PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES
COMPUTER SCIENCES

Parallel trade-offs in human cognition and neural networks: The dynamic interplay between in-context and in-weight learning

Jacob Russin[a,b,1], Ellie Pavlick[a,2], and Michael J. Frank[b,c,2]

Edited by Richard Shiffrin, Indiana University Bloomington, Bloomington, IN; received April 30, 2025; accepted July 14, 2025

Human learning embodies a striking duality: Sometimes, we can rapidly infer and compose logical rules, benefiting from structured curricula (e.g., in formal education), while other times, we rely on an incremental approach or trial-and-error, learning better from curricula that are randomly interleaved. Influential psychological theories explain this seemingly conflicting behavioral evidence by positing two qualitatively different learning systems—one for rapid, rule-based inferences (e.g., in working memory) and another for slow, incremental adaptation (e.g., in long-term and procedural memory). It remains unclear how to reconcile such theories with neural networks, which learn via incremental weight updates and are thus a natural model for the latter, but are not obviously compatible with the former. However, recent evidence suggests that metalearning neural networks and large language models are capable of in-context learning (ICL)—the ability to flexibly infer the structure of a new task from a few examples. In contrast to standard in-weight learning (IWL), which is analogous to synaptic change, ICL is more naturally linked to activation-based dynamics thought to underlie working memory in humans. Here, we show that the interplay between ICL and IWL naturally ties together a broad range of learning phenomena observed in humans, including curriculum effects on category-learning tasks, compositionality, and a trade-off between flexibility and retention in brain and behavior. Our work shows how emergent ICL can equip neural networks with fundamentally different learning properties that can coexist with their native IWL, thus offering an integrative perspective on dual-process theories of human cognition.

neural networks | in-context learning | cognitive flexibility | curriculum effects | compositionality

Significance

The neural networks dominating AI in recent years have achieved a remarkable level of behavioral flexibility, in part due to their capacity to learn new tasks from only a few examples. These in-context learning abilities are analogous to human inference and have different properties than the usual in-weight learning. Here, we show that when both are present in a single network, their dynamic interplay can tie together several key learning phenomena observed in humans, including curriculum effects, compositional generalization, and a trade-off between flexibility and retention. Our findings highlight important computational principles that may be shared between human and artificial intelligence.

Humans are capable of two qualitatively distinct kinds of learning (1–9). The first involves slow, incremental adaptation and storage in long-term or procedural memory (2, 6, 9–12). The second is much more advanced and involves rapid inference of rules or structure from information available in the environment or held in working memory (WM; 10, 13–18). For example, when learning chess, it can take years to develop the subtle, tacit intuitions required to evaluate complex positions, though initially one may make rapid inferences about how the pieces move.

Many findings support the idea that humans exhibit different learning and generalization behaviors in different domains (1, 5, 9, 10, 14, 19–21). On the one hand, in tasks that are readily described by simple rules (e.g., inferring how the knight moves in chess), humans learn efficiently from only a few examples ("few-shot learning"), appearing to make rapid inferences about the latent structure governing the task (15, 22, 23). When this latent structure is compositional, humans can generalize by flexibly recombining familiar elements according to inferred rules (24–32). In such settings, people exhibit a blocking advantage, learning better when information is organized into blocks of related examples that make this underlying structure more salient (5, 19, 24, 33). On the other hand, when a task is not governed by simple rules, learning may require integrating across multiple task dimensions. This kind of learning proceeds more incrementally (1, 5, 34), but can also be associated with greater retention after a delay (10, 21, 35, 36). In these contexts, compositional generalization is not possible, and, as shown in both laboratory (5, 37) and real-world contexts (38, 39), people exhibit an interleaving advantage, learning better when trials are randomly shuffled over time.

The traditional way of explaining how such contrasting effects can arise in different learning contexts is to posit two separate systems with qualitatively different properties. Prominent dual-process accounts (1, 4, 5, 40) hypothesize that when a task can be solved by inferring simple rules, a symbolic or rule-based system is deployed, whereas in the absence of such simple rules, a procedural or subsymbolic system is recruited. However, recent findings in machine learning have shown that a single neural network,

Author affiliations: [a]Department of Computer Science, Brown University, Providence, RI 02906; [b]Department of Cognitive and Psychological Sciences, Brown University, Providence, RI 02912; and [c]Carney Institute for Brain Science, Brown University, Providence, RI 02906

Author contributions: J.R., E.P., and M.J.F. designed research; J.R. performed research; J.R. analyzed data; and J.R., E.P., and M.J.F. wrote the paper.

Competing interest statement: E.P. is a paid consultant for Google (unrelated to this work).

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: jake_russin@brown.edu.

[2]E.P. and M.J.F. contributed equally to this work.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2510270122/-/DCSupplemental.

Published August 28, 2025.

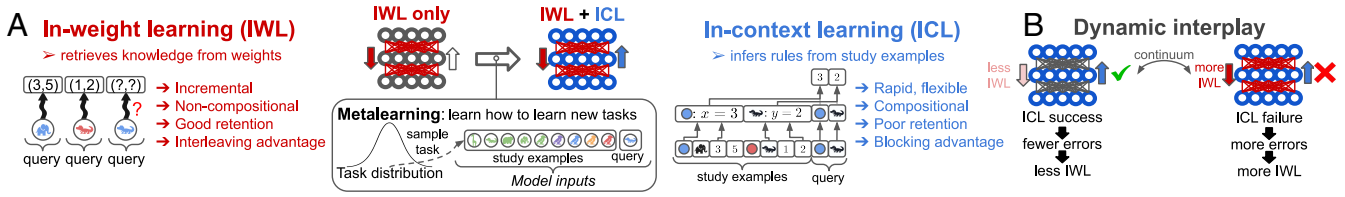

Fig. 1. (*A*) Properties of in-context learning (ICL) and in-weight learning (IWL). IWL (red) is the usual form of learning in neural networks, wherein errors are backpropagated to update weights. IWL can support better retention but is noncompositional, depicted here as failing to generalize on a compositional task (Fig. 3) where color and animal features determine x and y coordinates, respectively. IWL exhibits an interleaving advantage, learning better when examples are randomly shuffled or interleaved. Metalearning over a distribution of tasks allows a neural network to *learn how to learn* new tasks from just a few examples given in context (46). Once it emerges, ICL (blue) is carried out through activation dynamics (blue arrow) and can have different learning properties than those realized in IWL. For example, ICL can be flexible and compositional, and is shown here leveraging the attention mechanism of the transformer to compose rules inferred from the study examples. ICL can also exhibit a blocking advantage, learning better when related examples are blocked over time. (*B*) Dynamic interplay between ICL and IWL. When ICL is successful, fewer errors are accumulated and therefore less IWL occurs.

whose default learning occurs via incremental weight updates or in-weight learning (IWL; 41, 42), can through extensive training on a variety of tasks acquire an emergent capacity for in-context learning (ICL; 43–47). ICL is the ability to flexibly adapt to the rules of new tasks from a few demonstrations or instructions provided in context. For example, if a novel task is demonstrated with contextual inputs (strawberry → red, banana → yellow), trained networks such as large language models (LLMs) often readily generalize to new inputs (plum → ??). This kind of learning can also be understood as a form of *inference* and does not require additional weight updates but can occur within the flow of information from inputs to outputs through the network's activation dynamics (43, 44, 48, 49). This is similar to how neural network models of the prefrontal cortex (PFC) utilize their activation-based dynamics to make inferences and flexibly adapt to the current context (13, 18, 50–52). While much work has focused on the factors that drive the emergence of ICL (e.g., 44), less emphasis has been placed on how it can interact with the usual IWL. Here, we consider how the interplay between these learning mechanisms can explain a variety of phenomena in human cognition, exploring how they may account for various findings motivating traditional dual-process theories.

IWL operates by backpropagating errors to incrementally update weights and is a natural model of procedural or subsymbolic learning (41, 42). Yet because neural networks have traditionally relied exclusively on IWL, they have often failed to explain key aspects of rule-based human cognition. Indeed, neural networks are notoriously data hungry compared to human learners (15) and have been criticized for failing to account for human few-shot learning and compositionality (25, 29, 53–55), as they do not explicitly represent symbols or infer rules (15, 56, 57). Furthermore, in contrast to the blocking advantage that humans exhibit in rule-based tasks (5, 19), neural networks suffer from "catastrophic forgetting" in these scenarios because new IWL can overwrite information stored in the same weights during previous blocks (58–60).

ICL offers a natural framework for understanding how a neural network can acquire such qualitatively different learning properties. The ICL abilities that emerge in LLMs have led to surprising success on rule-governed tasks involving reasoning (43, 61, 62), analogy (63, 64), and compositionality (49, 65, 66). However, ICL does not come for free, but emerges only after extensive training: The usual IWL mechanisms, when applied over a large and diverse dataset, support a form of "metalearning," in which the network learns to efficiently utilize its activation dynamics to flexibly adapt to new scenarios (see Fig. 1; 48, 67–71). In other words, over the course of learning many different tasks, these networks *learn how to learn* new tasks efficiently from a few examples provided as inputs ("in context"). Such metalearning can reproduce human-like compositional generalizations (31, 49, 72) and can also induce neural network models of the PFC to acquire rule-based inference abilities (13, 18, 71).

Once ICL has emerged, the network's learning behavior will not be exclusively determined by ICL or IWL but will typically lie on a continuum between them. A natural trade-off governs this relationship (44, 73, 74): When ICL succeeds, fewer errors accumulate, resulting in fewer weight updates and therefore less IWL (Fig. 1*B*). We therefore hypothesized that when ICL was effective, the network would exhibit its metalearned properties, but when ICL made errors, the resulting weight updates would expose the network's default IWL behavior. Notably, cognitive neuroscience studies have demonstrated a related trade-off in human learning. When information can be learned rapidly within working memory, neural prediction errors are suppressed (9, 14, 75). This neural signature predicts enhanced generalization of rule-like structure (14) but *degraded* retention when information is no longer available in working memory (21, 35, 36). Thus, the dynamic interaction between ICL and IWL may allow a single neural network to express different learning properties in different situations, while simultaneously reproducing a trade-off observed in human learning.

In this work, we demonstrate how a single neural network capable of both ICL and IWL can replicate the behavioral effects associated with the two systems posited in traditional dual-process theories (1, 4, 5), producing compositional generalizations and a blocking advantage in rule-based tasks, while exhibiting an interleaving advantage in tasks lacking simple rules. Moreover, we show how the very same mechanisms give rise to the trade-off between flexibility and retention observed in human learning (14, 21, 36). Our theoretical framework comprises three key principles, summarized in Table 1.

