longer random baseline correlates with slower learning of the structured phase, as depicted in Figure S6 of the SI. However, learning rate improvements after the vocabulary switch remain significantly pronounced, even without exposure to random sequences. Figure S7 in the SI compares the learning at the first 100 iterations at the start of the structured phase versus the first 100 iterations following the vocabulary switch, indicating a significantly increased learning after the switch. This reinforces our conclusion that even when the edge labels are changed, the model learns a statistical structure and can generalize the DFA structure.

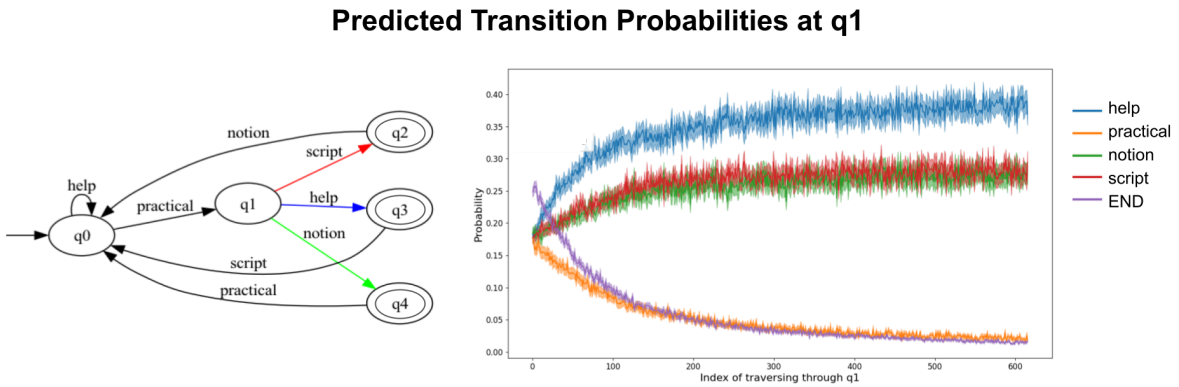

Figure 5: The model predictions for transition probabilities at q1. The true transition probabilities at q0 are 0.33 to script, help, and notion. We indeed see that the model-assigned probabilities converge around the true probabilities.

A fundamental trait of a regular language that is induced by some DFA, is that a prefix of a sequence has several valid continuations, which is represented in the DFA with multiple outgoing edges from a certain state. A model that learns a regular language should also reflect this aspect, as demonstrated in Fig. 5. In each iteration, we extract the probabilities of all possible continuations and compare them with the graph's transition probabilities at each state. In Fig. 5, we show the probabilities the model assigns each word given some prefix that ends at q1, and in Fig. S7 in SI, we provide a similar analysis for all other DFA states.

We indeed see that the model successfully differentiates between two groups of words: those that are not valid continuations from q0, whose probabilities decrease to zero, and

those that are valid continuations from q0, whose probabilities increase and stabilize roughly around 0.3. This finding demonstrates the LLM's ability to identify multiple valid continuations at each step, supporting the claim that it captures all possible valid sequences in the regular language defined by the underlying structure.

Additionally, there is a noticeable difference in the probabilities assigned to "help" compared to "script" and "notion." This disparity may be attributed to the higher frequency of "help" in the regular language (half of the valid sequences in the language start with "help"), which could lead to a model bias toward this word.

Discussion

This work introduces a novel in-context learning approach for tasks that diverge from natural language patterns. The results show the flexibility of pre-trained large language models (LLMs) in learning out-of-distribution abstract structures without additional model training. Throughout our different experiments, we ruled out potential confounds to the model's learning, such as memorization or spurious associations between words, establishing the model's capability to capture underlying structural patterns.
While previous research has showcased LLMs' in-context learning abilities on tasks aligned with their pre-training distribution,(e.g., question answering (20)), our findings unveil a substantially broader capability. We reveal that LLMs can abstract over arbitrary structures fundamentally different from natural language, a phenomenon not previously reported. In contrast to prior work leveraging LLMs for non-natural language data, e.g., LLMs for biological encoding (21, 22), which typically involves fine-tuning (i.e., plasticity), we showcase the efficacy of in-context learning for such tasks.
Our results demonstrate the ability of sufficiently powerful language models to generalize to tasks fundamentally different from their training data without explicit adaptation. A potential explanation for this capability is the hypothesis that in-context learning is a form of implicit gradient descent (23–25). This perspective posits that the model's apparent adaptation during inference is akin to a single step of gradient descent, without altering the model's underlying parameters. However, this interpretation remains speculative, as existing research supporting this claim is limited to simplified settings such as linear regression and has yet to be conclusively established for complex setups.

