In-Context Learning Agents Are Asymmetric Belief Updaters

Johannes A. Schubert[1] Akshay K. Jagadish[1 2] Marcel Binz[* 1 2] Eric Schulz[* 1 2]

Abstract

We study the in-context learning dynamics of large language models (LLMs) using three instrumental learning tasks adapted from cognitive psychology. We find that LLMs update their beliefs in an asymmetric manner and learn more from better-than-expected outcomes than from worse-than-expected ones. Furthermore, we show that this effect reverses when learning about counterfactual feedback and disappears when no agency is implied. We corroborate these findings by investigating idealized in-context learning agents derived through meta-reinforcement learning, where we observe similar patterns. Taken together, our results contribute to our understanding of how in-context learning works by highlighting that the framing of a problem significantly influences how learning occurs, a phenomenon also observed in human cognition.

1. Introduction

Large language models (LLMs) are powerful artificial systems that excel at many tasks (Radford et al., 2019). They can, among other things, write code (Roziere et al., 2023), help to translate from one language to another (Kocmi & Federmann, 2023), and play computer games (Wang et al., 2023). Their abilities are so far-reaching that some (Bubeck et al., 2023) have argued that they "could reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence (AGI) system". At the same time, they are notoriously difficult to interpret which becomes especially aggravating as these models permeate through our society.

In the present paper, we aim to shed light on the in-context learning abilities of LLMs (Brown et al., 2020). Making use of the two-alternative forced choice (2AFC; Fechner, 1860) paradigm from cognitive science, we show that in-context learning implements an asymmetric updating rule when learning about the values of options. In particular, we find that – when provided with outcomes from freely chosen options – in-context learning exhibits an optimism bias (Sharot, 2011), meaning that it learns more from positive than from negative prediction errors. We additionally find that this effect is mediated by two factors. First, when the outcome of the unchosen option is also provided, the bias for that option reverses and the model learns more from negative than from positive prediction errors. Furthermore, when no agency is implied and the query says *someone else* does the sampling instead of *you* sampled, the bias disappears. Interestingly, similar behavioral effects have been observed in human subjects (Lefebvre et al., 2017; Chambon et al., 2020).

Why do these tendencies for asymmetric belief updating emerge in both natural and artificial agents? Previous work has suggested that asymmetric belief updating might be a rational strategy to implement as it allows agents to achieve maximum rewards in a given task. However, these claims have been limited by the use of a restricted model class (Lefebvre et al., 2022; Cazé & van der Meer, 2013). To investigate this idea further, we study the behavior of idealized in-context learning agents trained specifically to solve 2AFC tasks using meta-reinforcement learning (Meta-RL). Meta-RL agents have been shown to implement Bayes-optimal learning strategies upon convergence (Ortega et al., 2019; Binz et al., 2023b) and enable us to test if displaying such a bias is rational. We again find the same behavioral effects in these agents: (1) they show an optimism bias when only observing outcomes of the chosen option, (2) the bias reverses when learning about the value of the unchosen option, and (3) it disappears when the agent has no control about its own choices.

Taken together, our results have broad implications for both natural and artificial agents. We have shown that in-context learning depends critically on how the problem is framed. There are many applications where practitioners have control over problem framing, and thus our results suggest that these design choices must be carefully considered to achieve desired outcomes. In the context of human cognition, our

[*] Equal contribution [1] Computational Principles of Intelligence Lab, Max Planck Institute for Biological Cybernetics, Tübingen, Germany [2] Institute for Human-Centered AI, Helmholtz Computational Health Center, Munich, Germany. Correspondence to: Johannes A. Schubert <johannes.schubert@tue.mpg.de>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

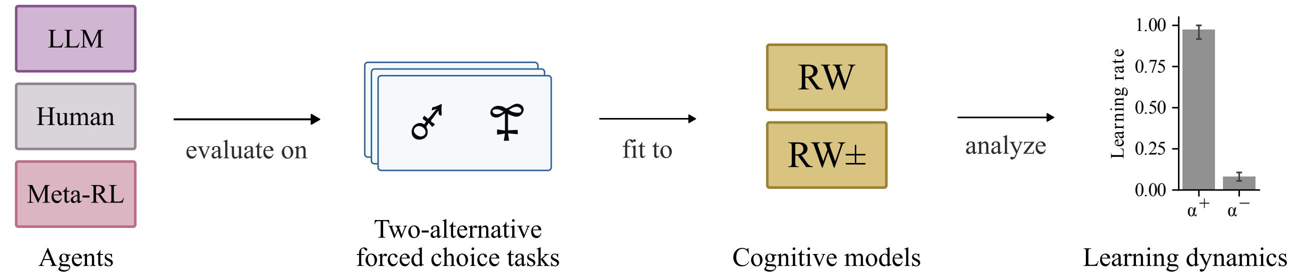

Figure 1. Schematic of our methodology, where we evaluate the learning dynamics of LLMs, humans, and meta-reinforcement learning (Meta-RL) agents on two-alternative forced choice tasks. After evaluating the agents on the tasks, we fit variants of cognitive models based on the Rescorla-Wagner (RW) model to the resulting behavior. Finally, we analyze the fitted models and extract and compare the learning rates.

simulations extend previous work suggesting that the optimism bias (and related effects) may not be a bias after all, as it can be considered a rational response to certain problems.

2. Related work

In-context learning in LLMs. In their seminal paper introducing GPT-3, Brown et al. (2020) demonstrated that LLMs can do tasks, such as text translation, question-answering, and arithmetic problems, after seeing just a few demonstrations – an ability they referred to as in-context learning. There has been a recent surge of research trying to better understand this phenomenon: investigating when and how in-context learning emerges (Chan et al., 2022; Min et al., 2022; Wei et al., 2023), identifying what algorithms LLMs implement during in-context learning (Xie et al., 2021; Von Oswald et al., 2023), and mapping out what can be learned in-context (Garg et al., 2022; Dong et al., 2022). For the present paper, the work of Binz & Schulz (2023) and Coda-Forno et al. (2023) is of particular relevance. They showed that LLMs can learn to perform simple multi-armed bandit problems – which were similar to the 2AFC problems used here – in-context and without requiring any weight updates.

Asymmertric belief updating in cognitive science. How humans integrate task-relevant information to update their beliefs has received a lot of attention in cognitive science (Jacobs & Kruschke, 2011; Nassar et al., 2010; Gershman, 2015). Traditionally, this question has been investigated using the 2AFC paradigm, which provides a controlled setup to study the various facets of human reinforcement learning, including generalization, exploration, and compositional inference (Jagadish et al., 2023; Binz & Schulz, 2022; Lefebvre et al., 2017; Behrens et al., 2007; Chambon et al., 2020; Palminteri & Lebreton, 2022; Schulz et al., 2019).

The experimental paradigms and analyses used in this paper are heavily inspired by the following studies. Lefebvre et al. (2017) showed that people have asymmetric belief updating tendencies in a reinforcement learning setting. Later on, Chambon et al. (2020) demonstrated that this tendency reverses when people observe outcomes for unchosen options and that it completely disappears in purely observational trials (i.e. when participants observe outcomes following a predetermined choice). Recently, Palminteri & Lebreton (2022) reviewed the influence of factors, such as reward magnitude (Lefebvre et al., 2022) and volatility (Gagne et al., 2020; Behrens et al., 2007), on asymmetric belief updating in people. Finally, Lefebvre et al. (2022) used simulations to show that asymmetric belief updating can lead to optimal performance under certain reward regimes. In a series of experiments, the authors manipulated the reward probabilities of options. They found that an optimism bias is optimal for low reward probabilities, while a pessimism bias is optimal for high reward probabilities. This occurs as the learning asymmetry enables better separation of learned reward probabilities in their respective reward regimes and decreases the probability of switching to the worse option after a certain amount of trials (Cazé & van der Meer, 2013).