We test this framework with metalearning transformer neural networks (76) trained on tasks from human studies (5, 9, 24). A distinguishing feature of the transformer is its "attention heads," which isolate and process subsets of the inputs and can learn which inputs are relevant for particular predictions. For example, when given the above task (plum → ??) a trained transformer may use its attention heads to focus on the fruits and colors provided in context, and subsequently constrain its activations to access color-relevant information (77, 78). Transformers develop ICL more reliably compared to traditional neural networks (44), perhaps because attention provides a natural way for contextual information to dynamically steer representations according to the context (e.g., "plum" ⇝ "purple"). As such, this mechanism resembles those in biological models of PFC,

Table 1. Theoretical framework

(1) Properties of IWL	IWL fails on compositional generalization problems, shows an interleaving advantage due to catastrophic forgetting when trials are blocked, and results in better retention.
(2) Properties of ICL	ICL can generalize compositionally and show a blocking advantage, but results in worse retention.
(3) Dynamic interplay	When ICL is possible, its properties dominate because few errors are made, suppressing IWL. But when ICL is difficult, the properties of IWL dominate because errors result in larger weight updates.

where activation-based dynamics can guide processing toward task-relevant information (13, 18). Although much remains unknown about the fine-grained computations carried out during ICL, it is clear that in transformers it must rely at least partially on the attention mechanism (77–79). However, we note that the basic principles governing the interplay between ICL and IWL (Fig. 1*B*) should apply to any network architecture.

Our experiments (80) show how this dynamic interplay offers a unified account of learning phenomena observed in humans across a wide range of studies from cognitive psychology and neuroscience.* First, we show in a category-learning setting (5) that, like humans, a single neural network capable of ICL and IWL produces a blocking advantage on rule-like tasks, and an interleaving advantage in the absence of simple rules. Second, we show that the same neural network can also produce compositional generalizations and their associated blocking advantage (24). Third, we test LLMs on this same compositional task without further training and show that ICL in these models exhibits both compositionality and a blocking advantage. Finally, we show how ICL depends on attention in our models through causal manipulations, demonstrating how the dynamic interplay between ICL and IWL naturally gives rise to the trade-off between flexibility and retention observed in recent human studies (9, 21, 35, 36).

The primary goal of this work is not simply to show that neural networks can perform well on these cognitive tasks but to demonstrate how the specific principles governing the dynamic interplay between ICL and IWL naturally reproduce learning phenomena observed in human studies. Our findings show how these two qualitatively distinct learning processes can interact within a single neural network model, and offer a framework for reconciling dual-process theories of cognition with a neural network perspective.

Results

Curriculum Effects in Category Learning. We first consider whether the principles above can account for the curriculum effects observed in human category learning, before turning to compositionality in the next section. As summarized above, human category learning exhibits an interaction ($\eta_p = 0.04$) between whether a task is rule-like, and the curriculum (blocked vs. interleaved), showing a blocking advantage ($d = 0.47$) when categories are governed by succinct rules, but an interleaving advantage ($d = 0.33$) otherwise (5).

We designed a category-learning task based directly on this previous work (5), but suitable for use with metalearning neural networks (Fig. 2 *A*–*C*). Stimuli varied along two feature dimensions (akin to line length and line orientation) with 8 possible values, yielding 64 possible items. Each item was assigned to one of two categories, indicated by an arbitrary category label (e.g., "A" or "B"). In the **Rule-like** (or "Rule-based") condition, one of the two feature dimensions determined category membership (e.g., lines with shorter lengths are in category "A" and lines with longer lengths are in category "B"), while in the **Rotated** (or "information-integration") condition, category membership was determined by both features. This simple rotation has been shown to challenge the search for a simple, verbalizable rule, and is thought to recruit the more incremental procedural learning system in humans (1, 5). Networks were presented with 16 items from each category (32 total), and tested on the remaining held-out items. The 32 items used during learning were either **Blocked**, where items from one category were presented first, followed by the items from the other, or **Interleaved**, where items were randomly shuffled. Both rotation conditions were tested with both curriculum conditions, yielding a 2 × 2 design.

IWL produces an interleaving advantage. To isolate the learning properties of IWL in this category-learning setting, a randomly initialized transformer was trained from scratch via the usual IWL in each of the four conditions (see *Materials and Methods* for details). Because IWL requires slow, incremental updates, this network was not capable of few-shot learning (Fig. 2*D*) even in the rule-like condition, where a few examples should suffice for inference of the simple rule. Consistent with our theoretical framework (principle 1), the model performed better when trials were interleaved compared to blocked ($P < 10^{-3}$; see Fig. 2 *E* and *F*; see *SI Appendix* for details about all statistical testing), in both the rule-like and rotated conditions (although slightly better in the rule-like condition). This interleaving advantage was due to catastrophic forgetting when trials were blocked (59, 60), which can be seen in the dramatic decrease in accuracy on examples of the category trained during the previous block (e.g., accuracy on category "A" train trials decreases as category "B" is trained in the second block). Thus, the default IWL behavior of neural networks can explain why an interleaving advantage would be observed in human category learning (5). However, a network capable of IWL alone cannot account for the blocking advantage that humans exhibit when categories are governed by simple rules (5, 19, 24).

ICL can produce a blocking advantage. Next, we endowed a transformer with ICL abilities by having it metalearn on a distribution of categorization tasks (see *Materials and Methods* for details), analogous to the way humans improve at rapidly learning new tasks throughout development. These ICL abilities allowed the network to solve unseen tasks given in context, even when weights were frozen and no IWL was allowed to occur. Rather, ICL was accomplished via activation dynamics enabled by attention to contextual information, which we confirm below with causal manipulations (Fig. 5).

To ensure that ICL would have the desired properties (see principle 2), we had the network metalearn on a distribution of categorization tasks with 1) rule-like structure and 2) blocked curricula. To isolate these learning properties of ICL, we evaluated the network in the few-shot setting, where the weights were frozen and the network had to learn new tasks from examples given in context (see *Materials and Methods* for details; see red vs. blue lines in Fig. 2*G*). As predicted, when the model had developed ICL

*All code used for simulations is available at: https://github.com/jlrussin/icl-iwl-interplay.

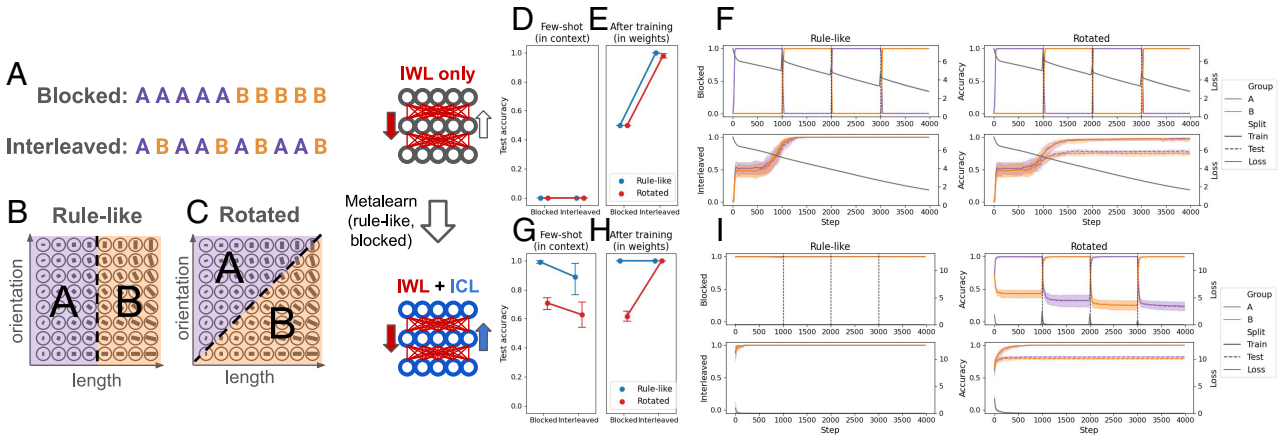

Fig. 2. Category-learning experiments. The task is derived from a human study (5). Transformer neural networks were presented with multifeature items along with their category labels and tested on unseen items. (*A*) Curriculum conditions. Trials were either blocked by category or randomly interleaved. (*B*) In the rule-like condition, category membership was determined by a simple rule based on only one of the two features. (*C*) In the rotated (information-integration) condition, category membership was jointly determined by both features. The original axes were rotated by 45° and a category boundary was chosen in the new coordinate system. (*D–F*) Category-learning with in-weight learning (IWL) only. Randomly initialized networks were trained from scratch on the task. (*D*) The few-shot evaluation tested networks' ability to learn the task from the 32 examples presented in context, before any weight updates were made. Unsurprisingly, networks without prior metalearning experience were incapable of utilizing examples given in context to learn the task. Note that model choices were not constrained to the two category labels, so chance performance here corresponds to $1/d_V$, where d_V is the vocabulary size. (*E*) Without prior metalearning, the network was able to learn via IWL, performing well on both the rule-like and rotated tasks after training. However, performance was much worse in the blocked condition due to catastrophic forgetting. Values correspond to accuracy on the 32 train items. (*F*) Accuracy and loss results over the course of task-specific training. Accuracy is split by category. (*G–I*) Category learning with both IWL and in-context learning (ICL). Randomly initialized networks metalearned on a distribution of rule-like tasks with blocked curricula and were subsequently trained on specific category-learning tasks. (*G*) After metalearning, the models exhibited strong ICL, as shown by high few-shot test accuracy. ICL exhibited a blocking advantage and also showed improved performance in the rule-like compared to the rotated condition. (*H*) After training had occurred on a specific task, the network exhibited an interleaving advantage in the rotated condition, due to catastrophic forgetting when trials were blocked. (*I*) Accuracy and loss results over the course of task-specific training. When trials were blocked in the rule-like condition, ICL achieved near-perfect accuracy, resulting in little loss and thus little IWL. When trials were interleaved, few-shot test accuracy was worse, but performance quickly recovered due to compensation by IWL. In the rotated condition, ICL failed, resulting in larger losses and increased IWL. This IWL resulted in catastrophic forgetting, as can be seen in the rapid decline in train accuracy on "A" items while training on "B," and vice versa. No such catastrophic forgetting occurred when trials were interleaved (although test performance was not perfect).

abilities by metalearning on rule-like category-learning problems, it could generalize to held-out test items on new rule-like tasks, but struggled to solve rotated tasks in context (main effect of rotation on test accuracy: $P < 10^{-3}$). Moreover, ICL exhibited a blocking advantage on unseen rule-like categorization tasks (main effect of curriculum on test accuracy: $P < 0.05$). This blocking advantage also emerged due to the fact that trials were blocked during metalearning (*SI Appendix*), but see *Discussion* for alternative explanations based on architectural constraints in human brains. In sum, these few-shot results suggest that it is possible to endow a network with ICL abilities that are sensitive to rule-like structure and learning curriculum: The network was capable of making inferences over the items provided in context, but was better at doing so when related items were organized into blocks.

