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*).

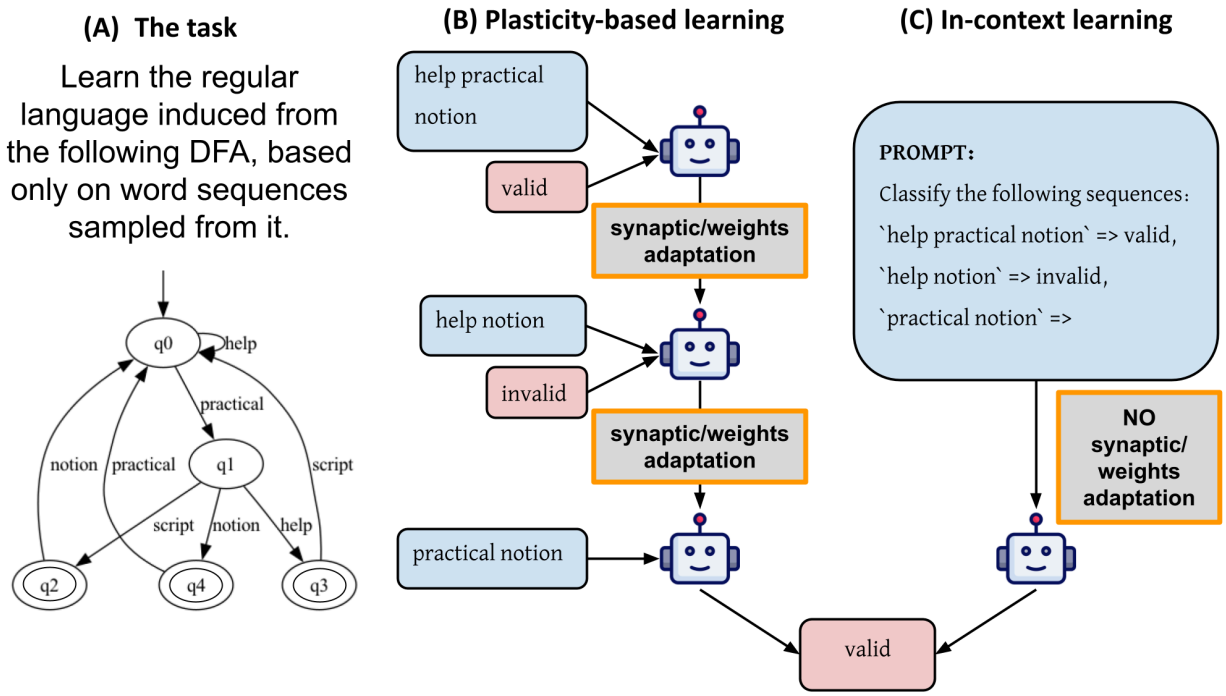

Figure 1: (A) Example of statistical learning task to identify a regular artificial language induced from a DFA. It can be learned via (B) plasticity-based learning, i.e., fine-tuning, or (C) In-context learning (ICL) paradigms. Fine-tuning involves plasticity, as the model updates its parameters following 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 statistical learning tasks: Artificial Grammar

Learning (AGL; (*9*)) and the Serial Reaction Time Task (SRTT; (*10*)). AGL involves acquiring implicit knowledge about structures in sequences and rule abstraction, while SRTT is a motor learning task that measures the reaction time to stimuli, which underlies learning the structure of stimuli sequences (*11*, *12*). These two tasks diverge from the natural language used for training LLMs. Yet, we show that LLMs can adapt to the new statistical patterns in inference without plasticity.

Several recent studies in the neuroscience literature align with our findings, albeit at a much smaller scale or requiring some limited plasticity and dedicated architectures. Kumaran & McClelland (*13*) and Schapiro et al. (*14*) showed models of the hippocampus in which generalization over bigrams of patterns can be performed in inference time without additional training. 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. Importantly, both works build dedicated models, curate special training sets, and test on similar instances to those in their training set.

In contrast, we show that off-the-shelf LLMs trained on massive amounts of natural language show generalization to complex synthetic structures without any form of explicit plasticity. Our findings can provide new insights into the potential generalization mechanisms in the brain. 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 formulate both Artificial Grammar Learning (AGL) and Serial Reaction Time Task (SRTT) as tasks of learning a deterministic finite automaton (DFA**;** Fig. 2A). We specifically choose the Mistral LLM as it is currently one of the best-performing open models (*16*), and in the supplementary materials (SM) we show that our findings hold for other off-the-shelf LLMs.