3. Methods

In this section, we first describe how we queried an LLM to perform 2AFC tasks and explain how models from cognitive psychology can be used to analyze the in-context learning dynamics of LLMs and idealized agents derived through Meta-RL. We provide an overview of our methodology in Figure 1.

3.1. LLM prompting

In a 2AFC task, an agent has to repeatedly choose between two options and receives a reward after each choice. The goal of the agent is to maximize the reward over all trials. We prompted an LLM to perform such 2AFC tasks. We used Claude-1.2 as the reference LLM for all our experiments via its API[1] with the temperature set to 0.0.

The prompt design was based on earlier work that had studied LLMs in similar settings (Binz & Schulz, 2023; Coda-Forno et al., 2023). Each prompt included an introduction to the task setup, a history of previous observations, and a question asking for the next choice. The tasks were framed in a gambling context, where an agent visits several casinos with two slot machines. More details on the prompting are provided in Appendix A.1.

The following prompt was used to run the default 2AFC task (see Section 4):

Prompt Task 1

> You are going to visit four different casinos (named 1, 2, 3, and 4) 24 times each. Each casino owns two slot machines which all return either 0.5 or 0 dollars stochastically with different reward probabilities. Your goal is to maximize the sum of received dollars within 96 visits.
>
> You have received the following amount of dollars when playing in the past:
>
> - Machine B in Casino 4 delivered 0.5 dollars.
> - Machine F in Casino 1 delivered 0.0 dollars.
> - Machine B in Casino 4 delivered 0.5 dollars.
>
> Q: You are now in visit 4 playing in Casino 4. Which machine do you choose between Machine R and Machine B?
>
> A: Machine [insert]

Prompts were updated dynamically after every trial. Slot machines were labeled with a random letter, excluding meaningful ones (U, I), and the order of slot machine labels was randomized. The selected slot machine returned a stochastic reward that was appended to the bulleted history of previous slot machine interactions in subsequent prompts. We simulated the behavior of the LLM for a certain number of trials and recorded the action-reward pairs.

We additionally considered two other variants of 2AFC tasks, which shared the same general structure but differed in how information was conveyed to the agent. One provided additional information by revealing the reward for the unchosen option (see Section 5) and the other included trials in which the choice of a particular option was forced (see Section 6).

3.2. Using cognitive models to analyze in-context learning dynamics

For the analysis of the learning dynamics of an agent, we built simplified but interpretable cognitive models of its choice behavior. We then identified which parameter setting in these models provides the best explanation for the observed data. The resulting parameter values can then offer a window into the behavior of the agent. This approach has its origins in cognitive science (Miller, 1956; Rescorla & Wagner, 1972; Wilson & Collins, 2019) but has recently been adopted to study the behavior of artificial agents as well (Dasgupta et al., 2022; Binz & Schulz, 2023; Bigelow et al., 2023).

The core of our analysis is the Rescorla-Wagner (RW) model (Rescorla & Wagner, 1972) – a classic model for studying learning in humans [2]. It formalizes learning as a dynamic process of minimizing prediction errors between expected and actual outcomes:

$$V_{t+1}(a) = V_t(a) + \alpha \cdot \delta_t$$
$$\delta_t = r_t - V_t(a)$$

where δ_t is the prediction error between the observed reward r_t and the expected reward value $V_t(a)$ in trial t. The learning rate α determines how much the expected value for action a changes after an observation. Thus, it quantifies the amount of learning that occurs based on the prediction error. The model maps learned values to choice probabilities using a softmax decision rule with an inverse temperature parameter β:

$$p(a) = \frac{\exp(\beta \cdot V_t(a))}{\sum_{k=1}^{K} \exp(\beta \cdot V_t(k))}$$

There are several extensions to the RW model that have been used to evaluate different hypotheses about how learning occurs. We relied on one such extension, namely the RW± model (Palminteri & Lebreton, 2022), which introduces separate learning rates for positive and negative prediction errors:

$$V_{t+1}(a) = V_t(a) + \begin{cases} \alpha^+ \cdot \delta_t, & \text{if } \delta_t > 0 \\ \alpha^- \cdot \delta_t, & \text{if } \delta_t < 0 \end{cases}$$
$$\delta_t = r_t - V_t(a)$$

Positive prediction errors occur in situations where the received reward is greater than the estimated value (i.e. $\delta_t > 0$), while negative prediction errors occur when the received reward is less than the estimated value (i.e. $\delta_t < 0$).

[1] https://www.anthropic.com/news/introducing-claude

[2] Note that while the RW model is considered one of the most canonical modeling choices for 2AFC tasks, other alternatives have been used to model human behavior in this setting, such as Bayesian models (Zhang & Yu, 2013; Gershman, 2018) and drift-diffusion models (Pedersen et al., 2017; Lefebvre et al., 2022).

This model allows us to study how the information of prediction errors is weighted during learning. In the case where the expected value is influenced symmetrically by both prediction errors (i.e. $\alpha^+ = \alpha^-$), the RW± model is equivalent to the classical RW model. If the learning rates differ, the beliefs are updated asymmetrically, either optimistically ($\alpha^+ > \alpha^-$) or pessimistically ($\alpha^- > \alpha^+$).

We relied on a maximum a posteriori estimation approach to fit the parameters of these models to the behavioral data generated by an LLM. For the RW model, this involves fitting parameters $\theta = [\alpha, \beta]$ (two parameters in total). For the RW± model, this involves fitting parameters $\theta = [\alpha^+, \alpha^-, \beta]$ (three parameters in total). Prior probabilities were based on Daw et al. (2011), using a beta distribution $\mathcal{B}(1.1, 1.1)$ for learning rates and a gamma distribution with $\mathcal{G}(1.2, 5)$ for the inverse temperature parameter. Each option was represented by a separate expected value V_t and the initial values V_0 were assigned to the average reward values. More details on the parameter estimation can be found in Appendix A.2.

For each task, we compared different cognitive models based on their average posterior probabilities (PP; Wu et al., 2020). To do this, we first computed the Bayesian Information Criterion (BIC; Schwarz, 1978) for each model m and each simulation:

$$\text{BIC}_m = k \cdot \ln(N) - 2 \cdot \ln p(a_{1:N}|\hat{\theta}, m)$$

where $\hat{\theta}$ are the estimated parameters based on the agent's choices, k is the number of estimated parameters, and N is the number of choices performed. Under the assumption of a uniform prior over models, the PP for each simulation can then be approximated as:

$$p(m|a_{1:N}) \approx \frac{\exp(-0.5 \cdot \text{BIC}_m)}{\sum_i \exp(-0.5 \cdot \text{BIC}_i)}$$

The BIC is a standard metric for model comparison and selection. It can be interpreted as an approximation to the model evidence (or marginal likelihood) which is obtained by integrating out the parameters of a model using Laplace's method (Bishop, 2006).

4. Partial information results in optimism bias

We started our investigations with the 2AFC task with default settings in which only the outcome for the chosen slot machine was revealed as illustrated in the prompt of the previous section and in Figure 2a. We adopted the experimental paradigm of an earlier study with humans from Lefebvre et al. (2017). It involved four different casinos with win probabilities of 0.25/0.25, 0.25/0.75, 0.75/0.25, and 0.75/0.75 for the respective slot machines. Winning led to a reward of 0.5 dollars, losing to a reward of 0.0 dollars. Each casino was visited 24 times in a random order resulting in 96 visits in total. We simulated the LLM 50 times on the task.