Concurrent ICL and IWL reproduce both curriculum effects. While the above explorations showed how IWL and ICL can produce different curriculum effects, we are now in a position to study how the two might interact in a single system capable of both. To do this, we took the network that developed ICL abilities through metalearning, and gave it unseen category-learning tasks, allowing it to learn by ICL (via forward activation dynamics) and IWL (via error backpropagation). Here, we predicted that the dynamic interaction between IWL and ICL would qualitatively reproduce the full set of curriculum effects observed in the original study (5): ICL would produce the blocking advantage in the presence of rule-like structure, while IWL would produce the interleaving advantage in the absence of such structure (see principle 3).

As we described above, when categories were governed by a simple rule, ICL succeeded and exhibited a blocking advantage on test trials in few-shot inference. But in the rotated task, where categories were not governed by simple rules, ICL struggled (Fig. 2*G*). The resulting errors drove an increase in IWL, producing an interleaving advantage due to catastrophic forgetting (Fig. 2*I*; interaction between curriculum and rotation on train accuracy: $P < 10^{-3}$).

Taken together, these experiments show that a single model capable of ICL and IWL can recapitulate the curriculum effects observed in human category learning (5). When the network is capable of making inferences over familiar rules, it can solve new tasks from a few examples given in context. However, when the environment does not afford such inferences or the network cannot make them, IWL can still compensate, allowing good performance. This IWL suffers from catastrophic forgetting, resulting in an interleaving advantage on the rotated task.

Curriculum Effects in a Compositional Task. As noted above, one of the most impressive recent developments in research on neural networks has been the demonstration that ICL can give rise to compositionality (49, 66, 72), traditionally considered to be a major theoretical challenge to neural networks (25, 54). Recent studies have shown that while standard IWL struggles to reproduce human-like compositional generalization behaviors (53, 81, 82), ICL can appear to compose inferred rules in order to generalize to new inputs (49, 61, 65, 66). Thus, a key goal of our framework is to leverage the distinction between ICL and IWL to provide a unified account of both the compositional generalization behaviors and the curriculum effects observed in humans. In particular, ICL should account for both the blocking advantage and the compositional generalization behaviors observed in tasks governed by rule-like structure, while IWL accounts for the interleaving advantage ob-

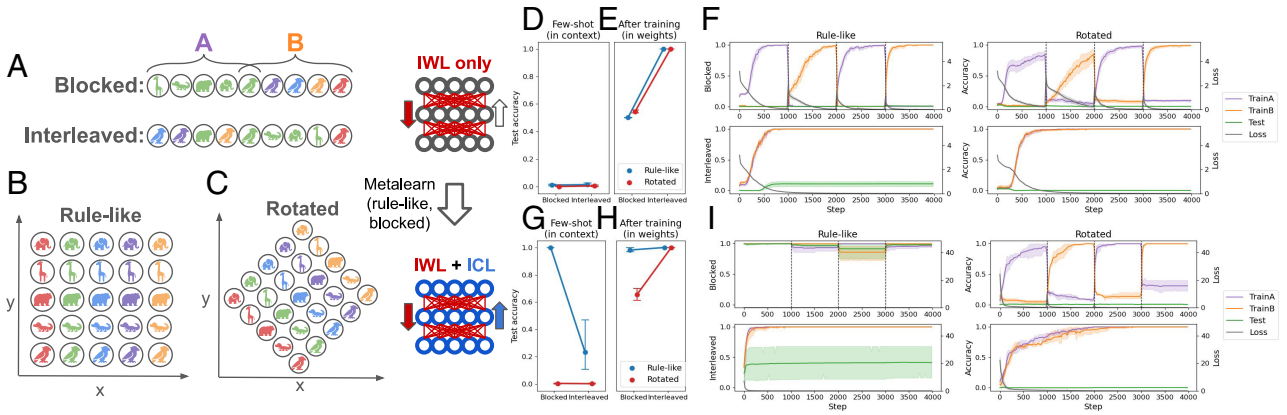

Fig. 3. Compositional task and results. The task is derived from a human study (24). Transformer networks were presented with the locations corresponding to particular cues (colored animals) and had to predict the locations of unseen cues. (*A*) Curriculum conditions. In both the blocked and interleaved conditions, the 9 study examples always included one full row and one full column. In the blocked condition, the row was presented before the column, or vice-versa. In the interleaved condition, the 9 examples were randomly shuffled. (*B*) In the rule-like condition, locations varied systematically with color/animal. (*C*) In our novel rotated condition, the original axes were rotated by 45°, so that any change in either color or animal resulted in a change to both coordinates. (*D–F*) Performance on the compositional task with in-weight learning (IWL) only, where networks were trained from scratch. (*D*) The few-shot evaluation tested networks' ability to solve the task in context based on the 9 study examples. Without prior metalearning, neural networks were incapable of solving the task, regardless of condition. (*E*) However, even without prior metalearning, the network was still able to learn via IWL, achieving high train accuracy on both the rule-like and the rotated tasks after task-specific training. IWL again exhibited an interleaving advantage due to catastrophic forgetting. (*F*) Accuracy and loss results over the course of IWL training. Accuracy is split by group, in this case corresponding to whether the cue was part of the row or the column. IWL exhibited catastrophic forgetting in train accuracy when trials were blocked, regardless of rotation condition. IWL also failed to generalize compositionally, failing on the 16 held-out test cues (green lines) in all conditions. (*G–I*) Experiments using networks capable of both IWL and ICL. (*G*) After metalearning, the models exhibited a blocking advantage, but also showed strong compositional generalization, as shown by the high few-shot test accuracy in the blocked condition. ICL failed in the rotated condition. (*H*) After task-specific training, the network exhibited an interleaving advantage in the rotated condition, due to catastrophic forgetting when trials were blocked. (*I*) When trials were blocked in the rule-like condition, train accuracy was nearly perfect, resulting in little loss and thus little IWL. In the rotated condition, ICL failed, resulting in larger losses and thus increased IWL, and increased catastrophic forgetting, as can be seen in the rapid drop in train accuracy on the first group ("TrainA," purple) while training on the second group ("TrainB," orange), and vice versa. No catastrophic forgetting occurred in the interleaved condition, but compositional generalization (green) was considerably worse.

served when such compositional generalization is challenging or impossible.

We focused our investigations on a recent study demonstrating compositional generalization in humans on a novel rule-governed task, where the goal was not to categorize stimuli but to learn a latent compositional coordinate system (see Fig. 3 *A–C*; 24). Notably, this study showed that compositional generalization improved when related trials were blocked (47/58 participants generalized) compared to interleaved (36/60 participants generalized)—consistent with the idea that the mechanisms underlying compositionality can be linked to those responsible for producing the blocking advantage. This task therefore provides an excellent testbed for our metalearning neural networks, allowing us to replicate the above curriculum-related results in a different paradigm while also studying their connection to compositionality.

In the original task, participants learned to pair colored animals with arbitrary xy-coordinates via trial-and-error. Importantly, the correct locations varied systematically with the two features: Color determined the x-coordinate (each of 5 different colors was linked to one of 5 different x-values) while the animal determined the y-coordinate, or vice-versa. Participants saw only 9 of the 25 possible color–animal pairs as study examples; they had to make novel inferences to generalize to the 16 remaining pairs during testing (without feedback). This task can be seen as rule-based in that a simple rule (e.g., color = x, animal = y) governs the locations, and can be seen as compositional in that good test performance requires composition of knowledge about a particular color (e.g., "blue" means x = 3) with knowledge about a particular animal (e.g., "alligator" means y = 2) into a novel combination (e.g., "blue alligator" means location is 3, 2).

The key experimental variable manipulated in the study was the curriculum—which 9 of the 25 cues were used as study examples, and the order in which they were presented (Fig. 3*A*). In the **Blocked** condition, all cues of a particular color (i.e., a single row/column) were presented before all the cues with a particular animal, or vice-versa. In the **Interleaved** condition, a single row and column were again chosen for study, but their order was randomly shuffled.

The experimenters found that human compositional generalization performance depended on which curriculum was used: Participants performed better in the blocked than the interleaved condition (24).[†] The original study did not manipulate the presence or absence of rule-like structure as the categorization task did (5), but we hypothesized that rotating the underlying coordinate grid (Fig. 3*C*) would cause a similar interleaving advantage to emerge. This is because when the underlying coordinate system is rotated, no simple rule (e.g., color = x, animal = y) is available. We therefore tested our metalearning models in both the original **Rule-like** setting, and in a **Rotated** version.

IWL is noncompositional and produces an interleaving advantage. As in the simulations with the categorization task, we first isolated the learning properties of IWL by training transformer neural networks without ICL capabilities on the task. Without ICL, performing the task in the few-shot setting was again impossible (Fig. 3*D*). The only way the network could learn was through IWL over more extensive task-specific training, which again exhibited an interleaving advantage (due to catastrophic forgetting when trials were blocked; main effect of curriculum on train accuracy: $P < 10^{-3}$; see Fig. 3 *E* and *F*). Furthermore, while the network learned the study examples well when trials

[†]Note that the original study also tested two other related conditions, where sampling of items was "Aligned" or "Misaligned;" we simulated these cases and reproduced similar results in *SI Appendix*, but here focus on the key blocked vs. interleaved contrast.

were interleaved, it performed poorly on test trials that required compositional generalization. Thus, in contrast to the categorization task where IWL showed good generalization performance (Fig. 3*F*), the compositional task allowed us to reproduce known failures in compositional generalization in networks capable only of standard IWL (25, 53, 54, 81, 82).

ICL can be compositional and can produce a blocking advantage. We then endowed the network with ICL abilities by first having it metalearn on a distribution of blocked, rule-like tasks (analogous to the developmental process prior to entering a psychology study; see *Materials and Methods* for details). After metalearning, these ICL abilities allowed the network to generalize compositionally on unseen tasks, achieving near-perfect test accuracy on color–animal combinations not included in the study examples. As in the previous simulations, ICL exhibited the blocking advantage observed in humans (24), performing better in the few-shot setting when trials were blocked compared to interleaved (main effect of curriculum on rule-like test accuracy: $P < 10^{-3}$). This was again due to the fact that the model metalearned on blocked curricula (*SI Appendix*).

These findings extend recent work (49) by showing that the ICL algorithm that emerges in metalearning neural networks can reproduce human-like compositional generalization behavior and its associated blocking advantage in this experimental paradigm (24). This is significant because it shows how neural networks, which have traditionally been criticized for lacking compositionality (25, 54), can through metalearning come to implement an ICL algorithm that is capable of human-like compositional generalization (31, 72).

Concurrent ICL and IWL produce compositionality and both curriculum effects. Finally, we allowed the metalearning networks to continue to train via IWL, and replicated the full set of human curriculum effects that we reproduced above in the category-learning setting. As predicted, ICL failed in our novel rotated version of the task, leading to more errors and thus greater IWL (Fig. 3*G*). This increase in IWL led to an interleaving advantage on the rotated task (Fig. 3*H*)—a testable prediction not evaluated in humans in the original study—whereas a blocking advantage was reproduced for the original rule-like task due to ICL (see Fig. 3*G*; interaction between rotation and curriculum on accuracy: $P < 10^{-3}$). Taken together, our findings on the compositional task are again consistent with our theoretical framework (see principle 3), and show how the distinction between in-context and in-weight learning can offer a unified account of human compositional generalization capabilities and their dependence on the learning curriculum (24).