Our findings were based on LLMs, which recently demonstrated strong resemblance with the brain's language representation (26–30) and similar behavioral performance during memory recall tasks (31). In conjunction with the findings of our study, this notable

similarity offers potentially new insights into generalization mechanisms in the brain, namely that prior knowledge of the learning system can be leveraged to perform new tasks without plasticity.

This paper questions some of the main assumptions regarding learning and its relation to change in the learning system. First, we suggest that the brain may rely on computations rather than weight adaptation more than previously acknowledged. This follows Schapiro et al.'s (14) findings, which proposed a model for the hippocampus that could support simple statistical learning before having the opportunity to encode the statistical regularities in its weights. The generalization in this network occurred during inference time, using recurrent connections. The idea of achieving generalization via computation was supported from LLMs perspective. Geva et al. (32) showed that transformer-based LLMs encode memories within their feed-forward layers. Still, the prediction is achieved in inference time via an efficient computation of those encoded memories. Our findings integrate the above ideas into a broader perspective without assuming specific neural mechanisms or cognitive processes.

Second, we show that sufficiently trained predictive systems can generalize to complex regularities apparently far from their training data distribution. Sorscher et al. (15) suggested biologically plausible modeling for the brain, showing that representations similar to those in the inferotemporal (IT) cortex can help to identify new visual concepts using a simple plasticity rule. They suggested that extensive pre-training yields rich representations that enable the applicability of such a simple plasticity rule. Our work shows even more impressive capabilities of a pre-trained learning system, as we show that they could generalize to patterns and structures far from any of their training data distribution without adaptation.

Last, our findings suggest a distinction between the computation producing the adapted output and the system's state change. While learning in the brain is assumed to happen in tandem (27, 28), in-context learning proposes that they can be observed disjointly. It is possible that the brain can adapt to a new input via computation alone, and this adaptation does not inherently require plasticity. Plasticity is then used to encode the adaptation within the system parameters, for a longer-term process. This raises further questions about the distinction between memory, learning, and generalization, and we believe that this work opens the door for future neuroscience research to use deep learning and AI models to provide novel insights into foundational concepts from the cognitive and neuroscience communities.

Acknowledgments
We gratefully acknowledge our colleagues' support and contributions to the reviewers' insightful feedback: Prof. Haim Sompolinksy, Prof. Baruch Eitam, Dr. Sebastian Michelmann, Dr. Noam Siegelman, Dr. Yael Bitterman, Mr. Itay Itzhak, Ms. Daria Lioubashevski.

Supporting Information Text

A Sequence Generation for The Experiments

To generate sequences for our experiment, we employ a random walk on a graph that is designed to reflect the structure of the DFA representing a regular language, with an additional final state incorporated into the design. This process is illustrated in Fig. S1.

The graph structure is similar to the DFA, as each node in the graph corresponds to a state in the DFA, and the edges correspond to the DFA transitions. We add an additional node, in which each node corresponding to an accepting state of the DFA has an outgoing edge to that additional node.

We then perform a random walk on this graph, starting at the node corresponding to q0, and ending at the additional node mentioned above. During each step of the random walk, a transition is made to a neighboring node according to a random choice of the available edges. Each edge in the graph is associated with a specific word, corresponding to the DFA transitions. So, in each step of the random walk, we add to the sequence the word corresponding to the edge that was randomly chosen between all outgoing edges. This randomness introduces variability in the sequences while maintaining the structural constraints of the DFA. The sequences generated by this process inherently conform to the structural rules defined by the DFA. The inclusion of the final state ensures that each sequence is complete and valid.

This approach allows us to create a variety of sequences that adhere to the regular language's structural constraints, enabling a robust evaluation of the LLM’s ability to learn and predict based on such sequences. The random walk method ensures that our sequences exhibit the complexity and variability necessary for meaningful experimentation while maintaining adherence to the underlying DFA structure.

B Calculating the theoretical upper bound for transition probabilities

The DFA representing the task induces transition probabilities for the graph traversing. In this paper, we compare how close are the probabilities that the LLM assigns to the transition probabilities that the DFA induces, referring to it as the theoretical upper bound.