First, we examined performance regarding regret, which is the amount of reward missed relative to the optimal choice. When averaged across all simulations and casinos, regret decreased significantly from 0.09 ± 0.01 in the first trial to 0.05 ± 0.01 in the last trial ($t(199) = 3.1, p = .002$; see Figure 2b). This difference between starting and final performance during in-context learning is similar to that observed in human studies (Lefebvre et al., 2017).

To investigate whether in-context learning is done symmetrically or asymmetrically, we fitted the classical RW and the RW± model to the simulated behavior of the LLM separately for each simulated run as described in Section 3.2. The model comparison indicated that the RW± model provided on average a better explanation of the data with a PP of 0.94 ± 0.03 (see Figure 2c). Analyzing the learning rates revealed that this was associated with a strong optimistic asymmetry: new information about the options was incorporated more readily when it was desirable (positive prediction errors) than undesirable (negative prediction errors) as shown in Figure 2d, with $\alpha^+ > \alpha^-$ ($\alpha^+ = 0.86 \pm 0.02, \alpha^- = 0.08 \pm 0.03$; $t(49) = 24.2, p < .0001$). In-context learning seems to overweigh new evidence that conveys a positive valence. Previous work with human subjects has observed a similar – although less pronounced – optimism bias (reproduced in Figure 2d for reference).

We additionally tested how well our findings generalize to different LLMs and task formats. We therefore repeated our experiment on seven LLMs, including Claude-2.1, Claude-3 Haiku, GPT-4, Llama-2-7B, Llama-2-7B-Chat, Llama-2-70B, and Llama-2-70B-Chat. Furthermore, we extended our analysis to include a broader range of task formats. In all of these settings, we found robust evidence for asymmetric belief updating. The results of these analyses are presented in Appendix B.

5. Pessimistic updating for unchosen options

Next, we investigated the in-context learning dynamics of LLMs in a setting that provided full feedback about the outcomes of both the chosen and unchosen slot machines. For this, we borrowed another experimental paradigm of an earlier study with humans from Chambon et al. (2020). The adapted task consisted of visits to multiple casinos, each containing two slot machines. Each of the casinos was prompted independently of the others. Half of the casinos provided full feedback, the other half only provided partial

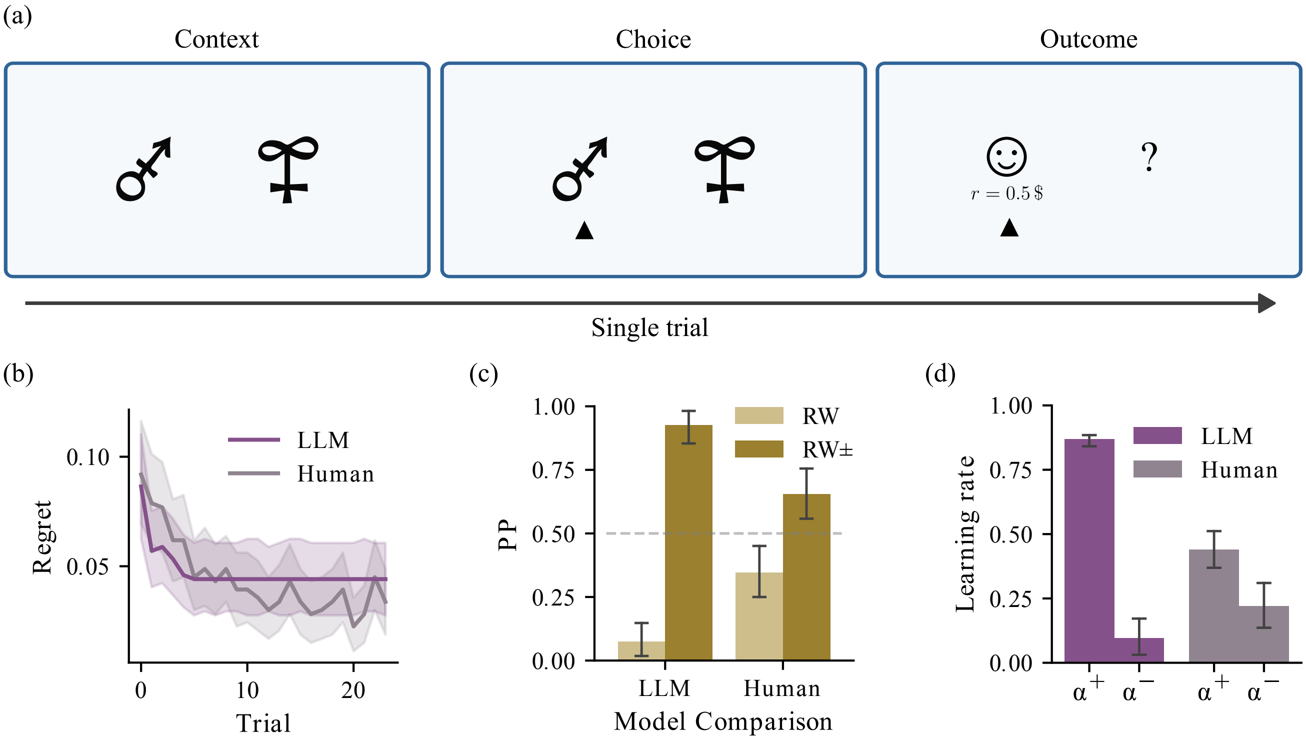

Figure 2. 2AFC task with partial feedback. (a) Presentation of a single trial. First, two slot machines, shown as symbols, are presented. After a choice is made, the outcome is shown. (b) Average performance of the LLM and humans measured in terms of regret. Performance improves over trials. (c) Model comparison of the Rescorla-Wagner (RW) model and the RW± model. For both the LLM (left) and human participants (right), RW± provides a better fit to the data, as indicated by the average posterior probability (PP). (d) Average learning rates of the RW± model for the LLM and human participants. Both agents show a stronger response to positive prediction errors than to negative prediction errors. Human participant data reproduced from (Lefebvre et al., 2017). Error bars and shaded areas and correspond to 95% CIs.

feedback about the chosen machine.[3] In the full feedback casinos, the foregone outcome of the unchosen slot machine was provided in addition to the outcome of the chosen slot machine in the history of prior interactions (see Figure 3a). We adapted the feedback items as follows to reflect this change: “On visit 1 you played Machine H and earned 1.0 point. On Machine E you would have earned -1.0 point.”

Half of the casinos were high-reward casinos with reward probabilities of 0.9 and 0.6 for both machines, and the other half were low-reward casinos with reward probabilities of 0.4 and 0.1. In all casinos, the outcome was either a gain or a loss of one point. The prompt in the full feedback casinos also mentioned that the observed outcome of the unchosen machine would not be added to the total rewards earned (see Appendix C.1). We simulated the LLM 24 times on the task consisting of 16 casinos with 40 trials each.

When comparing the partial to the full feedback casinos, we saw a decrease in final regret from partial feedback (0.07 ± 0.01) to full feedback casinos (0.02 ± 0.01), with $t(191) = 2.9, p = .004$ (see Figure 3b). This suggests that in-context learning was able to incorporate the additional information conveyed by the unchosen option. Like in the previous section, performance in terms of regret was comparable to that observed in an earlier study with human subjects as shown in Figure 3c.

For analyzing the dynamics of in-context learning, we extended the RW± model to also account for the additional information provided by the unchosen option. This extended RW± model included separate learning rates for positive and negative prediction errors of the chosen and unchosen options (i.e. four learning rates in total).