We do not perform model parameter updates in our experimental setup; we only use an *incremental input* setup. In each iteration, we perform one step of automaton graph

traversing, which produces a single word (i.e., the edge's label). This word is concatenated to the sequence of all words from previous steps, forming a gradual concatenation of words fed to the model in each iteration and congruent with the DFA (Fig. 2B). Then, we extract the probability that the model assigns to the next word (Fig. 2C). We will show model flexibility in learning the DFA without changing its inner state, but only by feeding it with a longer input congruent with the DFA regularity.

(A)

	Serial Reaction Time Task (SRTT)	Artificial Grammar Learning (AGL)
Alphabet	perspective, literally, process, editor	help, practical, script, notion
Valid sequences examples	"perspective literally process process perspective process editor literally"	"help help practical help" "practical notion" "practical script notion practical help"
Invalid sequences examples	"literally process editor perspective perspective editor process" "editor perspective editor literally" "perspective literally literally process perspective editor literally"	"notion help" "help practical script notion" "script help notion notion"

(B)

Design	Baseline (random)					structured																Post-learning (random)			
Iteration	1	2	3	4	...	...	1500	1501	1502	1503	1504	1505	1506	1507	1508	1509	1510	1511	1512	1513	...	3200	3201	3202	...
Full input	practical	notion	practical	END	...	...	practical	script	notion	practical	help	END	practical	notion	END	help	help	practical	help	END	...	notion	help	END	...
DFA path						...	q0	q1	q2	q0	q1	q3	q0	q1	q4	q0	q0	q0	q1	q0	...				
True probability for next word						...	0.5	0.33	0.5	0.5	0.33	0.5	0.5	0.33	0.5	0.5	0.5	0.5	0.33	0.5	...				
Model assigned probability						...	0.4	0.14	0.15	0.33	0.4	0.28	0.47	0.21	0.23	0.43	0.5	0.29	0.39	0.41	...				

(C)

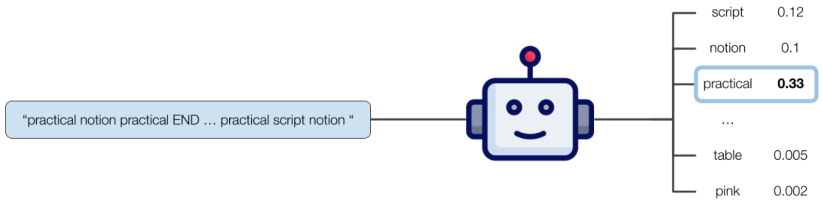

Figure 2: (A) The DFAs that simulate SRTT and AGL, with the graph alphabet and the induced valid and invalid sequences. The AGL is inspired by Pothos (*17*). (B) An input example for artificial grammar learning consists of random sequences followed by valid sequences induced from the DFA and random sequences. 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). 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). 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 random control and structured parts. For robustness, we provide results on more language models and different DFAs in Fig. S2-S3.

