Human Curriculum Effects Emerge with In-Context Learning in Neural Networks

Jacob Russin (jake_russin@brown.edu)
Department of Computer Science
Department of Cognitive & Psychological Sciences
Brown University

Ellie Pavlick
Department of Computer Science
Brown University

Michael J. Frank
Carney Institute for Brain Science
Department of Cognitive & Psychological Sciences
Brown University

Abstract

In tasks governed by succinct rules, human learning is more robust when related examples are blocked, but in the absence of such rules, interleaving is more effective. To date, no neural model has simultaneously captured these seemingly contradictory effects. Here we show that these effects spontaneously emerge in neural networks capable of "in-context learning" (ICL). In both language models and metalearning networks, ICL explains the observed blocking advantage while concurrent in-weight learning explains the interleaving advantage.

Keywords: curriculum; language models; metalearning

Introduction

Human learning is sensitive to "curriculum" — the particular examples used to demonstrate a task and the order in which they are presented. When the task is governed by simple rules, humans benefit when related trials are blocked over time (Dekker, Otto, & Summerfield, 2022), but when it isn't, humans benefit when trials are randomly shuffled or interleaved over time (Noh, Yan, Bjork, & Maddox, 2016).

Classic neural network models of memory predict an interleaving advantage (McClelland, McNaughton, & O'Reilly, 1995), but a blocking advantage can also emerge in neural network models of gating and working memory in the prefrontal cortex (Rougier, Noelle, Braver, Cohen, & O'Reilly, 2005; Russin, Zolfaghar, Park, Boorman, & O'Reilly, 2022; Flesch, Nagy, Saxe, & Summerfield, 2022). However, no neural network model has explained how both of these effects can coexist in a single network, nor why they would depend on the presence of rule-like structure. Furthermore, previous models have been narrowly specialized to perform specific tasks, making it unclear whether their underlying principles are general enough to scale to real-world scenarios.

Language models (LMs), neural networks that are trained to predict the next word on large datasets of text (Brown et al., 2020), have recently demonstrated impressive performance on many real-world tasks (Bubeck et al., 2023). Many of these sophisticated behaviors depend on "in-context learning" (ICL): the ability to learn new tasks from a few examples given in their context window without weight changes. ICL can be differentiated from the usual in-weight learning (IWL) in neural networks, where weights are updated by backpropagating errors. ICL abilities spontaneously emerge in LMs trained on next-word prediction, but can also be acquired through metalearning (von Oswald et al., 2023), which is thought to be a key aspect of the functioning of the prefrontal cortex and basal ganglia (Wang et al., 2018; O'Reilly & Frank, 2006). In either case, the trained network can be understood as implementing an ICL algorithm in its forward activation dynamics that is fundamentally distinct from the IWL algorithm that was used to train the network in the first place (Chan et al., 2022).

These two separate ICL and IWL algorithms can have different learning properties, thus offering a novel perspective on how two different learning "systems" (Ashby & Maddox, 2011) can be implemented by a single network. We hypothesized that this distinction between ICL and IWL might explain how both the blocking and interleaving advantages could coexist within one neural network. In particular, we hypothesized that ICL, which has been shown to generalize well in the presence of rule-like or compositional structure (Lake & Baroni, 2023), would explain the blocking advantage, while IWL, which is known to suffer from catastrophic forgetting when trials are blocked, would explain the interleaving advantage. We investigated this hypothesis by testing both LMs and metalearning neural networks on a task shown in a recent experiment to elicit a blocking advantage in humans (Dekker et al., 2022).

Task Design

In the original task (Dekker et al., 2022), participants learned the 2D coordinates of reward locations associated with particular cues. Each cue was one of five animals shown in one of five colors. The reward locations were systematic, with color determining the x-coordinate and animal determining the y-coordinate or vice versa. 9 of the 25 cues were used for training, and the other 16 were used to test generalization. The experimenters manipulated the training curriculum and showed that participants generalized better with an **Aligned** curriculum than a **Misaligned** one, and better with a **Blocked** curriculum than an **Interleaved** one (see Figure 1A). However, Dekker et al. (2022) did not include a condition where the task was not governed by simple rules, where participants might instead be predicted to show an interleaving advantage. We therefore performed an additional manipulation to prevent the application of simple rules, by rotating the space of reward locations such that a change in either color or animal resulted in changes to both coordinates (see Figure 1B). Text-based versions of both the **Unrotated** and **Rotated** tasks with all four conditions were used to test our hypotheses in both pretrained LMs and metalearning neural networks (see Figure 1C).