Let S be a sequence induced by traversing the following DFA states:

$$Q_S = \left\{ q_{t_0}, q_{t_1}, q_{t_2}, \ldots, q_{t_n} \right\}$$

Then, the theoretical upper bound for the sequence is the average of all transition probabilities between each pair of consecutive states:

$$P\left(Q_S\right) = \frac{1}{n} \sum_{i=0}^{n-1} p\left(q_{t_i} \rightarrow q_{t_{i+1}}\right)$$

In this paper, the theoretical upper bound is calculated over the complete input sequence, which is 1 for the SRTT, as each state has only one outcoming edge, and for the AGL is 0.45, as most states have two outcoming edges (with probability 0.5), and one state has three outcoming edges (with probability 0.33).

C Probabilities over sequences

In the paper, we presented the probability over each word, and below, we accumulated the word probabilities into sequence probability. Each sequence is separated by a special 'END' token, which represents traversing from an accepting node back to q0, such that between each two 'END' tokens, there is a valid sequence that belongs to the vocabulary of the DFA.

C.1 Mean on sequence. First, we calculate the probability of a sequence as the mean of all predicted probabilities of its words, given all previous tokens:

$$p(sequence) = \frac{1}{|sequence|} \cdot \sum_{i=1}^{|sequence|} p(sequence[i] \mid prefix, sequence[: i - 1])$$

For example, if we want to calculate the sequence probability of "practical notion" from Fig 2B (iterations 1506-1608):

$$P(\text{"}practical\ notion\ END\text{"}) = \frac{1}{3} \cdot [p(\text{"}practical\text{"} \mid prefix) + p(\text{"}notion\text{"} \mid prefix, \text{"}practical\text{"}) + p(\text{"}END\text{"} \mid prefix, \text{"}practical\ notion\text{"})]$$

Where $prefix$ is the concatenation of the 1505 input words until this sequence?

In Fig S1a, we see that the model also shows learning at the structured part when calculating the mean of a sequence.

C.2 Multiply the probability of the tokens combining the sequence. In LM evaluation, it is common to factorize the joint probability over a sequence as the product of conditional probabilities of the words and symbols combining this sequence (35, 36).
E.g., we calculate the probability of a sequence as follows:

$$P(\text{"}practical\ notion\ END\text{"}) = p(\text{"}practical\text{"} \mid prefix) \cdot p(\text{"}notion\text{"} \mid prefix, \text{"}practical\text{"}) \cdot p(\text{"}END\text{"} \mid prefix, \text{"}practical\ notion\text{"})$$

In Fig S1b, we see the learning in the structured part, as well as in calculating the production of a sequence.

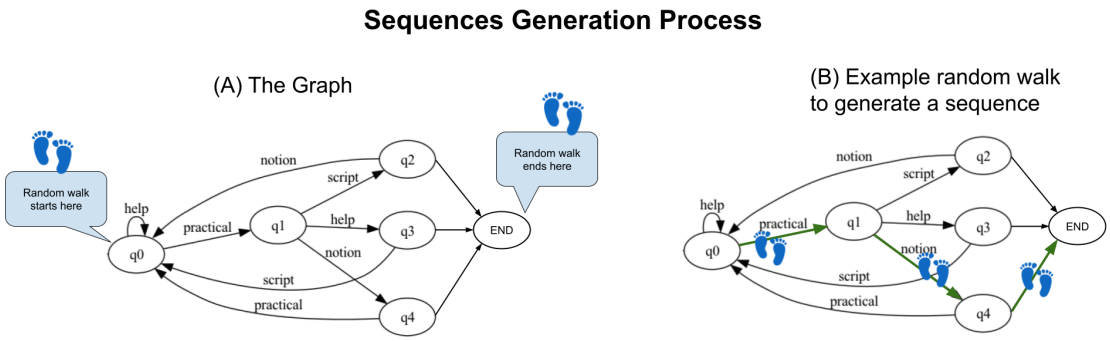

Figure S1: Sequence Generation Process. A illustrates the graph used for sequence generation. A single random walk from the initial state (q0) to the terminal state (END) produces a sequence. B provides an example of such a walk. At each node, an outgoing edge is selected uniformly. For instance, the edge leading to 'practical' is chosen with a probability of 0.5, followed by 'notion' with a probability of 0.33. The walk terminates at END with a probability of 0.5,resulting in the sequence "practical help END".