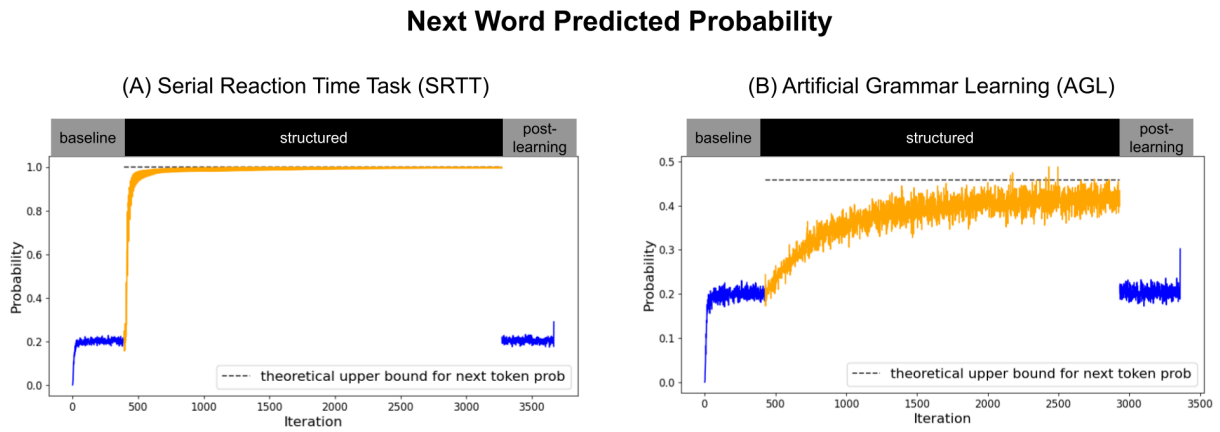

Figure 3: Next word probability for SRTT and AGL. Blue is the random control, random words in random sequences, and orange is the probability of words combining sequences induced by the corresponding DFA. The black dashed line is the theoretical upper bound of the next word probability, which for SRTT 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. 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 A in SM, materials and methods. 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 B in SM 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. S4 In the SM.

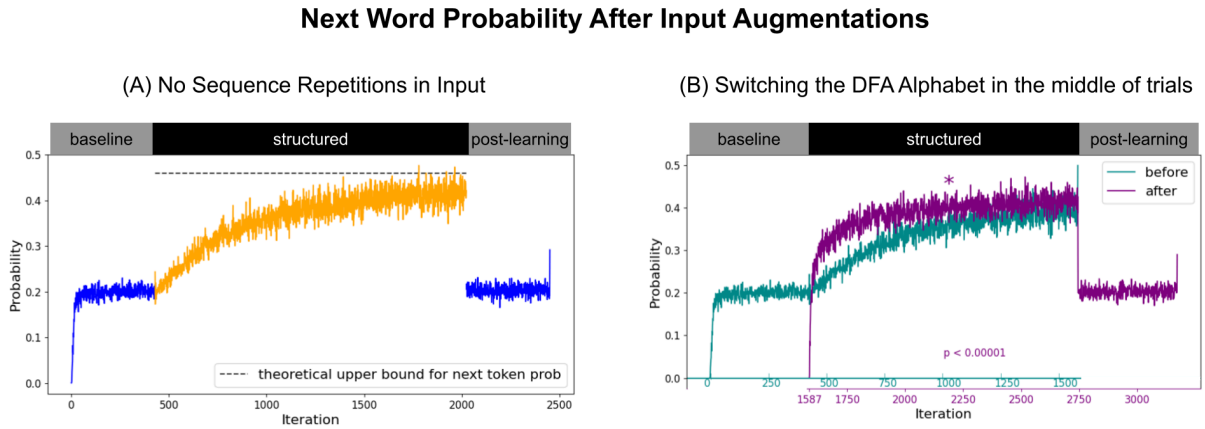

Figure 4: Next word probability for two experiments with some input augmentation. Figure A on the left presents the probability for an input with no repeating sequences, indicating that the model does not use memorization in its predictions. On the right, we present the probability of switching the DFA 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 structure of the DFA. 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. 4A, 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 DFA 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 DFA structure. In the experiment presented in Fig. 4B, 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.

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 DFA, 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 S5 of the SM. However, learning rate improvements after the vocabulary switch remain significantly pronounced, even without exposure to random sequences. Figure S6 in the SM 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.