Results

ICL in LMs

We predicted that LMs with advanced ICL abilities would show a blocking advantage on the Unrotated task, but would not be capable of solving the Rotated task in context.

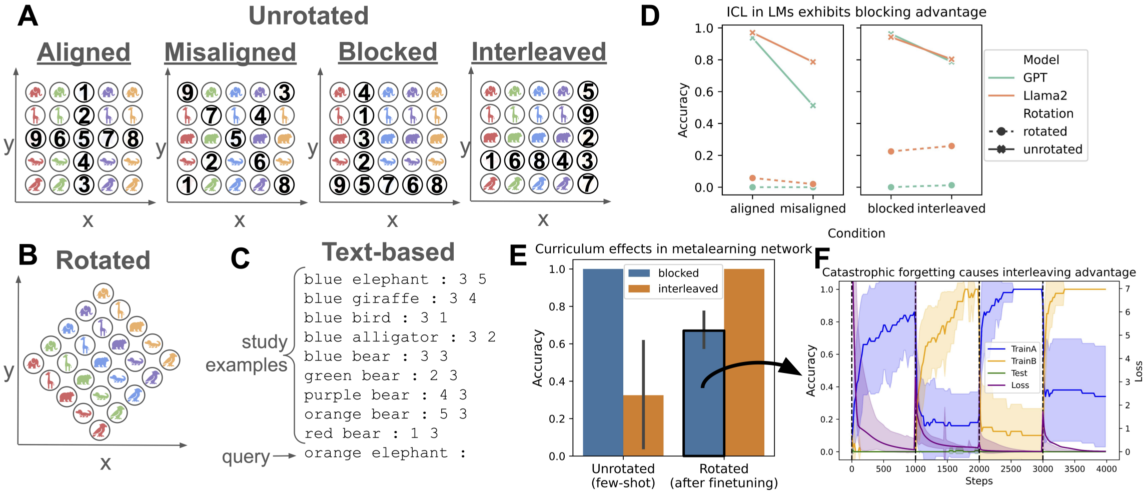

Figure 1: **(A)** Task from original experiment (Dekker et al., 2022). Cues used for training were presented in the order shown by overlaid numbers. **(B)** Rotated task. **(C)** Text-based version. **(D)** LM results. ICL explains the blocking advantage on the Unrotated task, and fails on the Rotated task. **(E)** Metalearning results. The blocking advantage was observed for few-shot generalization accuracy in the Unrotated task, while the interleaving advantage was observed after finetuning in the Rotated task. **(F)** Performance was worse after finetuning in the Rotated task when trials were blocked because IWL suffers from catastrophic forgetting: cues trained in the first block ("TrainA") are forgotten during the second block ("TrainB").

We tested two LMs, GPT-3.5 (gpt-3.5-turbo-instruct; Brown et al., 2020; Ouyang et al., 2022) and Llama 2 (70 billion; Touvron et al., 2023), on all conditions. As expected, ICL allowed both LMs to generalize well in the Unrotated task, but not in the Rotated task (see Figure 1D). Furthermore, both LMs exhibited the same blocking advantage observed in humans, generalizing better in the Aligned than the Misaligned condition, and in the Blocked than the Interleaved condition.

This pattern of results is consistent with our hypothesis that ICL would be capable of solving tasks governed by rule-like structure, and would exhibit a blocking advantage on such tasks. We also predicted that when ICL fails, IWL would become critical for solving the task, as more ICL errors would be backpropagated to the weights of the network. In this case, an interleaving advantage might result due to catastrophic forgetting (McClelland et al., 1995). As it is expensive to activate IWL in LMs for finetuning, we chose to investigate this hypothesis with smaller neural networks in a metalearning setting where both pretraining and finetuning can be controlled.

ICL and IWL in Metalearning Networks

To study the interplay between ICL and IWL, we adopted a metalearning framework where a network *learned how to in-context learn* by training on a distribution of tasks ("episodes"). New tasks were generated by randomly permuting the particular locations of the colors and animals. The network was trained on 12,000 such episodes. 100 episodes were held out for validation and 10 episodes were held out for testing.

To investigate how the blocking advantage exhibited by ICL in the LMs might interact with IWL, the metalearned ICL algorithms were trained on unrotated episodes where trials were blocked across the context window. Thus, we interpreted the metalearning stage as simulating the experiences shaping the biases present in participants coming into the experiment, but subsequent ICL and IWL on the specific task as simulating the learning they demonstrated during the experiment itself. In this latter stage, trials were either blocked or interleaved across the context window, and the metalearned network could learn in context in its forward activation dynamics, or in weights via the backpropagation of errors.