Figure 3d illustrates that fitted learning rates for chosen and unchosen options show opposite asymmetric patterns. While we again observed an optimism bias for learning about the chosen machine ($t(23) = 5.7, p < .0001$), information for the unchosen machine was integrated such that negative prediction errors were preferentially taken into account, relative to positive ones ($t(23) = 5.7, p < .0001$).

[3] Note that the task also contained forced-choice trials in which the agent has to select a predetermined machine. We ignored these trials for the analysis presented in this section, but come back to them in the next section.

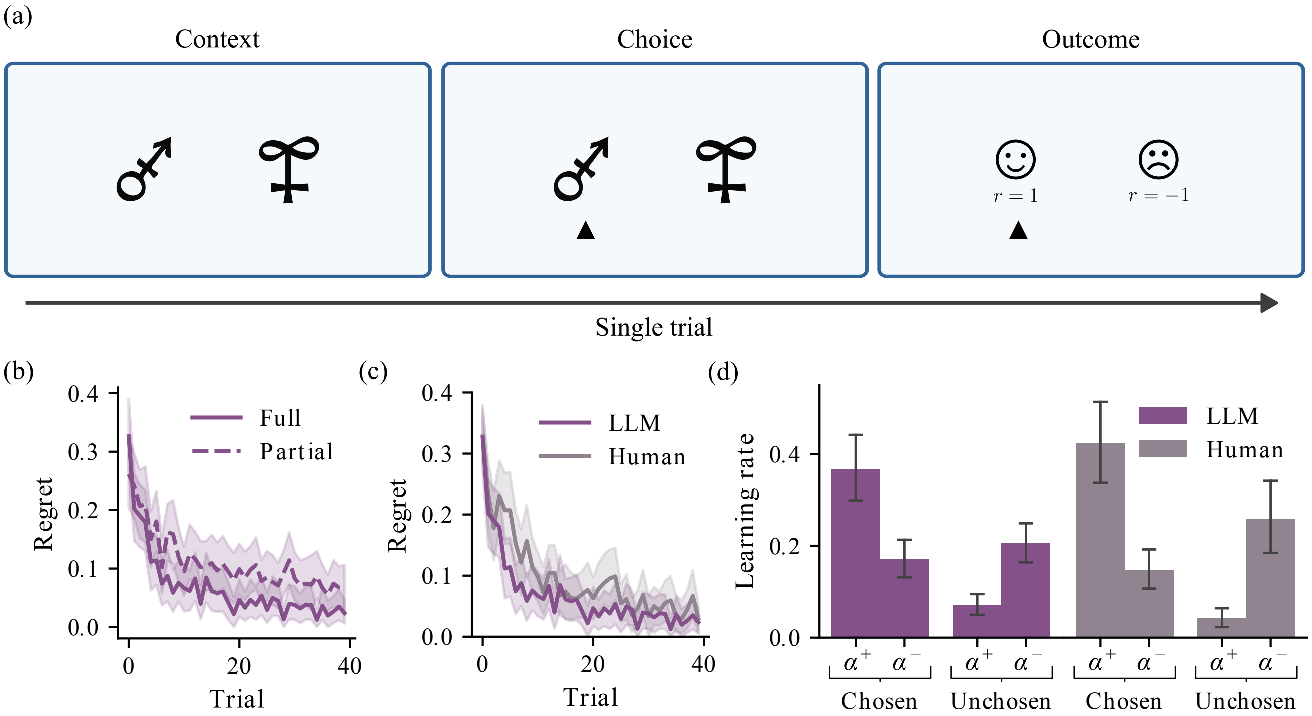

Figure 3. 2AFC task with full feedback. (a) Presentation of a single trial: The two slow machines are again shown as symbols. After a choice is made, the outcome of both the chosen and the unchosen option is shown. (b) Average regret of the LLM in partial and full feedback casinos, showing that the additional information of full feedback casinos leads to improved performance. (c) Average regret of the LLM and humans for full feedback casinos. The performance of the LLM improves over trials and is on par with human performance. (d) Learning rates of the full feedback model with two learning rates – for positive and negative prediction errors – for the chosen and unchosen slot machine. Both agents have an optimism bias for the chosen option and a pessimism bias for the unchosen option. Human participant data reproduced from (Chambon et al., 2020). Error bars and shaded areas correspond to 95% CIs.

This pattern is known as confirmation bias as it refers to integrating information in a way that confirms prior beliefs (Palminteri et al., 2017; Nickerson, 1998). Earlier studies with people in a matched experiment (Chambon et al., 2020) found similar behavioral characteristics (see Figure 3d for reference).

The pattern of learning rates suggests that the four learning rates can be compressed into just two: a confirmatory learning rate that combines the positive chosen and negative unchosen learning rates and a disconfirmatory learning rate that combines the negative chosen and positive unchosen learning rates. We therefore fitted a second cognitive model that was obtained by merging the learning rates depending on the confirmatory or disconfirmatory outcome. This simplified model provided a better fit to the data with an average PP of 0.93 ± 0.03 (refer to Figure 8 for the learning rates of this model). This implies that LLMs update their beliefs about certain outcomes more when new evidence confirms their prior beliefs and past decisions than when it disconfirms or contradicts them.

6. No asymmetric updating without agency

In our final analysis, we examined the influence of agency on the in-context learning dynamics of LLMs by providing additional information about slot machines through observational trials. We used the same general task structure as in the previous section (Chambon et al., 2020). However, half of the visited casinos now included randomly interleaved forced-choice trials of another agent playing in the casino (mixed-choice casinos; see Figure 4a). The other half contained only free-choice trials to assess the performance improvements resulting from the additional information provided by the forced-choice trials in the mixed-choice casinos. Both types of casinos provided partial feedback about the outcome of the selected machine.

We adapted the prompt structure of the force-choice trials as follows: "On visit 1 someone else played Machine H and received 1.0 point". To avoid biasing the agent towards a particular machine, the forced-choice trials sampled both slot machines equally. The prompt of the mixed-choice casinos mentioned that rewards from forced-choice trials would not be added to the total reward (see Appendix D.1). As in the previous section, half of the casinos were high-