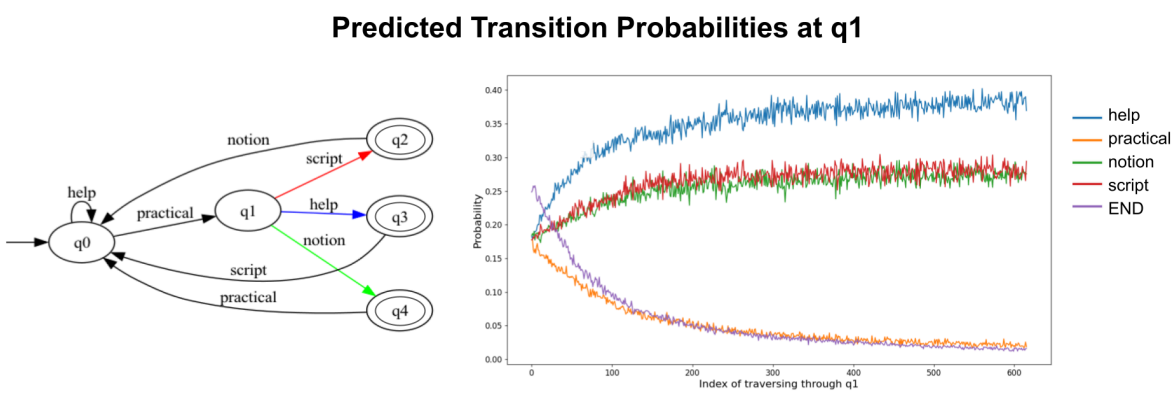

Figure 5: The model predictions for transition probabilities at q1. The true transition probabilities at q0 are 0.33 to script, help, and notion. We indeed see that the model-assigned probabilities converge around the true probabilities.

A fundamental trait of the DFA is that the model can transition to several possible next states in some states. A model that learns a DFA 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 when at q1, and in Fig. S7 in SM, we provide a similar analysis for all other DFA states. The transition probabilities at state q1 are:

$$p(script) = p(help) = p(notion) = \frac{1}{3}$$

We indeed see that the model assigns probabilities ~0.3 to each word with an edge from q1 and close to 0 probabilities for the other words, supporting the claim that the model captures the multiple possible continuations in each step, i.e., the actual DFA structure.

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. Previous work that leveraged LLMs for data that diverges from natural language, e.g., LLMs for biological encoding (*18*, *19*), usually involves fine-tuning and hence plasticity, while our work is the first to show in-context learning's efficiency for data that is fundamentally different from the model's training data. Our finding in this work offers that sufficiently trained learning systems can generalize their knowledge to completely new tasks without plasticity.

Recent studies have identified a strong resemblance between the brain's language representation and the representation generated by LLMs (*20–24*) and similar behavioral performance during memory recall tasks (*25*). In conjunction with the findings of this study, this notable similarity offers new insights into generalization mechanisms in the brain. The rich representation produced by LLMs may facilitate generalization to novel scenarios. Given the parallels between the brain and LLMs, we suggest that certain new environments, perceived as unfamiliar and necessitating learning, can be generalized without specific adaptation, thanks to extensive pre-training.

This paper suggests 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. (*26*) 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.

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. Our work opens the door for further research into the role of plasticity in learning. Also, it suggests a broader perspective that leverages deep learning models to provide novel insights into foundational concepts from the cognitive and neuroscience communities.

Supplementary Materials

Materials and Methods

A 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).

B 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.

B.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.

B.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 (Bengio, Ducharme, & Vincent, 2000; Radford, Alec, et al., 2019).

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.

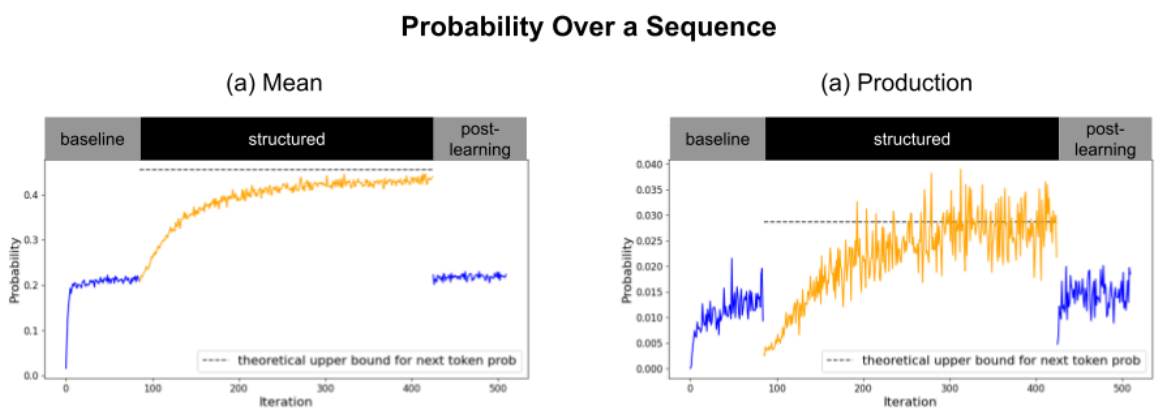

Figure S1: (a) predicted probabilities of a sequence via mean on tokens. (b) predicted probabilities of a sequence via the production of conditioned probabilities of its tokens.