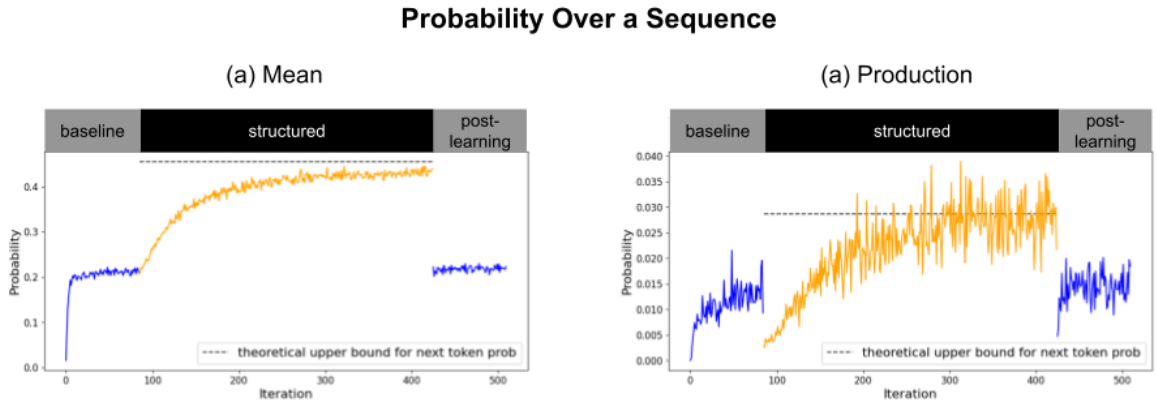

Figure S2: (a) predicted probabilities of a sequence via mean on tokens. (b) predicted probabilities of a sequence via the production of conditioned probabilities of its tokens.

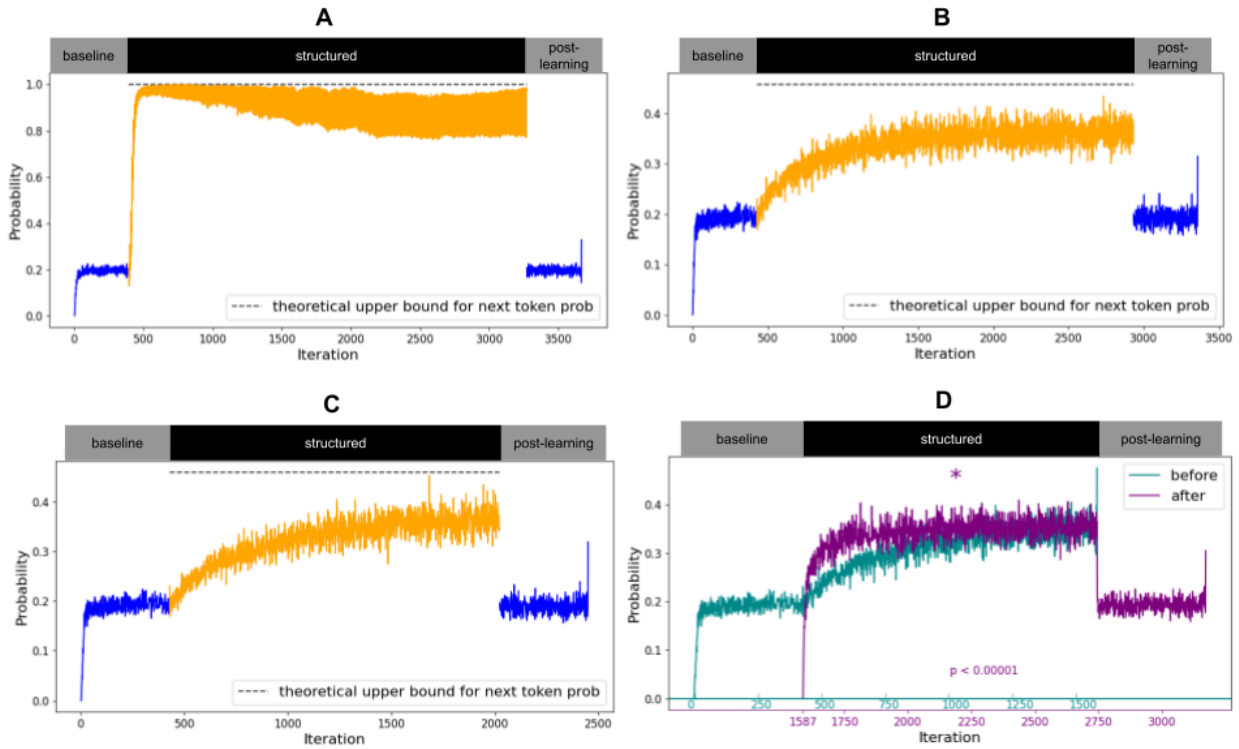

Figure S3: Results of LLaMA-2 language model (37). (A) SRTT. (B) AGL base setup. (C) AGL switched vocabulary in the middle of the trial. (D) AGL has no duplicates.