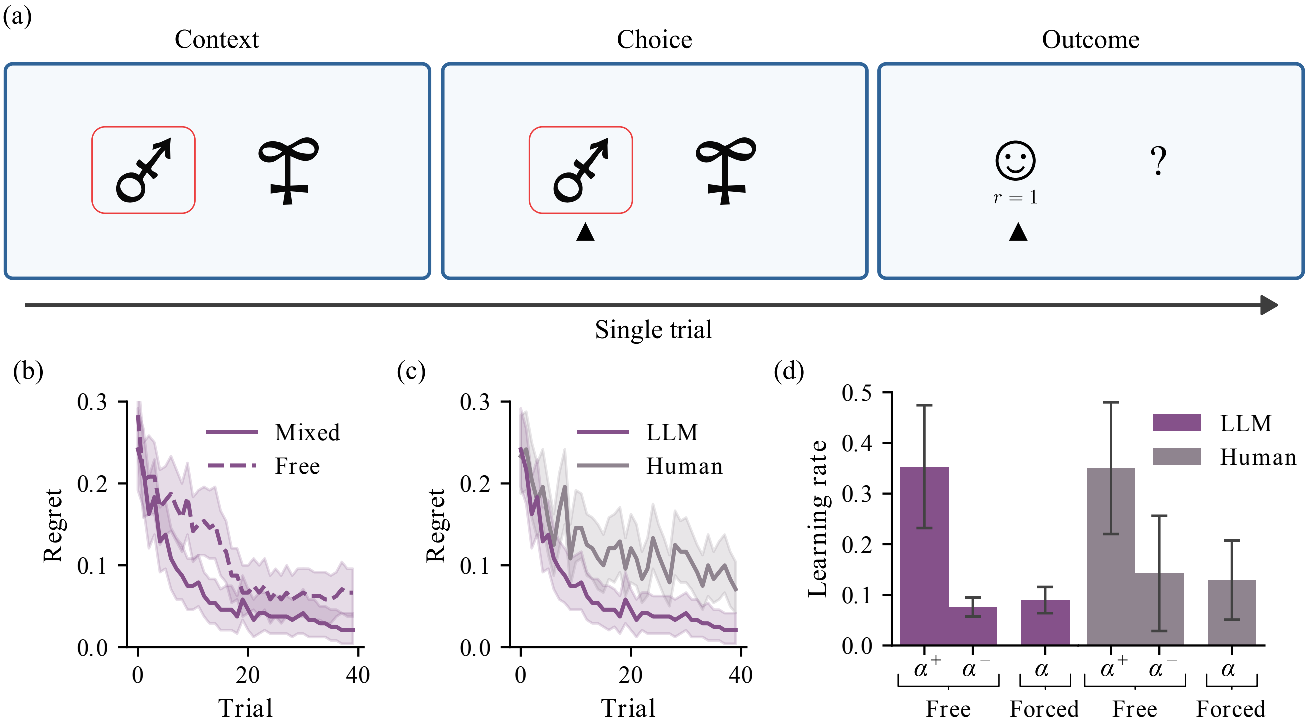

Figure 4. 2AFC task for the agency condition (a) Presentation of a single forced-choice trial: In forced-choice trials, one of the two slot machines is preselected (red square) and its outcome is presented directly to the LLM. (b) Average regret of the LLM in mixed-choice and free-choice casinos, showing that the additional information of forced-choice trials leads to improved performance in mixed-choice casinos. (c) Average LLM and human regret for mixed-choice casinos. The performance of the LLM outperforms the human participants over trials. (d) Learning rates of the 3α model with two learning rates for the free-choice trials and one learning rate for the forced-choice trials. Both agents integrate feedback for positive and negative prediction errors in free-choice trials asymmetrically, whereas feedback from forced-choice trials is integrated symmetrically. Human participant data reproduced from (Chambon et al., 2020). Error bars and shaded areas correspond to 95% CIs.

reward casinos, the other half were low-reward casinos. The six mixed-choice casinos consisted of 80 trials with 40 free and 40 forced-choice trials, while the six free-choice casinos consisted of only 40 free-choice trials. We simulated the LLM on this task 24 times.

When comparing the free-choice casinos with the mixed-choice casinos, we found that the LLM incorporated the additional information from the forced-choice trials, leading to performance improvements with a decrease in regret from 0.07 ± 0.2 to 0.02 ± 0.01 ($t(143) = 2.5, p = 0.01$; see Figure 4b). In comparison to human data from an earlier study (Chambon et al., 2020), the LLM learned significantly faster in this setting as shown in Figure 4c (final regret for LLMs: 0.02 ± 0.01, final regret for humans: 0.07 ± 0.02; $t(143) = 2.7, p = .008$).

To analyze the effects of choice types on learning we fitted two different cognitive models to the behavior of the LLM in the mixed-choice casinos: (1) a 4α model which consisted of separate learning rates for positive and negative prediction errors for both free-choice and forced-choice trials and (2) a 3α model which consisted of separate learning rates for positive and negative prediction errors for free-choice trials and only one learning rate for forced-choice trials. The model comparison indicated that the 3α model represented the behavior of the LLM best ($\text{PP}_{3\alpha} = 0.78 \pm 0.06$; see Figure 4c), suggesting that it seems to integrate the information from forced-choice trials symmetrically (i.e. $\alpha^+ = \alpha^-$). In contrast, information from free-choice trials is asymmetrically weighted with $\alpha^+ = 0.36 \pm 0.06$ being greater than $\alpha^- = 0.08 \pm 0.01$ ($t(23) = 4.7, p = .0001$), as seen in Figure 4d. This implies the extent to which the LLM can control its environment changes how it integrates received information.

7. Idealized in-context learning agents also display asymmetric updating

To better understand why in-context learning exhibits these behavioral characteristics, we tested whether they also emerge in a more controlled setting. For this, we trained idealized transformer-based agents to solve our previously examined 2AFC tasks via in-context learning. The agent

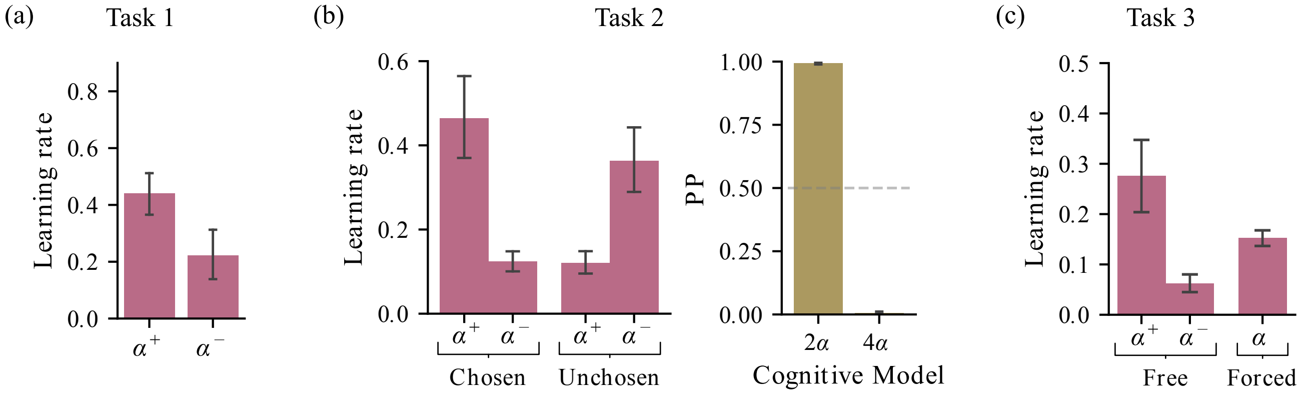

Figure 5. Learning rate analyses for the Meta-RL agent. (a) In the partial feedback task, the RW± provided a better fit to the Meta-RL agents' behavior ($\text{PP}_{\text{RW}\pm} = 0.97 \pm 0.02$) and showed an optimistic tendency to integrate information. (b) In the full feedback task, the tendency to integrate positive outcomes for chosen options optimistically and negative outcomes for unchosen options pessimistically is even more pronounced than in the LLMs (left). Model comparison showed that the simplified confirmatory model (2α) fits the data better (right). (c) In the agency condition, the 3α best fit the simulated behavior ($\text{PP}_{3\alpha} = 0.85 \pm 0.04$), implying that information was integrated asymmetrically in free-choice trials and symmetrically in forced-choice trials. Error bars correspond to 95% CIs.

received the previously selected action a_{t-1} and the reward that followed r_{t-1}, alongside the current context c_t which varied slightly for each task (see Appendix E.1), as inputs in each time-step. We trained the agent using Meta-RL (Duan et al., 2016; Wang et al., 2016) to learn a history-dependent policy $\pi_{\mathbf{\Theta}}$ that maximizes the expected sum of rewards over a prespecified task distribution:

$$\max_{\mathbf{\Theta}} \mathbb{E}_{p(\theta, c_{1:T}) \prod p(r_t|a_t, c_t, \theta) \pi_{\mathbf{\Theta}}(a_t|a_{1:t-1}, r_{1:t-1}, c_{1:t})} \left[\sum_{t=1}^{T} r_t \right]$$

where $a_{1:t-1}$, $r_{1:t-1}$, and $c_{1:t}$ denote sequences of actions, rewards, and contexts respectively, while $\mathbf{\Theta}$ denotes the weights of the underlying transformer. The task distribution used for training is specified by $p(\theta, c_{1:T})$ and $p(r_t|a_t, c_t, \theta)$ with θ corresponding to a vector containing win probabilities for all slot machines.