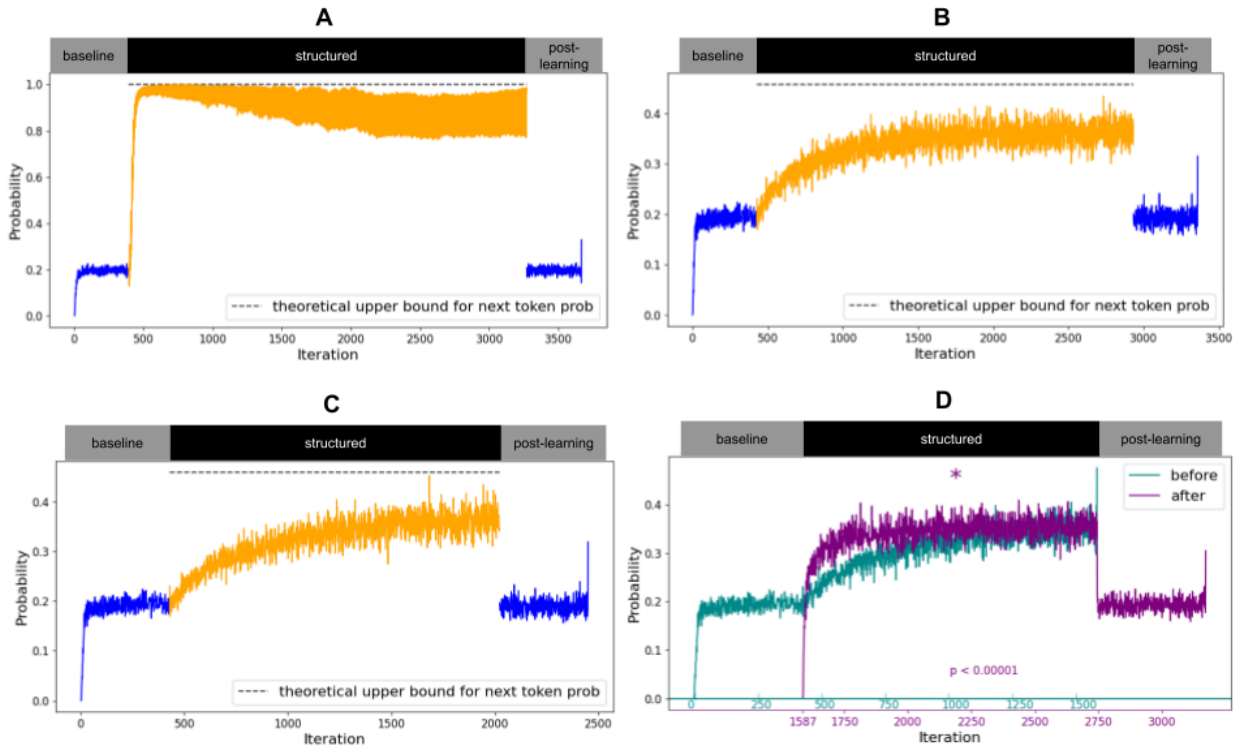

Figure S2: results of LLaMA-2 language model (Touvron, et al., 2023). (A) SRTT. (B) AGL base setup. (C) AGL switched vocabulary in the middle of the trial. (D) AGL has no duplicates.