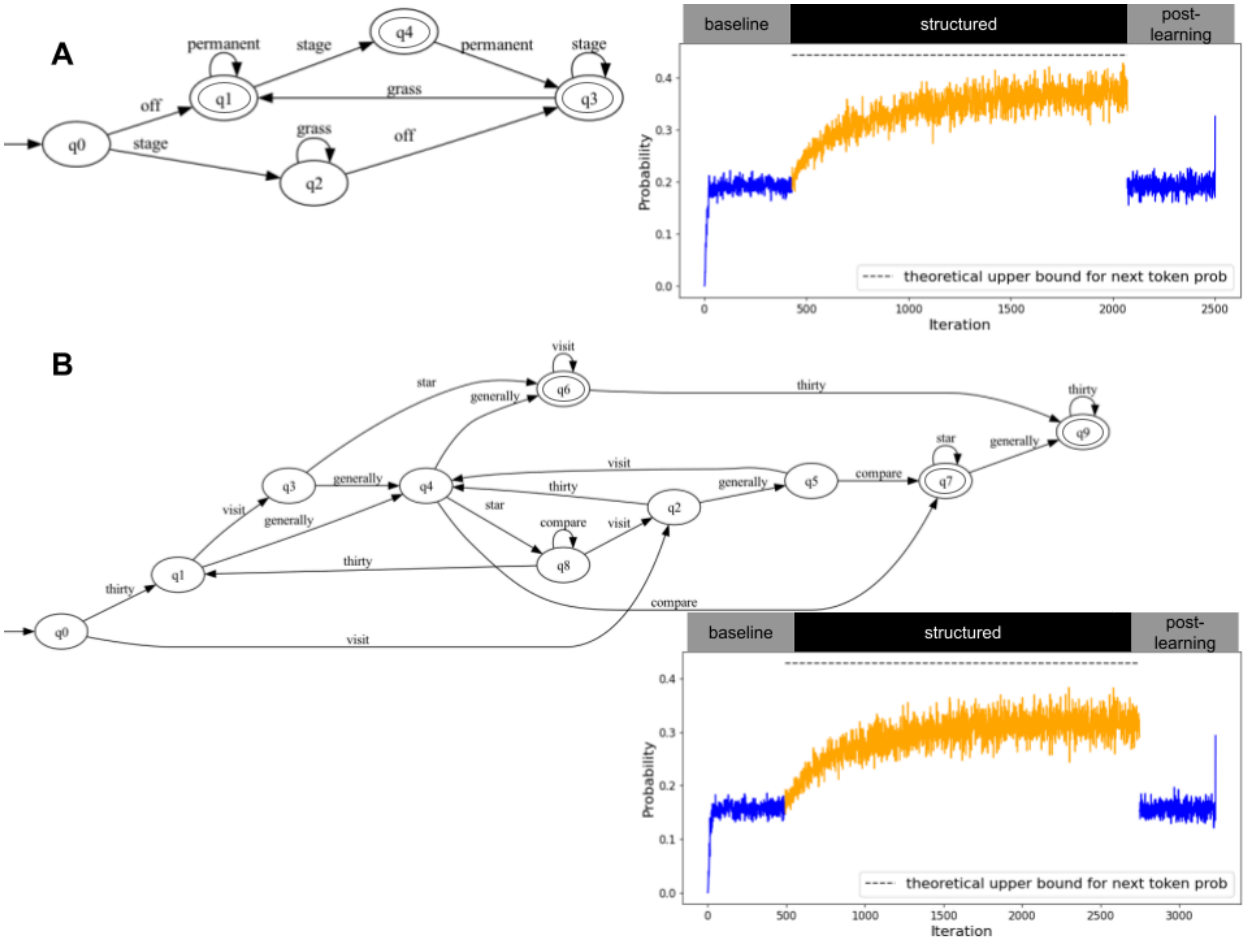

Figure S4: (A) Results of Llama 2 over DFA inspired by Knowlton & Squire (38). (B) results of Llama 2 over DFA inspired by Vokey & Brooks (39).