LLMs Exhibit Compositionality and a Blocking Advantage. So far, we have established that it is possible for an ICL algorithm to exhibit compositionality and a blocking advantage, and that a single neural network implementing this kind of ICL alongside its usual IWL will reproduce the full set of empirical results that we have been targeting. A separate question one can ask is *why* a network would develop an ICL algorithm with these particular properties in the first place. In our metalearning experiments, we used task distributions that promote these properties (*Materials and Methods*), but there may be more naturalistic distributions that could give rise to them.

Although the datasets used for training LLMs are developmentally unrealistic in many ways (83–85), they are more naturalistic in the sense that they are largely made up of natural language text, rather than content that is specifically relevant to our tasks. These corpora are not purposefully designed to encourage ICL or any of our hypothesized properties to emerge. Nevertheless, impressive ICL abilities do arise in these models, giving them the flexibility to accomplish many tasks in context (43, 61). Given the scale and complexity of their training datasets, it is unclear a priori what ICL properties LLMs should develop, but prior work has shown that they can exhibit compositional generalization in some settings (64–66), and can be sensitive to the order in which in-context examples are provided (86, 87).

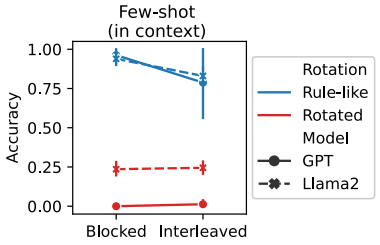

Fig. 4. LLM results. Large language models (LLMs) are capable of in-context learning (ICL) on the text-based version of the compositional task based on the human study (24). Both GPT-3.5 (solid lines) and Llama 2 (dashed lines) achieved good compositional generalization performance (i.e., test accuracy) on the rule-like version of the task (blue), and also exhibited a blocking advantage, performing better when trials were blocked than interleaved (Fig. 3*A*). ICL test accuracy was much worse on the rotated task (red), consistent with our theoretical framework.

We thus hypothesized that the properties of ICL assumed by our theoretical framework (i.e., compositionality and a blocking advantage; see principle 3) may emerge in LLMs. We tested this hypothesis by evaluating two LLMs, Llama 2 (88) and GPT-3.5 (43, 89), on the same compositional task used above. We evaluated the emergent ICL abilities of these models by presenting color–animal pairs from the compositional task only in context.

Both LLMs showed strong compositional generalization performance on the task (Fig. 4), even though they were only given the 9 study examples and had not been explicitly trained on variants of the task. This shows that the emergent ICL abilities in these models can produce the kinds of generalization behaviors that standard IWL in neural networks struggles to achieve (see test accuracy in Fig. 3*F*).

Notably, both LLMs also showed a blocking advantage in the rule-like version of the task (main effect of curriculum on test accuracy: $P < 10^{-3}$).[‡] This again shows that although the ICL capabilities in the LLMs was not specifically sculpted to produce this blocking advantage, it emerges nonetheless via large-scale next-word (next-token) prediction.

Finally, both LLMs performed poorly on the rotated task (main effect of rotation on test accuracy: $P < 10^{-3}$). This is also consistent with our theoretical framework (see principle 3), which predicts that ICL should be more difficult in the absence of rule-like structure because inferences are more complex. IWL would be required to compensate for this failure, as we showed in our metalearning experiments.[§]

Thus, neural networks can come to implement an ICL algorithm with the properties of compositionality, a blocking advantage, and a preference for rule-like structure—even when their training does not specifically target these properties, but consists in next-token prediction on naturalistic text.

[‡]Like our metalearned neural networks, the LLMs also showed the full pattern of curriculum effects described in the human study; see *SI Appendix* for details.

[§]In principle, these models should also show IWL properties like any other neural network, but it is expensive to finetune them, and our main questions here pertain to ICL.

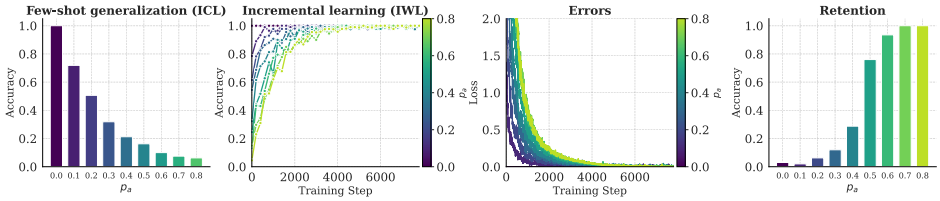

Fig. 5. Trade-off between flexibility and retention. Ablating attention to examples in the context (analogous to limiting the number of items accessible in WM) hurts cognitive flexibility but improves retention. When attention is ablated (i.e., p_a is high), few-shot generalization via ICL suffers (*Left*) and incremental learning via IWL is slower (*Middle Left*). This results in more errors (*Middle Right*), consistent with human EEG data showing prolonged presence of neural signatures of prediction errors under higher WM load (9). When more errors are made, more IWL occurs, resulting in better retention in the absence of contextual information (*Right*), consistent with human results showing better retention after learning under higher WM load (21, 36). Results are from the compositional task. p_a is the probability that attention to each example in the context was inhibited.

Trade-Off Between Flexibility and Retention. So far, we have highlighted the advantages of ICL over IWL in supporting rapid few-shot learning and generalization, and only noted that IWL is necessary in cases where ICL is less effective (e.g., when trials are interleaved or in rotated tasks). However, an additional benefit of IWL (and relatedly, episodic and semantic memory compared to WM) is that retrieval of information stored in synaptic weights does not require persistent activity (e.g., throughout a delay), and can therefore operate over longer timescales and in the absence of contextual cues (44, 74). Thus, we hypothesized that a natural consequence of the coexistence of ICL and IWL would be the emergence of a trade-off between *flexibility*, or rapid adaptation to new tasks from only a few examples, and *retention*, or the ability to recall information after longer delays or in the absence of contextual information. The key insight here is that if learning takes place in a setting where ICL is successful, generalization based on latent rules may be improved, but fewer prediction errors will update weights. When latent rules are identified, they should give rise to reduced prediction errors, and this suppression should in turn relate to better generalization but worse retention.

In fact, analogous findings have been reported in two lines of research on reinforcement learning (RL) in humans (9, 14, 21, 35, 36). When the task was structured with a hidden rule that could be inferred from the context, neural reward prediction errors were suppressed, and this suppression predicted better generalization of that rule (14). In simpler RL tasks, participants also show particularly rapid learning when they can acquire contingencies using an ICL-like WM strategy, achieving near-perfect performance within only a few presentations of each stimulus (9). However, when WM load was increased, many more presentations were required to achieve the same performance, consistent with an IWL-like incremental learning strategy relying more on RL. Electroencephalogram (EEG) recordings showed that neural signatures of reward prediction errors were suppressed when WM load was low (9) or when the underlying structure of the task had been inferred (14). Moreover, this neural marker of an ICL-like WM strategy was predictive of both better generalization (14) and reduced retention (21). To date, however, no single model has accounted for both sets of findings.

We tested whether a similar trade-off between flexibility and retention would emerge in our neural networks. As a proxy for WM load, we ablated the networks' ability to maintain contextual information throughout learning by inhibiting attention to each in-context example with probability p_a (similar results were obtained by adding noise; see *SI Appendix*). Retention was evaluated by testing networks on examples they had seen during training, but in the absence of any contextual information (analogous to testing after a delay when relevant information is not available in WM).

The results for the compositional task are shown in Fig. 5 (results for the category-learning task can be found in *SI Appendix*). In both tasks, we observed a robust trade-off between flexibility and retention. When the ablation was stronger (p_a was higher), few-shot generalization via ICL was worse (Fig. 5, *Left*) and incremental learning via IWL took longer to reach optimal performance (*Middle Left*). This meant that more errors were made throughout learning (*Middle Right*), consistent with the stronger neural signatures of prediction errors observed in humans under higher WM load (9, 75). More errors resulted in greater IWL, leading the model to perform better on the retention test (*Right*), consistent with the superior retention observed in humans that had learned under higher WM load (21, 36).

Thus, the same principles that allow the networks to reproduce compositional generalizations and curriculum effects can also explain the trade-off observed in human RL experiments, recapitulating both its neural and behavioral signatures (9, 21, 36). This trade-off suggests that ICL and IWL have distinct advantages whose relative importance depends on whether flexibility or retention is prioritized.

Discussion

Influential theories in cognitive science posit two distinct systems to account for findings suggesting a duality in human learning (1–10, 90). Prominent theories leverage distinctions between controlled vs. automatic processing (6, 91, 92), model-based vs. model-free reinforcement learning (3, 93, 94), WM in PFC vs. striatal synaptic learning (9, 21, 34, 50), system 2 vs. system 1 thinking (40), and rule-based vs. procedural learning (1, 5). These theories explain why human learning exhibits different phenomena under different conditions. Here, we have focused on three such phenomena: 1) compositionality 2) curriculum effects, and 3) the trade-off between flexibility and retention. Humans are capable of utilizing rule-like structure to generalize compositionally (15, 24, 26–29), and of integrating over multiple dimensions and making arbitrary associations when no rule-like structure is present (1, 5, 60, 95). In the former case, learning can be rapid and flexible, and tends to benefit when related trials are blocked over time (5, 19, 24). In the latter case, it benefits when trials are interleaved (5, 37–39), and can result in improved retention but limited flexibility and generalization.

Our work shows how these phenomena can be explained by a single neural network capable of two qualitatively distinct learning processes. In particular, we have shown how metalearning can endow a network with a capacity to learn *in context*, and how this capacity can capture compositionality and the blocking advantage on tasks governed by rule-like structure. This is analogous to how humans improve at learning new

tasks and using WM to make sophisticated rule-based inferences over the course of development (49, 96, 97). ICL operates when tasks are consistent with readily identifiable rules but can be unsuccessful on tasks lacking such structure, triggering error-driven IWL and producing an interleaving advantage due to catastrophic forgetting (59, 60). This dynamic interaction between ICL and IWL naturally recapitulates the trade-off between flexibility and retention observed in humans: WM can be leveraged to rapidly learn new stimulus–response rules, but causes reductions in neural prediction errors driving incremental reinforcement learning, resulting in worse retention after longer delays (9, 21, 36).

ICL has recently emerged as an important topic in machine learning (43, 98). Studies have investigated what data-distributional properties (44, 74, 99) or architectures (100–102) drive its emergence, as well as the learning algorithm it implements (47, 103, 104), and the internal circuits underlying it (77–79). Here, we link this recent work to longstanding issues in cognitive science, showing how the dynamic interplay between ICL and IWL can offer a unified perspective on phenomena related to prominent dual-process theories.