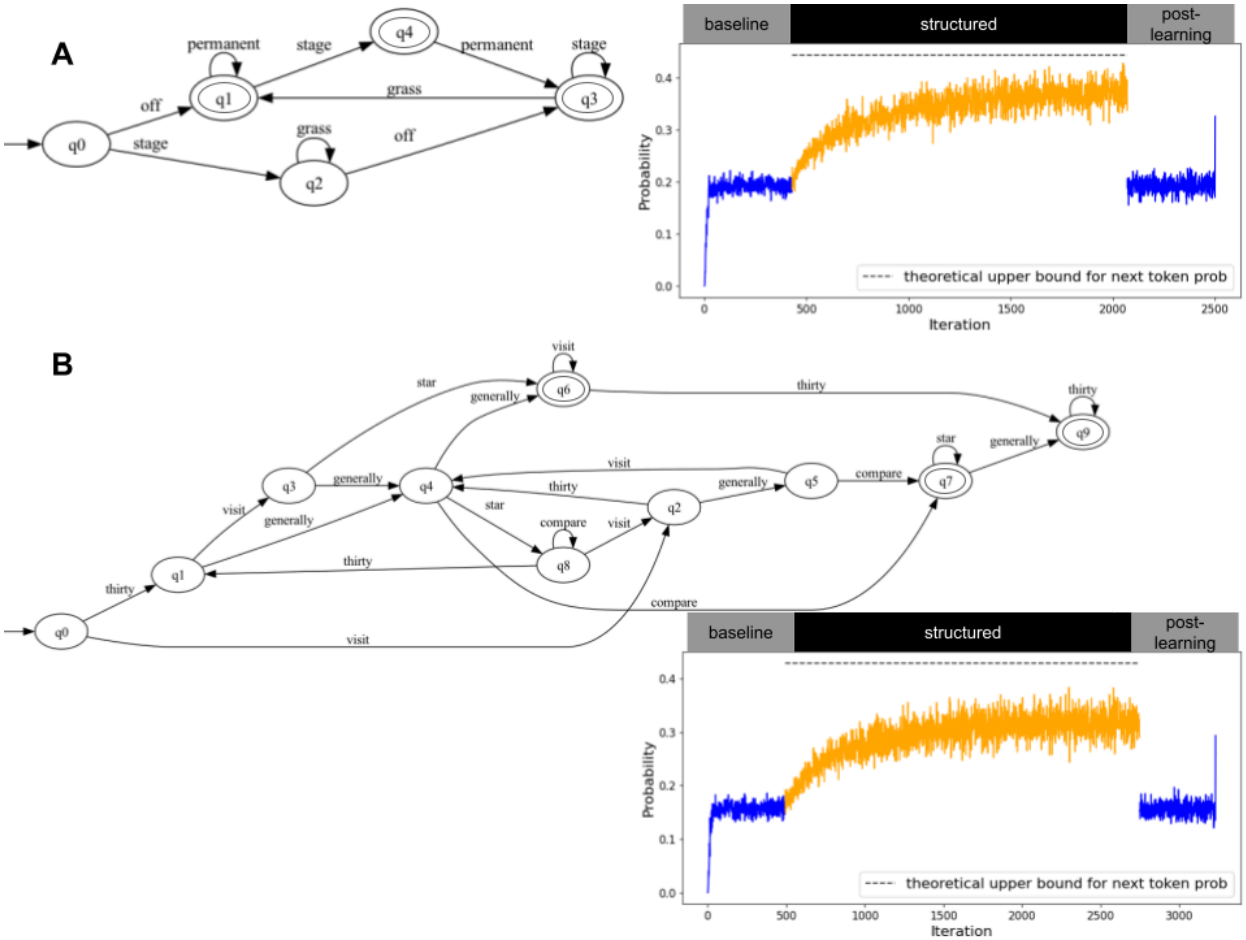

Figure S3: (A) Results of Llama 2 over DFA inspired by Knowlton & Squire, (1994). (B) results of Llama 2 over DFA inspired by Vokey & Brooks, (1992).

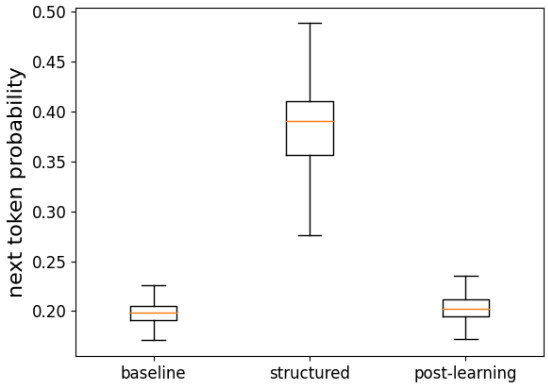

Figure S4: The box plot illustrates the distribution of predicted probabilities for each part of the experiment. We find a significant difference between structured and random parts, supported by a permutation t-test (t=135.43, p=10.0e-05).