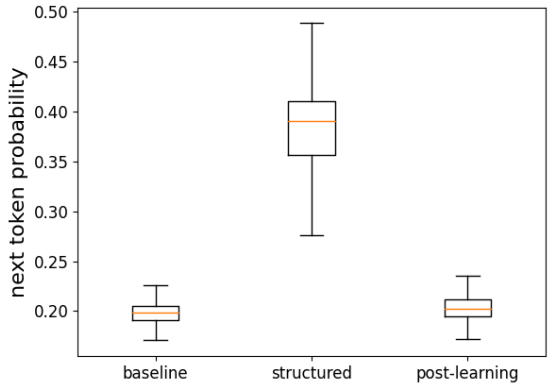

Figure S5: The box plot illustrates the distribution of predicted probabilities for each part of the experiment. We find a significant difference between structured and random parts, supported by a permutation t-test (t=135.43, p=10.0e-05).

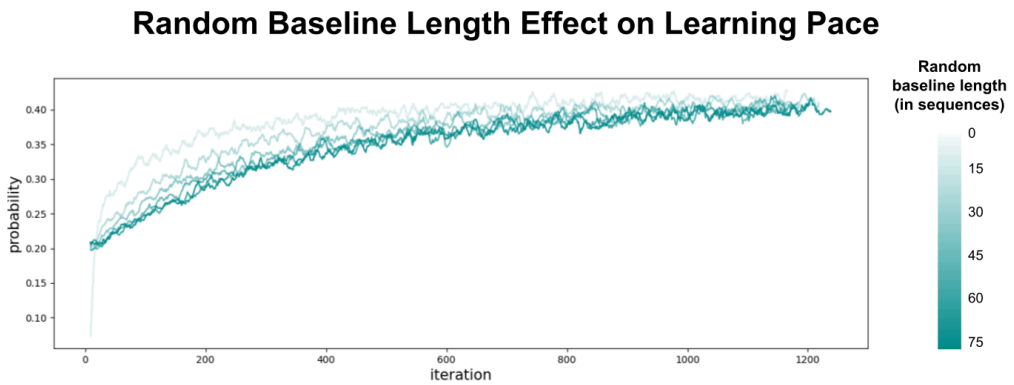

Figure S6: Learning curve for different baseline part lengths (smooth with window size=10). Light represents a shorter random baseline length before the structured part, which shows faster learning, compared to darker green, which represents a longer random baseline at the beginning.

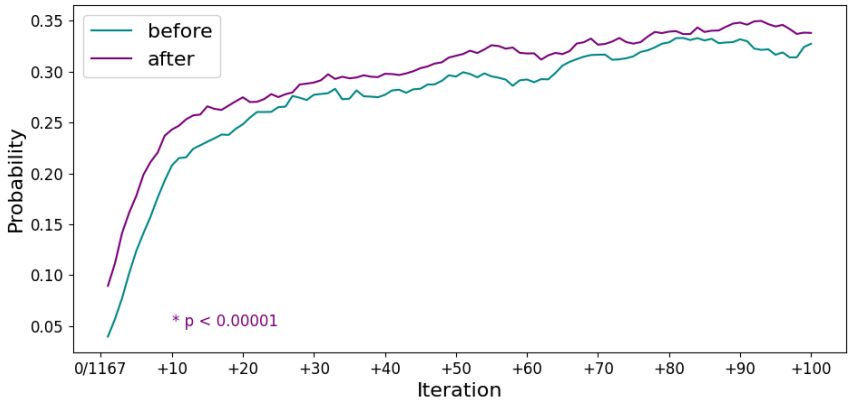

Figure S7: zoom-in over the first 100 iterations (smooth with window size=7) in a setup without any noise at the beginning, i.e., skipping the baseline part. We find that the probabilities after the switch are significantly higher than at the beginning of the structured part, even without any noise.