Curriculum Effects. There has been some debate on whether humans learn better when related content is blocked or interleaved over time, with some studies finding a blocking advantage (5, 19, 24, 33) and others finding an interleaving advantage (5, 37–39). Multiple factors may distinguish these cases (e.g., between-category and within-category similarity; 105), but one important variable may be the presence of rule-like structure: Humans have been shown to exhibit a blocking advantage when the task is governed by succinct rules, and an interleaving advantage when the task does not afford such rules (5, 24). These effects are explained by a dual-process account in which a rule-based learning system operates by an explicit hypothesis-testing strategy and a procedural learning system operates by incrementally integrating information over time (1, 5). Our work offers an integrative perspective on this dual-process account, showing how a similar duality can emerge in neural networks capable of both ICL and IWL.

In our framework, the interleaving advantage arises because of catastrophic forgetting (59), which is a natural property of IWL in neural networks due to their use of overlapping distributed representations (60). Might this kind of forgetting explain the interleaving advantage observed in humans? The brain is thought to mitigate catastrophic forgetting through the use of sparse, pattern-separated representations in the hippocampus (60, 106). However, this effect is unlikely to be eliminated completely, so a similar principle may still underlie the modest interleaving advantage observed in humans (5). Future work could directly investigate the extent to which the interleaving advantage observed in the absence of rule-like structure is due to this kind of forgetting.

The blocking advantage, on the other hand, does not emerge by default in standard neural networks, but a number of studies have explored the neural mechanisms that might underlie it. For example, a neural network model of rule-based inference and WM in the PFC showed that blocking related trials over time can encourage abstract rule-like representations to emerge in the network's activations (18). More recent work (58) showed that a PFC-like neural network augmented with a gating mechanism and a bias for active maintenance produces a blocking advantage on a task involving cognitive maps (107). Related work has shown how a neural network equipped with a specialized Hebbian gating mechanism (108) can reproduce a blocking advantage observed in humans on an analogous task (19). A similar Hebbian mechanism was then used to explain the blocking advantage observed in the compositional task studied here (24). Another recent study showed how the blocking advantage observed in humans on a next-state prediction task (33) was reproduced by a neural network model that actively maintained distinct contextual representations over time (109). Overall, these studies emphasize how a blocking advantage can emerge when inferences are made through forward activation dynamics (i.e., *in context*), such as those made over items maintained in WM in PFC.

Our theoretical account of the blocking advantage is broadly consistent with previous models but has some advantages. First, we have shown how it can emerge in a model that also produces the interleaving advantage on other tasks. Furthermore, while our framework is consistent with previous models in suggesting that the blocking advantage is related to activation dynamics (e.g., WM in PFC; 18, 58), we show how these dynamics can be metalearned (71), thus providing a conceptual link between these prior models and ongoing work investigating metalearning and cognitive flexibility in natural and AI (49, 64, 67, 69, 110, 111).

Indeed, we also observed a blocking advantage in LLMs, which appear to exhibit high levels of cognitive flexibility (43, 61, 112). These results show that a blocking advantage can emerge with ICL even when networks are trained on natural text rather than metalearning datasets specifically designed to promote it. Although it is difficult to know exactly why this blocking advantage emerges, we speculate that it is driven by distributional properties of natural text, such as the tendency for human writing to afford inferences best made by assimilating consecutive examples in a sequential manner. Further work is needed to better understand the mechanisms responsible for the blocking advantage in LLMs.

In general, our work does not directly address whether the human blocking advantage emerges due to strong constraints imposed by neural architecture (e.g., recurrence, limited WM capacity), rather than the statistical properties of the environment. Our metalearning networks and LLMs utilized the transformer architecture (76), which is not recurrent and does not have hard constraints in WM capacity. Both the blocking advantage and the preference for rule-like tasks emerged in these models due to the statistical properties of their training data. This was especially clear in the metalearning experiments, where we had full control over the data distribution and confirmed that it determines when the blocking advantage emerges (*SI Appendix*). Consistent with these findings, prior work has shown that metalearning networks trained on category-learning problems that match the natural statistics of real-world tasks perform poorly on the same problems humans struggle with ref. 69. Furthermore, the human blocking advantage has been shown to depend on the extent to which feature dimensions relevant to the rule-like structure of the task are represented in a strongly segregated manner (19), a factor likely to depend on an individual's prior learning experiences. However, we think the human blocking advantage is also likely to depend on key architectural features of the brain, such as its recurrence and mechanisms for gating and serial attention in PFC and basal ganglia (17, 18, 34, 58). These, in turn, might affect the distributional properties of natural language produced by humans and provided as training data for the LLMs (113). Further work is required to fully resolve the extent to which architectural features rather than distributional properties drive the blocking advantage in humans. However, regardless of its origins, our work shows that the simultaneous presence of ICL and IWL can explain how a blocking advantage on rule-like tasks

can coexist with an interleaving advantage on other tasks within a single neural network.

Compositionality. Compositionality is thought to be a key property underlying human cognitive flexibility, permitting familiar rules or concepts to be combined in novel ways, thus facilitating a powerful form of generalization (15, 25, 72, 114). Recent work has shown that although compositionality may not be a natural property of standard IWL (25, 53, 54, 82), it can emerge with ICL (49, 72, 115, 116). Our results build on this work, showing that it is possible to endow a neural network with an ICL algorithm that is capable of reproducing compositional generalization behaviors observed in humans in a recent study (24), even when standard IWL fails (see test accuracy in Fig. 3*F*). We showed that this kind of ICL algorithm can be metalearned from a distribution of related tasks, but also emerges in LLMs trained on large corpora of text (Fig. 4). While metalearning offers a clear understanding of how a neural network can come to implement an emergent compositional learning algorithm (49, 72), it is less clear why this would emerge in LLMs. One suggestion is that at large enough scales, the language modeling objective used in LLMs can itself be seen as engendering a kind of metalearning (43, 110), where some subset of training samples puts pressure on these models to learn how to compose novel concepts or reasoning steps in context (72). This is consistent with the hypothesis that human compositionality is metalearned—a conjecture that, while difficult to study, may yield specific empirical predictions (31, 49, 96, 97). Finally, a key contribution of our work is that it builds on studies linking compositionality to curriculum effects in humans (24), providing a unified account of compositional generalization and its dependence on curriculum.

One Network or Two Systems for Category Learning? While our neural networks are not meant to be comprehensive models of human category learning (see e.g., ref. 117), they may be relevant to other phenomena observed in category-learning studies. One ongoing debate in this area has been about whether human category learning is best characterized by a single learning system or by multiple systems (1, 95, 118–125). Single-system theories emphasize the principle of parsimony, and argue that a system that relies on stimulus similarity and selective attention can explain most of the available findings (120). Multiple-systems theories argue that a single system is not sufficient to account for double dissociations evident in human behavior (1, 126, 127), such as the one pertaining to the curriculum effects discussed above (5).

Our work may help resolve this debate by showing how such double dissociations can be explained by a single network that can learn in two different ways. ICL and IWL are not separate learning *systems*, but nevertheless manifest fundamentally different properties and compete to drive learning behavior, with each taking precedence at different times. Our approach arguably maintains the parsimony of a single-system theory in the sense that these two distinct sets of learning properties emerge from the natural dynamics of a single network, rather than being independently posited as part of separate systems. However, as discussed above, the properties of ICL and IWL align well with the two systems proposed in prominent multiple-system theories (1, 118), with ICL corresponding to the explicit, verbal system and IWL corresponding to the implicit, procedural system.

In addition to the curriculum effects we observed in our experiments, the distinction between ICL and IWL may help to explain other findings motivating multiple-system theories of category learning. For example, some studies have shown that increased WM load can impair rule-based learning (126, 128–130, although see 122, 131). This finding parallels our results showing that ICL-mediated generalization suffers when access to contextual information is restricted (Fig. 5). There is also some evidence that children struggle specifically with rule-based category-learning tasks (132–135), but can perform at adult levels when categories are based on family resemblances (136). This is consistent with our neural networks, which are inherently capable of IWL but only develop sophisticated ICL through metalearning (31).

Our models may also clarify certain outstanding questions for current multiple-system theories. For example, behavioral evidence suggests that the verbal or explicit system operates by default initially in humans, but it is unclear a priori why this would be the case (1, 5). In our neural network models, ICL operates by default because it can occur at a much faster timescale (through activation dynamics), and because IWL only occurs when errors are made. Another unresolved question concerns evidence from neuroimaging studies on category learning suggesting that there is substantial overlap in the brain regions active during rule-based and information-integration tasks (137, 138). This can seem to contradict the predictions of a multiple-system theory that posits completely independent learning modules. The distinction between ICL and IWL provides a natural explanation for this finding, as these two learning processes coexist throughout the network and therefore need not be localizable to separate regions.

In fact, our neural networks are likely to be unrealistically homogeneous, as they have no inherent modularity at all. Many findings suggest that specific brain regions such as PFC are particularly important for cognitive functions such as WM, rule-based inference, and modulating processing according to the current context or goal (18, 52, 139–141). We speculate that the organization of the human PFC, which has an intrinsic bias to robustly maintain information over longer timescales until it is actively updated (17, 142, 143), may encourage ICL abilities, along with their specific properties, to become partially localized to this area (58, 144).

Although our models did not contain any separate PFC-like system, we note that the ICL algorithms implemented in their activation dynamics can be seen as analogous to those observed in neural models of PFC trained on multiple tasks (13, 18, 71). Just as in our models, these ICL-like abilities only emerge through IWL-like learning of abstract representations in PFC and gating policies in the basal ganglia. Recent work has shown that transformers can mimic the frontostriatal gating mechanisms in these biological models when trained on human WM tasks, and exhibit effective capacity limitations despite the lack of any inherent architectural constraint imposing such a limitation (145, 146). While we did not directly investigate the fine-grained computations carried out in ICL in our models, the results of the ablation experiments (Fig. 5) illustrate how ICL relates to attentional access to contextual information. Future work could more thoroughly investigate whether emergent PFC-like computational mechanisms also explain the ICL-related phenomena in our metalearning networks.

Materials and Methods

Model Details. All metalearning experiments used the transformer architecture (76, 88). An informal hyperparameter search was conducted over number of layers, hidden size, dropout, and learning rate. The size of the feedforward layers was always twice the hidden size. The best model was selected based on

validation accuracy. In the category task, the best model had 4 layers, 8 heads, a hidden size of 64, and no dropout. In the compositional task, the best model had 12 layers, 8 heads, a hidden size of 64, and dropout of 0.1. Models were evaluated on exact-match accuracy.