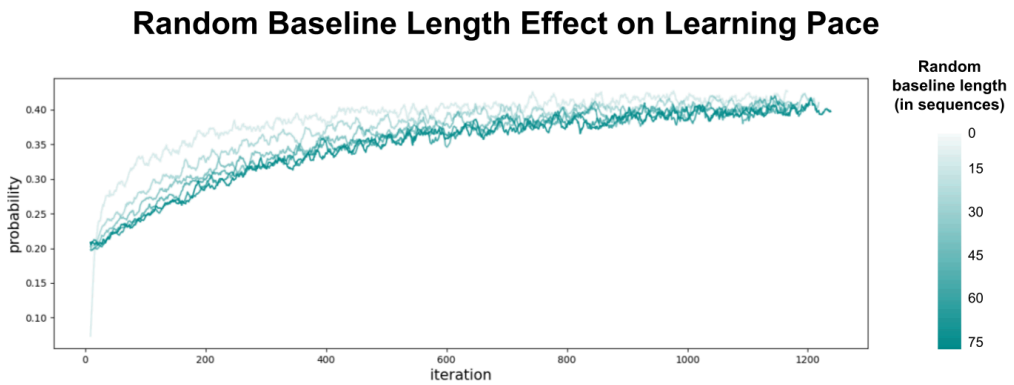

Figure S5: Learning curve for different baseline part lengths (smooth with window size=10). Light represents a shorter random baseline length before the structured part, which shows faster learning, compared to darker green, which represents a longer random baseline at the beginning.

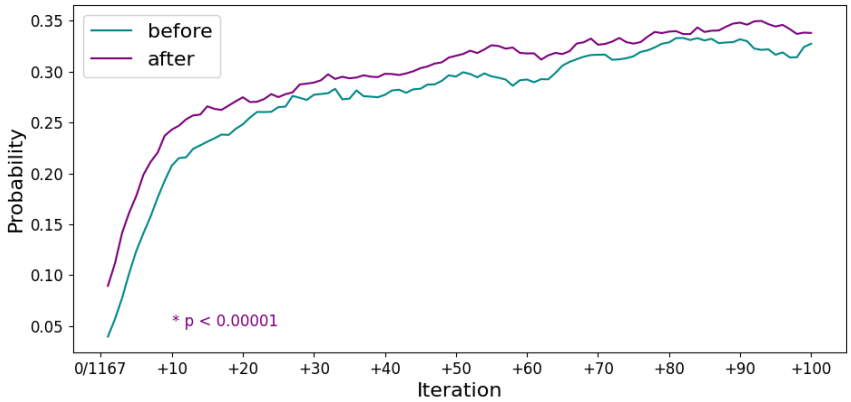

Figure S6: zoom-in over the first 100 iterations (smooth with window size=7) in a setup without any noise at the beginning, i.e., skipping the baseline part. We find that the probabilities after the switch are significantly higher than at the beginning of the structured part, even without any noise.