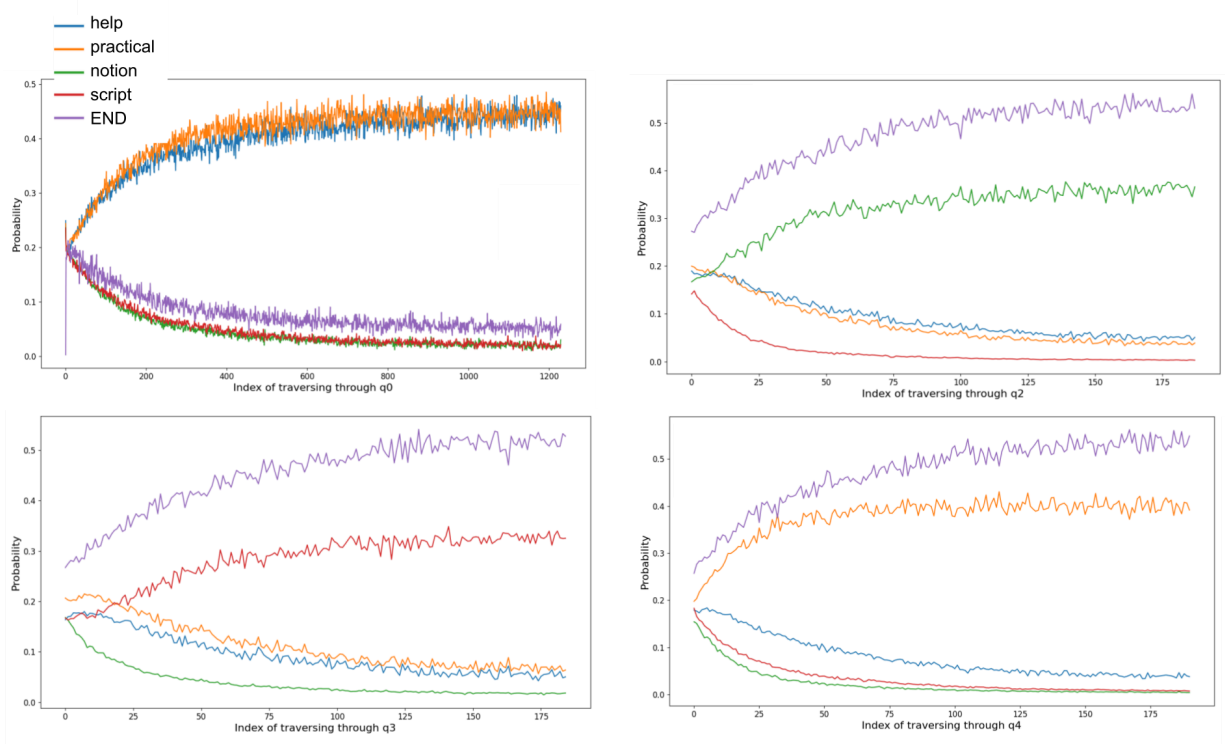

Figure S8: The model predictions for transition probabilities at each stage of the DFA, in addition to q1, are presented in Fig 4. The true transition probabilities are 0.5 for two different words at each state, as each state has exactly two outcoming edges with different labels. We see that, indeed, in all states, the model captured the two words with higher probabilities than the other words that are not the edge labels.

References

1. H. C. Ellis, The transfer of learning. 200 (1965).

2. R. von Bernhardi, L. E. Bernhardi, J. Eugenín, What is neural plasticity? Adv. Exp. Med. Biol. 1015, 1–15 (2017).

3. T. Brown, et al., Language models are few-shot learners. Adv. Neural Inf. Process. Syst. 33, 1877–1901 (2020).

4. C. M. Conway, How does the brain learn environmental structure? Ten core principles for understanding the neurocognitive mechanisms of statistical learning. Neurosci. Biobehav. Rev. 112, 279–299 (2020).

5. J. R. Saffran, R. N. Aslin, E. L. Newport, Statistical learning by 8-month-old infants. Science 274, 1926–1928 (1996).

6. A. Schapiro, N. Turk-Browne, Statistical learning. Brain mapping (2015).

7. T. F. Brady, A. Oliva, Statistical learning using real-world scenes: extracting categorical regularities without conscious intent. Psychol. Sci. 19, 678–685 (2008).

8. R. N. Aslin, E. L. Newport, Statistical learning: From acquiring specific items to forming general rules. Curr. Dir. Psychol. Sci. 21, 170–176 (2012).

9. A. S. Reber, Implicit learning of artificial grammars. Journal of Verbal Learning and Verbal Behavior 6, 855–863 (1967).

10. M. J. Nissen, P. Bullemer, Attentional requirements of learning: Evidence from performance measures. Cogn. Psychol. 19, 1–32 (1987).

11. E. M. Robertson, The serial reaction time task: implicit motor skill learning? J. Neurosci. 27, 10073–10075 (2007).

12. T. A. Forest, N. Siegelman, A. S. Finn, Attention Shifts to More Complex Structures With Experience. Psychol. Sci. 33, 2059–2072 (2022).

13. D. Kumaran, J. L. McClelland, Generalization through the recurrent interaction of episodic memories: a model of the hippocampal system. Psychol. Rev. 119, 573–616 (2012).