The distribution shared a similar structure in all three experiments. We assumed that win probabilities for each option were sampled from a uniform distribution at the beginning of each training episode to capture the assumption of an uninformative prior. The agent consisted of a Transformer network (Vaswani et al., 2017) with a model dimension of $8 \cdot \text{input\_size}$, two feedforward layers with a dimension of 128, and eight attention heads, followed by two linear layers that output a policy and a value estimate, respectively. The network weights were adjusted by gradient-based optimization using ADAM (Kingma & Ba, 2014) on a standard actor-critic loss (Mnih et al., 2016) at the end of each training episode. Further details about the training are provided in Appendix E.

We trained all Meta-RL agents until convergence and then tested their in-context learning abilities without performing any further updates to the network weights. It has been shown that the history-dependent policy learned in this setting can solve new but similar tasks in an approximately Bayes-optimal way (Mikulik et al., 2020; Ortega et al., 2019). The intuition behind this result is that the Meta-RL protocol incentivizes the agent to maximize a Monte-Carlo approximation of the Bayes-optimal objective. While analytical solutions to this objective are intractable for most cases, Meta-RL allows us to derive a tractable approximation, and thereby to investigate belief updating in an idealized setting.

We simulated the agents on all three tasks and analyzed their behavior as described in the previous sections. We found that the idealized transformer-based agents learned strategies that outperformed both LLMs and humans as shown in Figure 11. Furthermore, the idealized agents showed similar learning characteristics to LLMs: (1) in the partial feedback task, they learned more from positive prediction errors, (2) the pattern reversed for the counterfactual option in the case of the full feedback task, and (3) asymmetric updating was limited to free-choice trials and was absent in forced-choice trials. These results are illustrated in Figure 5.

Taken together, these results indicate that behavior observed in humans and LLM shares key characteristics with idealized in-context learning agents trained specifically on 2AFC tasks.

8. Discussion

People change their learning strategies based on how the problem is framed (Palminteri & Lebreton, 2022). In this paper, we have shown that this also holds for in-context learning agents. In particular, we found that LLMs exhibit an optimism bias, i.e. they learn more from better-than-expected outcomes (positive prediction errors) than from worse-than-expected ones (negative prediction errors). However, this bias was only present when the prompt was formulated in a way that implied agency. Furthermore, we found that for counterfactual feedback for unchosen options, the bias reversed and the model learned more from negative than positive errors for these options.

We conducted these analyses in a highly controlled setting, providing high internal validity to our results. However, claims regarding their external validity must be taken with care for now, and future studies will have to investigate whether we can also find similar patterns in more naturalistic settings. Furthermore, our findings relied on an inference from observed learning rate patterns in cognitive models. It is unclear if the link is causal as another auxiliary computational process could potentially be explaining the observed pattern. Nevertheless, opposing interpretations of the pattern have not yet been substantiated by experimental evidence in human experiments (Palminteri & Lebreton, 2022). To test the causal relationship between the latent processes underlying cognitive models and those of the LLM, we see two possibilities. The first is to lesion the neurons in the LLM that encode trial-by-trial negative prediction errors to determine if it results in a stronger optimistic bias. The second is to swap the activations between positive and negative prediction error encoding neurons to determine if it results in a pessimism bias instead of an optimism bias.

In the tasks we have investigated, it is rational to perform asymmetric belief updating – as indicated by our simulations with Meta-RL agents. It remains an open question if this also holds in situations where asymmetric belief updating is suboptimal. Future work should aim to characterize whether in-context learning also displays asymmetric belief updating in such situations. The study of Globig et al. (2021) could be a starting point for this question as it provides an example of a setting where people show an optimism bias even though it is not rational.

From a methodological perspective, our work demonstrates that it is possible to fit simpler computational models to the behavior of LLMs and use the resulting parameter values to infer *how* they behave. Fitting and interpreting parameters of simpler computational methods provides us with a tool that complements existing techniques for explainable machine learning (Roscher et al., 2020). We believe that there are further exciting applications in this research field.

Taken together, our results contribute to our understanding of how in-context learning in LLMs works, which is especially important as the number of applications of these models in real-world scenarios is increasing (Binz et al., 2023a; Eloundou et al., 2023; Kasneci et al., 2023). If the biases found in this work also emerge in tasks where they are not optimal, as has been shown in humans (Shepperd et al., 2013), it will be important to develop techniques to mitigate them.

Acknowledgements

We thank the members of the "Computational Principles of Intelligence Laboratory" (CPI Lab), and Stefano Palminteri for their comments, discussions, and support. We would also like to thank the authors of Lefebvre et al. (2017), and Chambon et al. (2020) for making the data from their study available. Further, we thank all reviewers for their constructive feedback that helped to improve our work. This work was supported by the Max Planck Society, the Volkswagen Foundation, and funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy–EXC2064/1–390727645.15/18.

Impact statement

Large language models (LLMs) excel in a wide range of applications due to their in-context learning abilities. Our research explores in-context learning and therefore contributes to our understanding of these models.

We have demonstrated certain behavioral patterns emerge depending on how a task is framed. This knowledge can be used for good and bad. On the good side, we now better understand how LLMs behave in certain situations. That means we can take precautionary measures if we see divergences from desired behaviors. However, bad actors may also exploit the fact that LLMs have an optimism bias, for example in the context of disinformation campaigns, risk assessments, or user recommendations. There may be other potential societal consequences of our work that we are not aware of.

References

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., and Rushworth, M. F. S. Learning the value of information in an uncertain world. *Nature Neuroscience*, 10(9):1214–1221, September 2007. ISSN 1546-1726. doi: 10.1038/nn1954.

Bigelow, E. J., Lubana, E. S., Dick, R. P., Tanaka, H., and Ullman, T. D. In-context learning dynamics with random binary sequences. *arXiv preprint arXiv:2310.17639*, 2023.

Binz, M. and Schulz, E. Modeling human exploration through resource-rational reinforcement learning. *Advances in Neural Information Processing Systems*, 35: 31755–31768, 2022.

Binz, M. and Schulz, E. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023.

Binz, M., Alaniz, S., Roskies, A., Aczel, B., Bergstrom, C. T., Allen, C., Schad, D., Wulff, D., West, J. D., Zhang, Q., et al. How should the advent of large language models affect the practice of science? *arXiv preprint arXiv:2312.03759*, 2023a.

Binz, M., Dasgupta, I., Jagadish, A., Botvinick, M., Wang, J. X., and Schulz, E. Meta-Learned Models of Cognition, April 2023b.

Bishop, C. M. Pattern recognition and machine learning. *Springer*, 2:5–43, 2006.

Botvinick, M., Ritter, S., Wang, J., Kurth-Nelson, Z., Blundell, C., and Hassabis, D. Reinforcement Learning, Fast and Slow. *Trends in Cognitive Sciences*, 23, April 2019. doi: 10.1016/j.tics.2019.02.006.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Cazé, R. D. and van der Meer, M. A. Adaptive properties of differential learning rates for positive and negative outcomes. *Biological cybernetics*, 107(6):711–719, 2013.