As expected, when both ICL and IWL were active within a single network, it reproduced the blocking advantage on the Unrotated task and the interleaving advantage on the Rotated task (see Figure 1E). As in the LMs, the blocking advantage observed on the Unrotated task was due to ICL, which allowed the network to generalize perfectly in the few-shot setting when trials were blocked, but struggled to do so when they were interleaved. In the Rotated task, ICL produced more errors, leading to increased IWL when these errors were backpropagated during finetuning. In this case, an interleaving advantage emerged because catastrophic forgetting occurred when trials were blocked (see Figure 1F).

Conclusion

In conclusion, our work shows that the duality between ICL and IWL offers a novel perspective on the curriculum effects observed in human learning. The blocking advantage emerges because the presence of rule-like structure allows ICL engagement, while the interleaving advantage emerges because the absence of such structure triggers IWL, which is susceptible to catastrophic forgetting.

Acknowledgments

We would like to thank all members of the Analogy Group, as well as members of the Language Understanding and Representation Lab and the Laboratory of Neural Computation and Cognition at Brown University for helpful discussions. MJF was supported by ONR grant N00014-23-1-2792. EP and JR were supported by COBRE grant #5P20GM103645-10.

References

Ashby, F. G., & Maddox, W. T. (2011). Human category learning 2.0. *Annals of the New York Academy of Sciences*, *1224*(1), 147–161. doi: 10.1111/j.1749-6632.2010.05874.x

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020, May). *Language Models are Few-Shot Learners.*

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... Zhang, Y. (2023, March). *Sparks of Artificial General Intelligence: Early experiments with GPT-4* (No. arXiv:2303.12712). arXiv.

Chan, S. C. Y., Santoro, A., Lampinen, A. K., Wang, J. X., Singh, A., Richemond, P. H., ... Hill, F. (2022, May). *Data Distributional Properties Drive Emergent In-Context Learning in Transformers* (No. arXiv:2205.05055). arXiv.

Dekker, R. B., Otto, F., & Summerfield, C. (2022, October). Curriculum learning for human compositional generalization. *Proceedings of the National Academy of Sciences*, *119*(41), e2205582119. doi: 10.1073/pnas.2205582119

Flesch, T., Nagy, D. G., Saxe, A., & Summerfield, C. (2022, March). Modelling continual learning in humans with Hebbian context gating and exponentially decaying task signals. *arXiv:2203.11560 [cs, q-bio]*.

Lake, B. M., & Baroni, M. (2023, October). Human-like systematic generalization through a meta-learning neural network. *Nature*, 1–7. doi: 10.1038/s41586-023-06668-3

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995, August). Why There Are Complementary Learning Systems in the Hippocampus and Neocortex: Insights from the Successes and Failures of Connectionist Models of Learning and Memory. *Psychological Review*, *102*(3), 419–457.

Noh, S. M., Yan, V. X., Bjork, R. A., & Maddox, W. T. (2016). Optimal sequencing during category learning: Testing a dual-learning systems perspective. *Cognition*, *155*, 23–29. doi: 10.1016/j.cognition.2016.06.007

O'Reilly, R. C., & Frank, M. J. (2006). Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation*, *18*(2), 283–328.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... Lowe, R. (2022, March). *Training language models to follow instructions with human feedback* (No. arXiv:2203.02155). arXiv.

Rougier, N. P., Noelle, D. C., Braver, T. S., Cohen, J. D., & O'Reilly, R. C. (2005, May). Prefrontal cortex and flexible cognitive control: Rules without symbols. *Proceedings of the National Academy of Sciences*, *102*(20), 7338–7343. doi: 10.1073/pnas.0502455102

Russin, J., Zolfaghar, M., Park, S. A., Boorman, E., & O'Reilly, R. C. (2022, February). A Neural Network Model of Continual Learning with Cognitive Control. In *Proceedings for the 44th Annual Meeting of the Cognitive Science Society.*

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... Scialom, T. (2023, July). *Llama 2: Open Foundation and Fine-Tuned Chat Models* (No. arXiv:2307.09288). arXiv. doi: 10.48550/arXiv.2307.09288

von Oswald, J., Niklasson, E., Schlegel, M., Kobayashi, S., Zucchet, N., Scherrer, N., ... Sacramento, J. (2023, September). *Uncovering mesa-optimization algorithms in Transformers* (No. arXiv:2309.05858). arXiv. doi: 10.48550/arXiv.2309.05858

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., ... Botvinick, M. (2018, June). Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, *21*(6), 860–868. doi: 10.1038/s41593-018-0147-8