14. A. C. Schapiro, N. B. Turk-Browne, M. M. Botvinick, K. A. Norman, Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. Philos. Trans. R. Soc. Lond. B Biol. Sci. 372 (2017).

15. B. Sorscher, S. Ganguli, H. Sompolinsky, Neural representational geometry underlies few-shot concept learning. Proc. Natl. Acad. Sci. U. S. A. 119, e2200800119 (2022).

16. A. Q. Jiang, et al., Mistral 7B. arXiv [cs.CL] (2023).

17. E. M. Pothos, Theories of artificial grammar learning. Psychol. Bull. 133, 227–244 (2007).

18. G. T. M. Altmann, Z. Dienes, A. Goode, Modality independence of implicitly learned grammatical knowledge. J. Exp. Psychol. Learn. Mem. Cogn. 21, 899–912 (1995).

19. R. J. Tunney, G. T. Altmann, Two modes of transfer in artificial grammar learning. J. Exp. Psychol. Learn. Mem. Cogn. 27, 614–639 (2001).

20. T. Li, et al., Few-shot In-context Learning on Knowledge Base Question Answering in Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), A. Rogers, J. Boyd-Graber, N. Okazaki, Eds. (Association for Computational Linguistics, 2023), pp. 6966–6980.

21. A. Elnaggar, et al., ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. IEEE Trans. Pattern Anal. Mach. Intell. 44, 7112–7127 (2022).

22. N. Ferruz, S. Schmidt, B. Höcker, ProtGPT2 is a deep unsupervised language model for protein design. Nat. Commun. 13, 4348 (2022).

23. J. von Oswald, et al., Transformers learn in-context by gradient descent. arXiv [cs.LG] (2022).

24. D. Dai, et al., Why Can GPT Learn In-Context? Language Models Secretly Perform Gradient Descent as Meta-Optimizers in Findings of the Association for Computational Linguistics: ACL 2023, A. Rogers, J. Boyd-Graber, N. Okazaki, Eds. (Association for Computational Linguistics, 2023), pp. 4005–4019.

25. A. Mahankali, T. B. Hashimoto, T. Ma, One Step of Gradient Descent is Provably the Optimal In-Context Learner with One Layer of Linear Self-Attention. arXiv [cs.LG] (2023).

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

27. C. Caucheteux, J.-R. King, Brains and algorithms partially converge in natural language processing. Commun Biol 5, 134 (2022).

28. A. Goldstein, et al., Brain embeddings with shared geometry to artificial contextual embeddings, as a code for representing language in the human brain. bioRxiv 2022.03.01.482586 (2022).

29. A. Goldstein, et al., Shared computational principles for language processing in humans and deep language models. Nat. Neurosci. 25, 369–380 (2022).

30. A. Goldstein, et al., Correspondence between the layered structure of deep language models and temporal structure of natural language processing in the human brain. bioRxiv 2022.07.11.499562 (2023).

31. A. Georgiou, T. Can, M. Katkov, M. Tsodyks, Using large language models to study human memory for meaningful narratives. bioRxiv 2023.11.03.565484 (2023).

32. M. Geva, R. Schuster, J. Berant, O. Levy, Transformer Feed-Forward Layers Are Key-Value Memories in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, M.-F. Moens, X. Huang, L. Specia, S. W.-T. Yih, Eds. (Association for Computational Linguistics, 2021), pp. 5484–5495.

33. G. Neves, S. F. Cooke, T. V. P. Bliss, Synaptic plasticity, memory and the hippocampus: a neural network approach to causality. Nat. Rev. Neurosci. 9, 65–75 (2008).

34. T. Takeuchi, A. J. Duszkiewicz, R. G. M. Morris, The synaptic plasticity and memory hypothesis: encoding, storage and persistence. Philos. Trans. R. Soc. Lond. B Biol. Sci. 369, 20130288 (2014).

35. Y. Bengio, R. Ducharme, P. Vincent, C. Janvin, A neural probabilistic language model. J. Mach. Learn. Res. 3, 1137–1155 (2003).

36. A. Radford, et al., Language Models are Unsupervised Multitask Learners (2019).

37. H. Touvron, et al., Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv [cs.CL] (2023).

38. B. J. Knowlton, L. R. Squire, The information acquired during artificial grammar learning. J. Exp. Psychol. Learn. Mem. Cogn. 20, 79–91 (1994).

39. J. R. Vokey, L. R. Brooks, Salience of item knowledge in learning artificial grammars. J. Exp. Psychol. Learn. Mem. Cogn. 18, 328–344 (1992).