Chambon, V., Théro, H., Vidal, M., Vandendriessche, H., Haggard, P., and Palminteri, S. Information about action outcomes differentially affects learning from self-determined versus imposed choices. *Nature Human Behaviour*, 4(10):1067–1079, October 2020. ISSN 2397-3374. doi: 10.1038/s41562-020-0919-5.

Chan, S., Santoro, A., Lampinen, A., Wang, J., Singh, A., Richemond, P., McClelland, J., and Hill, F. Data distributional properties drive emergent in-context learning in transformers. *Advances in Neural Information Processing Systems*, 35:18878–18891, 2022.

Coda-Forno, J., Binz, M., Akata, Z., Botvinick, M., Wang, J. X., and Schulz, E. Meta-in-context learning in large language models. *arXiv preprint arXiv:2305.12907*, 2023.

Dasgupta, I., Lampinen, A. K., Chan, S. C., Creswell, A., Kumaran, D., McClelland, J. L., and Hill, F. Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*, 2022.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., and Dolan, R. J. Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6):1204–1215, 2011.

Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J., and Sui, Z. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

Duan, Y., Schulman, J., Chen, X., Bartlett, P. L., Sutskever, I., and Abbeel, P. Rl $\hat{2}$: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016.

Eloundou, T., Manning, S., Mishkin, P., and Rock, D. Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*, 2023.

Fechner, G. T. *Elemente der Psychophysik*, volume 2. Breitkopf u. Härtel, 1860.

Gagne, C., Zika, O., Dayan, P., and Bishop, S. J. Impaired adaptation of learning to contingency volatility in internalizing psychopathology. *eLife*, 9:e61387, December 2020. ISSN 2050-084X. doi: 10.7554/eLife.61387.

Garg, S., Tsipras, D., Liang, P. S., and Valiant, G. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022.

Gershman, S. J. A unifying probabilistic view of associative learning. *PLoS computational biology*, 11(11):e1004567, 2015.

Gershman, S. J. Deconstructing the human algorithms for exploration. *Cognition*, 173:34–42, 2018.

Globig, L. K., Witte, K., Feng, G., and Sharot, T. Under threat, weaker evidence is required to reach undesirable conclusions. *Journal of Neuroscience*, 41(30):6502–6510, 2021.

Jacobs, R. A. and Kruschke, J. K. Bayesian learning theory applied to human cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(1):8–21, 2011.

Jagadish, A. K., Binz, M., Saanum, T., Wang, J. X., and Schulz, E. Zero-shot compositional reinforcement learning in humans, July 2023.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103: 102274, 2023.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kocmi, T. and Federmann, C. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*, 2023.

Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., and Palminteri, S. Behavioural and neural characterization of optimistic reinforcement learning. *Nature Human Behaviour*, 1(4):0067, March 2017. ISSN 2397-3374. doi: 10.1038/s41562-017-0067.

Lefebvre, G., Summerfield, C., and Bogacz, R. A Normative Account of Confirmation Bias During Reinforcement Learning. *Neural Computation*, 34(2):307–337, January 2022. ISSN 0899-7667, 1530-888X. doi: 10.1162/neco_a_01455.

Mikulik, V., Delétang, G., McGrath, T., Genewein, T., Martic, M., Legg, S., and Ortega, P. Meta-trained agents implement bayes-optimal agents. *Advances in neural information processing systems*, 33:18691–18703, 2020.

Miller, G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., and Zettlemoyer, L. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. Asynchronous Methods for Deep Reinforcement Learning. In *Proceedings of The 33rd International Conference on Machine Learning*, pp. 1928–1937. PMLR, June 2016.

Nassar, M. R., Wilson, R. C., Heasly, B., and Gold, J. I. An approximately bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *Journal of Neuroscience*, 30(37):12366–12378, 2010.

Nickerson, R. S. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220, 1998.

Ortega, P. A., Wang, J. X., Rowland, M., Genewein, T., Kurth-Nelson, Z., Pascanu, R., Heess, N., Veness, J., Pritzel, A., Sprechmann, P., et al. Meta-learning of sequential strategies. *arXiv preprint arXiv:1905.03030*, 2019.

Palminteri, S. and Lebreton, M. The computational roots of positivity and confirmation biases in reinforcement learning. *Trends in Cognitive Sciences*, 26(7):607–621, July 2022. ISSN 13646613. doi: 10.1016/j.tics.2022.04.005.

Palminteri, S., Lefebvre, G., Kilford, E. J., and Blakemore, S.-J. Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLOS Computational Biology*, 13(8):e1005684, August 2017. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005684.

Pedersen, M. L., Frank, M. J., and Biele, G. The drift diffusion model as the choice rule in reinforcement learning. *Psychonomic bulletin & review*, 24:1234–1251, 2017.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Rescorla, R. and Wagner, A. A theory of Pavlovian conditioning: The effectiveness of reinforcement and nonreinforcement. *Classical Conditioning: Current Research and Theory*, January 1972.

Roscher, R., Bohn, B., Duarte, M. F., and Garcke, J. Explainable machine learning for scientific insights and discoveries. *Ieee Access*, 8:42200–42216, 2020.

Roziere, B., Gehring, J., Gloeckle, F., Sootla, S., Gat, I., Tan, X. E., Adi, Y., Liu, J., Remez, T., Rapin, J., et al. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023.

Schulz, E., Bhui, R., Love, B. C., Brier, B., Todd, M. T., and Gershman, S. J. Structured, uncertainty-driven exploration in real-world consumer choice. *Proceedings of the National Academy of Sciences*, 116(28):13903–13908, July 2019. doi: 10.1073/pnas.1821028116.

Schwarz, G. Estimating the dimension of a model. *The annals of statistics*, pp. 461–464, 1978.

Sharot, T. The optimism bias. *Current Biology*, 21(23): R941–R945, December 2011. ISSN 0960-9822. doi: 10.1016/j.cub.2011.10.030.

Shepperd, J. A., Klein, W. M. P., Waters, E. A., and Weinstein, N. D. Taking Stock of Unrealistic Optimism. *Perspectives on Psychological Science*, 8(4): 395–411, July 2013. ISSN 1745-6916. doi: 10.1177/1745691613485247.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., and Vladymyrov, M. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pp. 35151–35174. PMLR, 2023.

Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., and Anandkumar, A. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., Blundell, C., Kumaran, D., and Botvinick, M. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*, 2016.

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., Blundell, C., Kumaran, D., and Botvinick, M. Learning to reinforcement learn, January 2017.

Wei, J., Wei, J., Tay, Y., Tran, D., Webson, A., Lu, Y., Chen, X., Liu, H., Huang, D., Zhou, D., et al. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023.

Williams, R. and Peng, J. Function Optimization Using Connectionist Reinforcement Learning Algorithms. *Connection Science*, 3:241, September 1991. doi: 10.1080/09540099108946587.

Wilson, R. C. and Collins, A. G. Ten simple rules for the computational modeling of behavioral data. *eLife*, 8:e49547, November 2019. ISSN 2050-084X. doi: 10.7554/eLife.49547.

Wu, H., Fai Cheung, S., and On Leung, S. Simple use of bic to assess model selection uncertainty: An illustration using mediation and moderation models. *Multivariate behavioral research*, 55(1):1–16, 2020.

Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021.

Zhang, S. and Yu, A. J. Forgetful bayes and myopic planning: Human learning and decision-making in a bandit setting. *Advances in neural information processing systems*, 26, 2013.

Zhu, C., Byrd, R. H., Lu, P., and Nocedal, J. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on mathematical software (TOMS)*, 23(4):550–560, 1997.