In the LLM experiments, we evaluated GPT-3.5 (43, 89) and Llama 2 (88). GPT-3.5 is an LLM trained on next-token prediction and finetuned to be more useful in a chat-based interface. Llama 2 had not been finetuned on instruction data. In GPT-3.5 ("gpt-3.5-turbo-instruct"), temperature was set to 0.1 and five runs were performed. A maximum of 7 tokens were generated, and no postprocessing was done except to strip extra spaces. Llama 2 is an open-source model with approximately 70 billion parameters. The model was run using resources from the Center for Computation and Visualization at Brown University. Different prompts were tested, but good performance was achieved with simple prompts containing only the study examples; prompts did not qualitatively change the pattern of results across conditions.

Metalearning. For the category-learning experiments, networks metalearned on a distribution of tasks with the same basic structure described above. Each individual task was sampled as follows: 2 feature dimensions were sampled uniformly without replacement from a set of 200 unique dimensions. Each of these dimensions had 8 possible values, making 64 possible items in the newly sampled task. One of two possible category labels was randomly assigned to each of the two categories. In each new task, 16 items from each category were included in the set of 32 study examples. The queries seen during metalearning could either be one of the 32 given in the context ("train"), or one of the remaining 32 ("test"). All samples in the metalearning distribution used the rule-like task and the blocked curriculum. The network metalearned on 12,000 tasks and was tested on a held-out set of 100 tasks that had not been seen during training. A further 10 held-out tasks were used for testing. In the category setting, networks metalearned for 20 epochs with cross-entropy loss, the Adam optimizer (147), a learning rate of 0.0001, and a batch size of 256.

The tasks used for metalearning on the compositional task (24) were sampled as follows: The orders of the lists of five colors and five animals were shuffled. Then, the two features were randomly assigned to the x- and the y-coordinates (color = x and animal = y, or vice versa). In the rotated condition, this 5 × 5 grid was rotated by 45° and scaled so that each coordinate landed on an integer. All samples in the metalearning distribution were rule-like and blocked. We again generated 12,000 tasks for metalearning, and used 100 held-out tasks for validation. A further 10 held-out tasks were used for testing. During metalearning in the compositional task, networks trained for 500 epochs with the Adam optimizer (147), a learning rate of 0.001, and a batch size of 256.

Task-Specific Training. Once the network acquired an ICL algorithm through metalearning, it was subsequently evaluated on its ability to learn new unseen tasks from each condition. This evaluation was conducted in two ways. In the **few-shot** evaluation, the weights of the network were frozen, ensuring that all learning was due to ICL on the study examples given in context. In **task-specific training**, the model's weights were not frozen, and any errors made were used to update weights. During task-specific training, the model learned a single task and only received feedback on the study examples, thus emulating the experience of the human participants (24). Note that this is unlike the metalearning phase, when the model learned how to generalize to queries not included in the study examples. This second task-specific learning phase that the model underwent can be understood as "finetuning" the model on a specific task, while the metalearning can be understood as "pretraining." During task-specific training, networks were again trained with cross-entropy loss and the Adam optimizer (147), with a learning rate of 0.00001 in the category-learning task, and a learning rate of 0.0001 in the compositional task. In both tasks, the batch size was equal to the total number of examples (i.e., queries) used in a given block (32 in the category-learning setting, 5 in the compositional setting).

During the task-specific training phase, samples were either blocked or interleaved in two distinct but congruent ways. In the blocked condition, related items were blocked over the context, but they were also blocked over the gradient steps (i.e., the model was trained for N gradient steps on samples containing queries from one stimulus group, then was trained for N gradient steps on samples containing queries from the other group, and so on). Likewise, in the interleaving condition, items from each group were interleaved both over the context and over the gradient steps. In the main experiments, the curriculum condition was always consistent during task-specific training—related items were either blocked over both the context and the gradients steps, or interleaved over both the context and the gradient steps. However, for the sake of completeness we experimented with all combinations and report these results in *SI Appendix*.

Data, Materials, and Software Availability. Code data have been deposited in GitHub: jlrussin/icl-iwl-interplay (https://github.com/jlrussin/icl-iwl-interplay) (148).

ACKNOWLEDGMENTS. We thank David Badre, Peter Hitchcock, Joonhwa Kim, the Language Understanding and Representation Lab, the Lab for Neural Computation and Cognition, and the Analogy Group for helpful discussions. An early version of this work was published in the Proceedings of the 46th Annual Meeting of the Cognitive Science Society. This work was supported by Office of Naval Research N00014-23-1-2792, NIH National Institutes of General Medical Sciences Centers of Biomedical Research Excellence #5P20GM103645-10, and Defense Advanced Research Projects Agency #D24AP00261.

1. F. G. Ashby, W. T. Maddox, Human category learning 2.0. *Ann. New York Acad. Sci.* **1224**, 147–161 (2011).
2. M. Botvinick *et al.*, Reinforcement learning, fast and slow. *Trends Cogn. Sci.* **23**, 408–422 (2019).
3. N. D. Daw, Y. Niv, P. Dayan, Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* **8**, 1704–1711 (2005).
4. J. S. B. T. Evans, Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.* **59**, 255–278 (2008).
5. S. M. Noh, V. X. Yan, R. A. Bjork, W. T. Maddox, Optimal sequencing during category learning: Testing a dual-learning systems perspective. *Cognition* **155**, 23–29 (2016).
6. R. C. O'Reilly, A. Nair, J. L. Russin, S. A. Herd, How sequential interactive processing within frontostriatal loops supports a continuum of habitual to controlled processing. *Front. Psychol.* **11**, 380 (2020).
7. M. Sablé-Meyer *et al.*, A geometric shape regularity effect in the human brain. *eLife* **14**, RP106464 (2025).
8. S. A. Sloman, The empirical case for two systems of reasoning. *Psychol. Bull.* **119**, 3–22 (1996).
9. A. G. E. Collins, M. J. Frank, Within- and across-trial dynamics of human EEG reveal cooperative interplay between reinforcement learning and working memory. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 2502–2507 (2018).
10. R. C. Atkinson, R. M. Shiffrin, Human memory: A proposed system and its control processes in *The Psychology of Learning and Motivation: Advances in Research and Theory*, K. W. Spence, J. T. Spence Eds. (Academic Press, 1968), pp. 89–195.
11. A. B. Nelson, R. M. Shiffrin, The co-evolution of knowledge and event memory. *Psychol. Rev.* **120**, 356–394 (2013).
12. E. Tulving, "Episodic and semantic memory" in *Organization of Memory*, E. Tulving, W. Donaldson, Eds. (Academic Press, San Diego, CA, 1972), pp. 381–403.
13. A. G. E. Collins, M. J. Frank, Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychol. Rev.* **120**, 190–229 (2013).
14. A. G. E. Collins, M. J. Frank, Neural signature of hierarchically structured expectations predicts clustering and transfer of rule sets in reinforcement learning. *Cognition* **152**, 160–169 (2016).
15. B. M. Lake, T. D. Ullman, J. B. Tenenbaum, S. J. Gershman, Building machines that learn and think like people. *Behav. Brain Sci.* **40**, e253 (2017).
16. G. A. Miller, *The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information* (Indiana: Bobbs-Merrill, 1956), vol. 101.
17. R. C. O'Reilly, M. J. Frank, Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia. *Neural Comput.* **18**, 283–328 (2006).
18. N. P. Rougier, D. C. Noelle, T. S. Braver, J. D. Cohen, R. C. O'Reilly, Prefrontal cortex and flexible cognitive control: Rules without symbols. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 7338–7343 (2005).
19. T. Flesch, J. Balaguer, R. Dekker, H. Nili, C. Summerfield, Comparing continual task learning in minds and machines. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E10313–E10322 (2018).
20. J. Pesnot Lerousseau, C. Summerfield, Space as a scaffold for rotational generalisation of abstract concepts. *eLife* **13**, RP93636 (2024).
21. R. Rac-Lubashevsky, A. Cremer, A. G. E. Collins, M. J. Frank, L. Schwabe, Neural index of reinforcement learning predicts improved stimulus-response retention under high working memory load. *J. Neurosci.* **43**, 3131–3143 (2023).
22. B. Lake, R. Salakhutdinov, J. Gross, J. Tenenbaum, One shot learning of simple visual concepts in *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*, L. Carlson, C. Hölscher, T. F. Shipley, Eds. (Cognitive Science Society, 2011), pp. 2568–2573.
23. B. M. Lake, R. Salakhutdinov, J. B. Tenenbaum, Human-level concept learning through probabilistic program induction. *Science* **350**, 1332–1338 (2015).
24. R. B. Dekker, F. Otto, C. Summerfield, Curriculum learning for human compositional generalization. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2205582119 (2022).