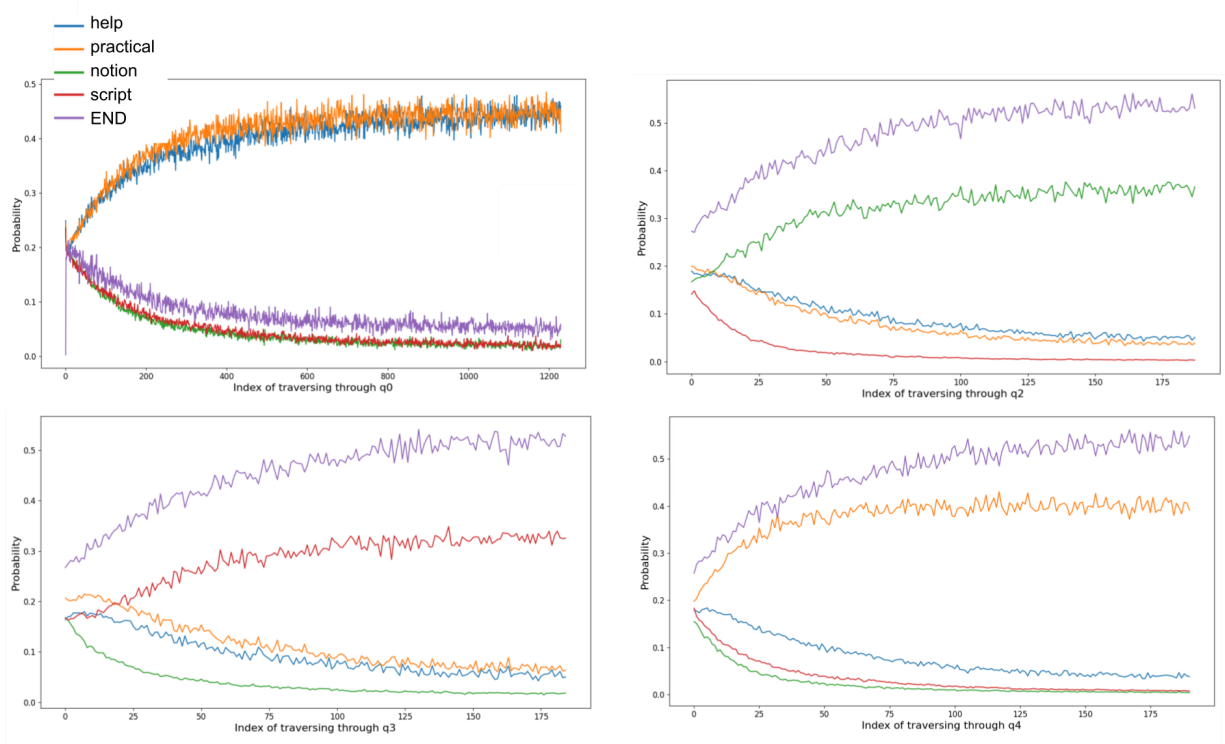

Figure S7: The model predictions for transition probabilities at each stage of the DFA, in addition to q1, are presented in Fig 4. The true transition probabilities are 0.5 for two different words at each state, as each state has exactly two outcoming edges with different labels. We see that, indeed, in all states, the model captured the two words with higher probabilities than the other words that are not the edge labels.

Bibliography

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, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, Others, 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, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. Le Scao, T. Lavril, T. Wang, T. Lacroix, W. El Sayed, Mistral 7B, *arXiv [cs.CL]* (2023). http://arxiv.org/abs/2310.06825.

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

18. A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik, B. Rost, ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 7112–7127 (2022).

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

20. M. Schrimpf, I. A. Blank, G. Tuckute, C. Kauf, E. A. Hosseini, N. Kanwisher, J. B. Tenenbaum, E. Fedorenko, The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences* 118, e2105646118 (2021).

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

22. A. Goldstein, A. Dabush, B. Aubrey, M. Schain, S. A. Nastase, Z. Zada, E. Ham, Z. Hong, A. Feder, H. Gazula, E. Buchnik, W. Doyle, S. Devore, P. Dugan, D. Friedman, M. Brenner, A. Hassidim, O. Devinsky, A. Flinker, U. Hasson, Brain embeddings with shared geometry to artificial contextual embeddings, as a code for representing language in the human brain, *bioRxiv* (2022)p. 2022.03.01.482586.

23. A. Goldstein, Z. Zada, E. Buchnik, M. Schain, A. Price, B. Aubrey, S. A. Nastase, A. Feder, D. Emanuel, A. Cohen, A. Jansen, H. Gazula, G. Choe, A. Rao, C. Kim, C. Casto, L. Fanda, W. Doyle, D. Friedman, P. Dugan, L. Melloni, R. Reichart, S. Devore, A. Flinker, L. Hasenfratz, O. Levy, A. Hassidim, M. Brenner, Y. Matias, K. A. Norman, O. Devinsky, U. Hasson, Shared computational principles for language processing in humans and deep language models. *Nat. Neurosci.* 25, 369–380 (2022).

24. A. Goldstein, E. Ham, S. A. Nastase, Z. Zada, A. Grinstein-Dabus, B. Aubrey, M. Schain, H. Gazula, A. Feder, W. Doyle, S. Devore, P. Dugan, D. Friedman, M. Brenner,

A. Hassidim, O. Devinsky, A. Flinker, O. Levy, U. Hasson, Correspondence between the layered structure of deep language models and temporal structure of natural language processing in the human brain, *bioRxiv* (2023)p. 2022.07.11.499562.

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

26. 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, Online and Punta Cana, Dominican Republic, 2021), pp. 5484–5495.