25. J. A. Fodor, Z. W. Pylyshyn, Connectionism and cognitive architecture: A critical analysis. *Cognition* **28**, 3–71 (1988).
26. N. T. Franklin, M. J. Frank, Compositional clustering in task structure learning. *PLoS Comput. Biol.* **14**, e1006116 (2018).
27. N. T. Franklin, M. J. Frank, Generalizing to generalize: Humans flexibly switch between compositional and conjunctive structures during reinforcement learning. *PLoS Comput. Biol.* **16**, e1007720 (2020).
28. S. M. Frankland, J. D. Greene, Concepts and compositionality. In search of the Brain's language of thought. *Annu. Rev. Psychol.* **71**, 273–303 (2020).
29. B. M. Lake, T. Linzen, M. Baroni, "Human few-shot learning of compositional instructions" in *Proceedings of the 41th Annual Meeting of the Cognitive Science Society, CogSci 2019: Creativity + Cognition + Computation, Montreal, Canada, July 24-27, 2019*, A. K. Goel, C. M. Seifert, C. Freksa, Eds. (Cognitive science society.org, 2019), pp. 611–617.
30. R. G. Liu, M. J. Frank, Hierarchical clustering optimizes the tradeoff between compositionality and expressivity of task structures for flexible reinforcement learning. *Artif. Intell.* **312**, 103770 (2022).
31. J. Russin, S. W. McGrath, E. Pavlick, M. J. Frank, Is human compositionality meta-learned? *Comment. Behav. Brain Sci.* **47**, e162 (2024).
32. P. Schwartenbeck *et al*., Generative replay underlies compositional inference in the hippocampal-prefrontal circuit. *Cell* **186**, 4885–4897.e14 (2023).
33. A. O. Beukers *et al*., Blocked training facilitates learning of multiple schemas. *Commun. Psychol.* **2**, 1–17 (2024).
34. M. J. Frank, D. Badre, Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: Computational analysis. *Cereb. Cortex* **22**, 509–526 (2012).
35. A. G. E. Collins, The tortoise and the hare: Interactions between reinforcement learning and working memory. *J. Cogn. Neurosci.* **30**, 1422–1432 (2018).
36. P. Hitchcock, J. Kim, M. Frank, How working memory and reinforcement learning interact when avoiding punishmentand pursuing reward concurrently. *J. Exp. Psychol. Gen.*, in press.
37. L. E. Richland, J. R. Finley, R. A. Bjork, Differentiating the contextual interference effect from the spacing effect. *Proc. Annu. Meet. Cogn. Sci. Soc.* **26**, 179–187 (2004).
38. S. Goode, R. A. Magill, Contextual interference effects in learning three badminton serves. *Res. Q. Exerc. Sport* **57**, 308–314 (1986).
39. D. K. Landin, E. P. Hebert, M. Fairweather, The effects of variable practice on the performance of a basketball skill. *Res. Q. Exerc. Sport* **64**, 232–237 (1993).
40. D. Kahneman, *Thinking, Fast and Slow* (Farrar, Straus and Giroux, ed. 1, 2011).
41. D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986).
42. D. E. Rumelhart, J. L. McClelland, P. R. Group, Eds., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 2: Psychological and Biological Models* (MIT Press, Cambridge, MA, USA, 1986).
43. T. Brown, Language Models are Few-Shot Learners in Advances in Neural Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin, Eds. (Curran Associates, Inc., 2020), pp. 1877–1901.
44. S. C. Y. Chan *et al*., "Data distributional properties drive emergent in-context learning in transformers" in Proceedings of the 36th International Conference on Neural Information Processing Systems, S. Koyejo *et al*., Eds. (Curran Associates Inc., 2022), pp. 18878–18891.
45. A. K. Lampinen, S. C. Y. Chan, A. K. Singh, M. Shanahan, The broader spectrum of in-context learning. arXiv [Preprint] (2025). https://doi.org/10.48550/arXiv.2412.03782 (Accessed 12 August 2025).
46. J. von Oswald *et al*., Uncovering mesa-optimization algorithms in Transformers. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2309.05858 (Accessed 12 August 2025).
47. S. M. Xie, A. Raghunathan, P. Liang, T. Ma, "An explanation of in-context learning as implicit Bayesian inference" in International Conference on Learning Representations, C. Finn, Y. Choi, M. Deisenroth, Eds. (ICLR, 2022), pp. 1-25.
48. J. X. Wang *et al*., *Learning to reinforcement learn*. arXiv [Preprint] (2017). https://doi.org/10.48550/arXiv.1611.05763 (Accessed 12 August 2025).
49. B. M. Lake, M. Baroni, Human-like systematic generalization through a meta-learning neural network. *Nature* **623**, 115–121 (2023).
50. M. J. Frank, E. D. Claus, Anatomy of a decision: Striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychol. Rev.* **113**, 300–326 (2006).
51. T. Kriete, D. C. Noelle, J. D. Cohen, R. C. O'Reilly, Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 16390–16395 (2013).
52. E. K. Miller, J. D. Cohen, An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* **24**, 167–202 (2001).
53. B. M. Lake, M. Baroni, "Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks" in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, Proceedings of Machine Learning Research*, J. G. Dy, A. Krause, Eds. (PMLR, 2018), vol. 80, pp. 2879–2888.
54. G. F. Marcus, Rethinking eliminative connectionism. *Cogn. Psychol.* **37**, 243–282 (1998).
55. G. Marcus, Deep learning: A critical appraisal. arXiv [Preprint] (2018). https://doi.org/10.48550/arXiv.1801.00631 (Accessed 12 August 2025).
56. S. Pinker, A. Prince, On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition* **28**, 73–193 (1988).
57. J. Quilty-Dunn, N. Porot, E. Mandelbaum, The best game in town: The reemergence of the language-of-thought hypothesis across the cognitive sciences. *Behav. Brain Sci.* **46**, e261 (2023).
58. J. Russin, M. Zolfaghar, S. A. Park, E. Boorman, R. C. O'Reilly, A neural network model of continual learning with cognitive control. arXiv [Preprint] (2012). https://doi.org/10.48550/arXiv.2202.04773 (Accessed 12 August 2025).
59. M. McCloskey, N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem" in *The Psychology of Learning and Motivation*, G. H. Bower, Ed. (Academic Press, San Diego, CA, 1989), vol. 24, pp. 109–164.
60. J. L. McClelland, B. L. McNaughton, R. C. O'Reilly, Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychol. Rev.* **102**, 419–457 (1995).
61. S. Bubeck *et al*., Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv [Preprint] (2023). http://arxiv.org/abs/2303.12712 (Accessed 12 August 2025).
62. A. Saparov *et al*., "Testing the general deductive reasoning capacity of large language models using OOD examples" in *Advances in Neural Information Processing Systems*, A. Oh *et al*., Eds. (NeurIPS, 2023). pp. 3083–3105.
63. S. Musker, A. Duchnowski, R. Millière, E. Pavlick, LLMs as models for analogical reasoning. *J. Mem. Lang. 145*, 104676 (2025).
64. T. Webb, K. J. Holyoak, H. Lu, Emergent analogical reasoning in large language models. *Nat. Hum. Behav.* **7**, 1526–1541 (2023).
65. O. Press *et al*., (2023). Measuring and Narrowing the Compositionality Gap in Language Models in Findings of the Association for Computational Linguistics: EMNLP 2023, H. Bouamor, J. Pino, K. Bali, Eds. (Association for Computational Linguistics, 2023), pp. 5687–5711.
66. D. Zhou *et al*., "Least-to-most prompting enables complex reasoning in large language models" in *The Eleventh International Conference on Learning Representations*, M. Nickel, M. Wang, N. F. Chen, V. Marivate, Eds. (ICLR, 2023), pp. 1-61.
67. M. Binz *et al*., Meta-learned models of cognition. *Behav. Brain Sci.* **47**, e147 (2024).
68. C. Finn, P. Abbeel, S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks" in *Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17*, D. Precup, Y. W. The, Eds. (JMLR.org, Sydney, NSW, Australia, 2017), pp. 1126–1135.
69. A. K. Jagadish, J. Coda-Forno, M. Thalmann, E. Schulz, M. Binz, "Human-like category learning by injecting ecological priors from large language models into neural networks" in *Proceedings of the 41st International Conference on Machine Learning*, K. Heller, Z. Kolter, N. Oliver, A. Weller, Eds. (ICML, 2024), pp. 21121–21147.
70. A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, T. Lillicrap, "Meta-learning with memory-augmented neural networks" in Proceedings of The 33rd International Conference on Machine Learning, M. Balcan, K. Q. Weinberger, Eds. (PMLR, 2016), pp. 1842–1850.
71. J. X. Wang *et al*., Prefrontal cortex as a meta-reinforcement learning system. *Nat. Neurosci.* **21**, 860–868 (2018).
72. J. Russin, S. W. McGrath, D. J. Williams, L. Elber-Dorozko, From Frege to chatGPT: Compositionality in language, cognition, and deep neural networks. arXiv [Preprint] (2024). https://doi.org/10.48550/arXiv.2405.15164 (Accessed 12 August 2025).
73. S. Anand, M. A. Lepori, J. Merullo, E. Pavlick, (2024). Dual process learning: Controlling use of in-context vs. in-weights strategies with weight forgetting. arXiv [Preprint] (2024). http://arxiv.org/abs/2406.00053 (Accessed 12 August 2025).
74. G. Reddy, The mechanistic basis of data dependence and abrupt learning in an in-context classification task. arXiv [Preprint] (2023). http://arxiv.org/abs/2312.03002 (Accessed 12 August 2025).
75. A. G. Collins, B. Ciullo, M. J. Frank, D. Badre, Working memory load strengthens reward prediction errors. *J. Neurosci.* **37**, 4332–4342 (2017).
76. A. Vaswani *et al*., "Attention is all you need" in *Advances in Neural Information Processing Systems 30*, I. Guyon *et al*., Eds. (long beach, CA, USA, 2017), pp. 5998–6008.
77. R. Hendel, M. Geva, A. Globerson, In-Context Learning Creates Task Vectors in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, K. Bali, Eds. (Association for Computational Linguistics, 2023), pp. 9318–9333.
78. E. Todd *et al*., Function vectors in large language models. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2310.15213 (Accessed 12 August 2025).
79. C. Olsson *et al*., In-context learning and induction heads. Transformer Circuits Thread. arXiv [Preprint] (2022). https://doi.org/10.48550/arXiv.2209.11895 (Accessed 12 August 2025).
80. J. Russin, E. Pavlick, M. J. Frank, "Human curriculum effects emerge with in-context learning in neural networks" in *Proceedings of the Annual Meeting of the Cognitive Science Society*, L. K. Samuelson, S. Frank, M. Toneva, A. Mackey, E. Hazeltine, Eds. (CogSci, 2024), pp. 1486–1493.
81. D. Keysers *et al*., "Measuring compositional generalization: A comprehensive method on realistic data" in *International Conference on Learning Representations*, D. Song, K. Cho, M. White, Eds. (ICLR, 2020), pp. 1–38.
82. N. Kim, T. Linzen, COGS: A Compositional Generalization Challenge Based on Semantic Interpretation in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, Y. Liu Eds. (Association for Computational Linguistics, 2020), pp. 9087–9105.
83. M. C. Frank, Bridging the data gap between children and large language models. *Trends Cogn. Sci.* **27**, 990–992 (2023).
84. T. Linzen, "How can we accelerate progress towards human-like linguistic generalization?" in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, J. Tetreault, Eds. (Association for Computational Linguistics, Online, 2020), pp. 5210–5217.
85. A. Warstadt *et al*., "Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora" in *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, A. Warstadt *et al*., Eds. (Association for Computational Linguistics, Singapore, 2023), pp. 1–34.
86. X. Chen, R. A. Chi, X. Wang, D. Zhou, "Premise order matters in reasoning with large language models" in *Proceedings of the 41st International Conference on Machine Learning*, K. Heller, Z. Kolter, N. Oliver, A. Weller, Eds. (JMLR, 2024), pp. 6596–6620.
87. Y. Lu, M. Bartolo, A. Moore, S. Riedel, P. Stenetorp, "Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity" in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, A. Villavicencio, Eds. (Association for Computational Linguistics, Dublin, Ireland, 2022), pp. 8086–8098.
88. H. Touvron *et al*., Llama 2: Open foundation and fine-tuned chat models. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2307.09288 (Accessed 12 August 2025).
89. L. Ouyang *et al*., "Training language models to follow instructions with human feedback" in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, S. Koyejo *et al*., Eds. (Curran Associates Inc., 2022), pp. 27730–27744.
90. J. Evans, K. E. Stanovich, Dual-process theories of higher cognition: Advancing the debate. *Persp. Psychol. Sci.* **8**, 223–241 (2013).
91. R. M. Shiffrin, W. Schneider, Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychol. Rev.* **84**, 127–190 (1977).

92. R. A. Fabio, T. Caprì, M. Romano, From controlled to automatic processes and back again: The role of contextual features. *Eur. J. Psychol.* **15**, 773–788 (2019).
93. N. D. Daw, S. J. Gershman, B. Seymour, P. Dayan, R. J. Dolan, Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69**, 1204–1215 (2011).
94. R. S. Sutton, A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA, 1998).
95. F. G. Ashby, W. T. Maddox, Human category learning. *Annu. Rev. Psychol.* **56**, 149–178 (2005).
96. S. Piantadosi, R. Aslin, Compositional reasoning in early childhood. *PLoS One* **11**, e0147734 (2016).
97. S. T. Piantadosi, H. Palmeri, R. Aslin, Limits on composition of conceptual operations in 9-month-olds. *Infancy Off. J. Int. Soc. Infant Stud.* **23**, 310–324 (2018).
98. Q. Dong *et al*., (2024). A Survey on In-context Learning in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, Y.-N. Chen, Eds. (Association for Computational Linguistics, 2024), pp. 1107–1128.
99. J. H. Lee, A. K. Lampinen, A. K. Singh, A. M. Saxe, Distinct computations emerge from compositional curricula in in-context learning. arXiv [Preprint] (2025). https://doi.org/10.48550/arXiv.2506.13253 (Accessed 12 August 2025).
100. R. Grazzi, J. Siems, S. Schrodi, T. Brox, F. Hutter, Is Mamba capable of in-context learning? arXiv [Preprint] (2024). https://doi.org/10.48550/arXiv.2402.03170 (Accessed 12 August 2025).
101. N. M. Sushma *et al*., State-space models can learn in-context by gradient descent. arXiv [Preprint] (2024). https://doi.org/10.48550/arXiv.2410.11687 (Accessed 12 August 2025).
102. N. Zucchet *et al*., Gated recurrent neural networks discover attention. arXiv [Preprint] (2024). https://doi.org/10.48550/arXiv.2309.01775 (Accessed 12 August 2025).
103. E. Akyürek, D. Schuurmans, J. Andreas, T. Ma, D. Zhou, What learning algorithm is in-context learning? Investigations with linear models. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2211.15661 (Accessed 12 August 2025).
104. J. von Oswald *et al*., Transformers learn in-context by gradient descent in *Proceedings of the 40th International Conference on Machine Learning*, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, Eds. (PMLR, 2023), pp. 35151–35174.
105. P. F. Carvalho, R. L. Goldstone, Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Mem. Cogn.* **42**, 481–495 (2014).
106. R. C. O'Reilly, R. Bhattacharyya, M. D. Howard, N. Ketz, Complementary learning systems. *Cogn. Sci.* **38**, 1229–1248 (2014).
107. S. A. Park, D. S. Miller, H. Nili, C. Ranganath, E. D. Boorman, Map making: Constructing, combining, and inferring on abstract cognitive maps. *Neuron* **107**, 1226–1238.e8 (2020).
108. T. Flesch, D. G. Nagy, A. Saxe, C. Summerfield, Modelling continual learning in humans with Hebbian context gating and exponentially decaying task signals. *PLoS Comput. Biol.* **19**, e1010808 (2023).
109. T. Giallanza, D. Campbell, J. D. Cohen, Toward the emergence of intelligent control: Episodic generalization and optimization. *Open Mind* **8**, 688–722 (2024).
110. K. Sandbrink, C. Summerfield, Modelling cognitive flexibility with deep neural networks. *Curr. Opin. Behav. Sci.* **57**, 101361 (2024).
111. J. X. Wang, Meta-learning in natural and artificial intelligence. *Curr. Opin. Behav. Sci.* **38**, 90–95 (2021).
112. S. Mirchandani et al., Large language models as general pattern machines. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2307.04721 (Accessed 12 August 2025).
113. W. Xu, R. Futrell, Strategic resource allocation in memory encoding: An efficiency principle shaping language processing (2025).
114. R. C. O'Reilly, C. Ranganath, J. L. Russin, The structure of systematicity in the brain. *Curr. Dir. Psychol. Sci.* **31**, 124–130 (2022).
115. S. C. Y. Chan *et al*., Transformers generalize differently from information stored in context vs in weights. arXiv [Preprint] (2022). http://arxiv.org/abs/2210.05675 (Accessed 12 August 2025).
116. A. K. Lampinen *et al*., On the generalization of language models from in-context learning and finetuning: A controlled study. arXiv [Preprint] (2025). https://doi.org/10.48550/arXiv.2505.00661 (Accessed 12 August 2025).
117. B. C. Love, D. L. Medin, T. M. Gureckis, SUSTAIN: A network model of category learning. *Psychol. Rev.* **111**, 309–332 (2004).
118. F. G. Ashby, L. A. Alfonso-Reese, A. U. Turken, E. M. Waldron, A neuropsychological theory of multiple systems in category learning. *Psychol. Rev.* **105**, 442–481 (1998).
119. F. G. Ashby, J. D. Smith, L. A. Rosedahl, Dissociations between rule-based and information-integration categorization are not caused by differences in task difficulty. *Mem. Cogn.* **48**, 541–552 (2020).
120. R. M. Nosofsky, M. K. Johansen, Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychon. Bull. Rev.* **7**, 375–402 (2000).
121. R. M. Nosofsky, "The generalized context model: An exemplar model of classification" in *Formal Approaches in Categorization*, E. M. Pothos, A. J. Wills, Eds. (Cambridge University Press, New York, NY, US, 2011), pp. 18–39.
122. J. P. Minda, C. L. Roark, P. Kalra, A. Cruz, Single and multiple systems in categorization and category learning. *Nat. Rev. Psychol.* **3**, 536–551 (2024).
123. B. R. Newell, J. C. Dunn, M. Kalish, "Systems of category learning: Fact or fantasy?" in *The Psychology of Learning and Motivation: Advances in Research and Theory, Vol. 54*, B. Ross, Ed. (Elsevier Academic Press, San Diego, CA, US, 2011), pp. 167–215.
124. R. A. Poldrack, K. Foerde, Category learning and the memory systems debate. *Neurosci. Biobehav. Rev.* **32**, 197–205 (2008).
125. R. D. Stanton, R. M. Nosofsky, Feedback interference and dissociations of classification: Evidence against the multiple-learning-systems hypothesis. *Mem. Cogn.* **35**, 1747–1758 (2007).
126. W. T. Maddox, J. V. Filoteo, K. D. Hejl, A. D. Ing, Category number impacts rule-based but not information-integration category learning: Further evidence for dissociable category-learning systems. *J. Exp. Psychol. Learn. Mem. Cogn.* **30**, 227–245 (2004).
127. S. J. Miles, K. Matsuki, J. P. Minda, Continuous executive function disruption interferes with application of an information integration categorization strategy. *Atten. Percept. Psychophys.* **76**, 1318–1334 (2014).
128. J. P. Minda, R. Rabi, Ego depletion interferes with rule-defined category learning but not non-rule-defined category learning. *Front. Psychol.* **6**, 35 (2015).
129. C. Quam, A. Wang, W. T. Maddox, K. Golisch, A. Lotto, Procedural-memory, working-memory, and declarative-memory skills are each associated with dimensional integration in sound-category learning. *Front. Psychol.*, 10.3389/fpsyg.2018.01828/full (2018).
130. E. M. Waldron, F. G. Ashby, The effects of concurrent task interference on category learning: Evidence for multiple category learning systems. *Psychon. Bull. Rev.* **8**, 168–176 (2001).
131. R. M. Nosofsky, J. K. Kruschke, Single-system models and interference in category learning: Commentary on Waldron and Ashby (2001). *Psychon. Bull. Rev.* **9**, 169–174 (2002).
132. C. L. Huang-Pollock, W. T. Maddox, S. L. Karalunas, Development of implicit and explicit category learning. *J. Exp. Child Psychol.* **109**, 321–335 (2011).
133. R. Rabi, S. J. Miles, J. P. Minda, Learning categories via rules and similarity: Comparing adults and children. *J. Exp. Child Psychol.* **131**, 149–169 (2015).
134. R. Rabi, J. P. Minda, Rule-based category learning in children: The role of age and executive functioning. *PLoS One* **9**, e85316 (2014).
135. C. L. Roark, E. Lescht, A. Hampton Wray, B. Chandrasekaran, Auditory and visual category learning in children and adults. *Dev. Psychol.* **59**, 963–975 (2023).
136. J. P. Minda, A. S. Desroches, B. A. Church, Learning rule-described and non-rule-described categories: A comparison of children and adults. *J. Exp. Psychol. Learn. Mem. Cogn.* **34**, 1518–1533 (2008).
137. K. L. Carpenter, A. J. Wills, A. Benattayallah, F. Milton, A Comparison of the neural correlates that underlie rule-based and information-integration category learning. *Hum. Brain Mapp.* **37**, 3557–3574 (2016).
138. F. Milton, P. Bealing, K. L. Carpenter, A. Bennattayallah, A. J. Wills, The neural correlates of similarity- and rule-based generalization. *J. Cogn. Neurosci.* **29**, 150–166 (2017).
139. B. R. Buchsbaum, S. Greer, W. L. Chang, K. F. Berman, Meta-analysis of neuroimaging studies of the Wisconsin Card-Sorting task and component processes. *Hum. Brain Mapp.* **25**, 35–45 (2005).
140. B. Milner, Effects of different brain lesions on card sorting: The role of the frontal lobes. *Arch. Neurol.* **9**, 90–100 (1963).
141. J. D. Wallis, K. C. Anderson, E. K. Miller, Single neurons in prefrontal cortex encode abstract rules. *Nature* **411**, 953–956 (2001).
142. L. T. Hunt, B. Y. Hayden, A distributed, hierarchical and recurrent framework for reward-based choice. *Nat. Rev. Neurosci.* **18**, 172–182 (2017).
143. S. E. Cavanagh, L. T. Hunt, S. W. Kennerley, A diversity of intrinsic timescales underlie neural computations. *Front. Neural Circ.* **14**, 615626 (2020).
144. J. Russin, R. C. O'Reilly, Y. Bengio (2020). "Deep learning needs a prefrontal cortex" in *Bridging AI and Cognitive Science (BAICS) Workshop*, J. Hamrick et al., Eds. (ICLR 2020), pp. 1-11.
145. A. Traylor, J. Merullo, M. J. Frank, E. Pavlick (2024). "Transformer mechanisms mimic frontostriatal gating operations when trained on human working memory tasks" in Proceedings of the Annual Meeting of the Cognitive Science Society, L. K. Samuelson, S. Frank, M. Toneva, A. Mackey, E. Hazeltine, Eds. (CogSci, 2024), pp. 1516-1523.
146. A. Soni, A. Traylor, J. Merullo, M. J. Frank, E. Pavlick, Transformer mechanisms mimic frontostriatal gating operations when trained on human working memory tasks. OpenReview.net [Preprint] (2024). https://openreview.net/pdf/84db2371ccf7431d6f8fa3811a92d6fa5ff3511f.pdf (Accessed 12 August 2025).
147. D. P. Kingma, J. Ba, "Adam: A method for stochastic optimization" in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings*, Y. Bengio, Y. LeCun, Eds. (ICLR, 2015).
148. J. Russin, Parallel trade-offs in human cognition and neural networks: The dynamic interplay between in-context and in-weight learning [Python]. GitHub. https://github.com/jlrussin/icl-iwl-interplay. Deposited 30 April 2025.