A. Methods: Additional details

A.1. Prompt engineering

We point out the two design decisions that were crucial to achieving an above-chance-level performance on all tasks: (1) We framed this task in a gambling context as this allowed us to explicitly mention the rewards the agent can expect and that they vary stochastically with different probabilities. (2) We randomly sample the letters for each slot machine for each task. We excluded certain letters (I, U, X, Y, Z) as we noticed that the LLM was positively (I, U) or negatively (X, Y, Z) biased towards choosing them.

A.2. Parameter estimation

The model parameters were fitted using a maximum a posteriori approach. For the RW model, this involves fitting parameters $\theta = [\alpha, \beta]$ (two parameters in total). For the RW± model, this involves fitting parameters $\theta = [\alpha^+, \alpha^-, \beta]$ (three parameters in total). The model parameters are fitted separately for each simulated run using the following objective:

$$\hat{\theta} = \arg\min_{\theta} \left[-\ln p(\theta) - \sum_{n=1}^{N} \ln p(a_n|\theta) \right]$$

Prior probabilities were based on (Daw et al., 2011), using a beta distribution $\mathcal{B}(1.1, 1.1)$ for learning rates and a gamma distribution with $\mathcal{G}(1.2, 5)$ for the inverse temperature parameter. We used a bound-constrained minimization that was implemented using scipy's `minimize` function, which internally relies on the L-BFGS-B algorithm (Zhu et al., 1997).

Parameter values were initialized in the range of 0 to 1 for α^+ and α^- and 0 to 10 for β. The fitting procedure was repeated 100 times and the time it took varied depending on the number of free parameters of the cognitive model. For one simulated run, the procedure took between ten seconds and three minutes on a standard desktop computer.

B. Generalizability across different LLMs and task formats

B.1. Further LLMs

We performed the partial feedback task on seven additional LLMs. Namely, we tested Claude-2.1, Claude-3 Haiku, GPT-4, and four versions of Llama-2, i.e. the 7 billion and the 70 billion parameter models with and without RLHF (Llama-2-7B, Llama-2-7B-Chat, Llama-2-70B, and Llama-2-70B-Chat). We analyzed the learning behavior of these models and found that the observed learning asymmetry is not unique to Claude-1.2 but is also significant in all tested LLMs (see Figure 6).

We also tested the influence of RLHF and model size on the emergence of the optimism bias. We used an ordinary least squares linear regression model for the Llama-2 model family, since the exact sizes for GPT-4 and the Claude models are unknown. We defined the optimism bias using the difference in learning rates (i.e., $\alpha^+ - \alpha^-$) as our dependent variable. We included RLHF, coded as 1 for presence and 0 for absence, model size (7 or 70), and a constant bias term as independent variables in the regression model. Our results show that RLHF tends to increase the optimism bias (with a coefficient of 0.2108, $p < 0.001$), while an increase in model size tends to decrease this effect (with a coefficient of -0.0043, $p < 0.001$). However, it is important to note that the scope of this analysis was limited to the Llama-2 models, which limits our ability to generalize these results to other models, such as GPT-4 or Claude.

B.2. Further task formats

We tested Claude-1.2 on variations of the partial feedback task to check the robustness of our results. We manipulated two aspects: For one, we added one or two additional slot machines to each casino, leading to three- and four-armed bandit problems. Furthermore, we used two different reward magnitudes for success and failure (0.5 and -0.5 and 1.0 and 0.0). This resulted in six new task formats. We found that the optimism bias remains persistent across these variations (see Figure7). This demonstrates that the underlying optimistic learning dynamics generalize to settings with different reward magnitudes and more than two options.

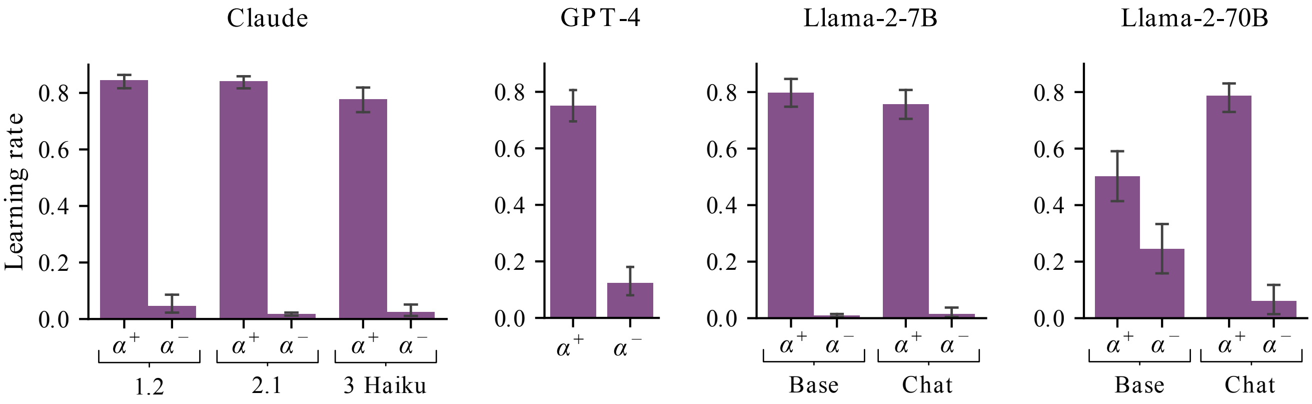

Figure 6. Learning rate comparison of eight LLMs in the partial feedback task. All analyzed LLMs show a significant learning asymmetry with a stronger response to positive prediction errors than to negative prediction errors, indicating an optimism bias. Plots are divided by model family. Error bars correspond to 95% CIs.

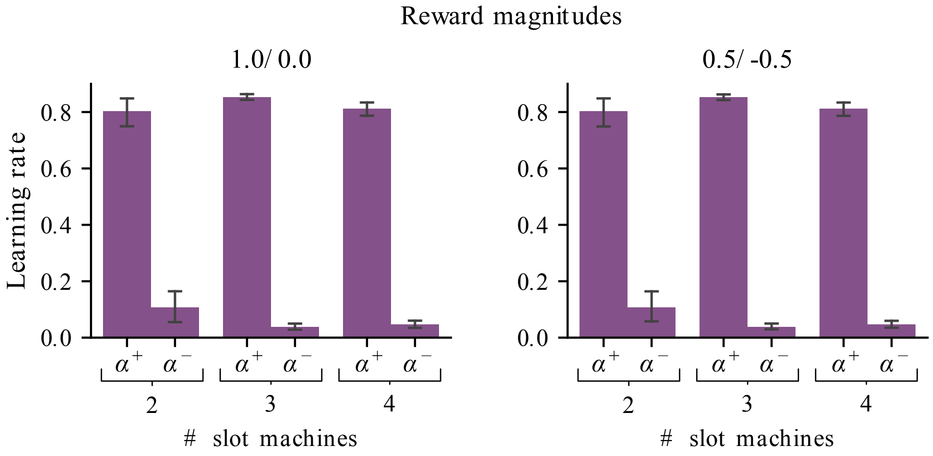

Figure 7. LLM learning rate analysis for partial feedback task variations. Slot machines returned two different pairs of rewards for success and failure of either 1.0 and 0.0 (left) or 0.5 and -0.5 (right). In addition, the casinos contained a varying number of slot machines (2, 3, and 4). The LLM shows an optimism bias in all settings tested. Error bars correspond to 95% CIs.

C. Task 2: Additional details

C.1. Prompts

The full information task consisted of 16 casinos, half of which were partial feedback casinos and half of which were full feedback casinos. Below is a sample prompt for both types of casinos. The differences between the partial and full feedback casinos are shown in bold.

Prompt Task 2: partial feedback casinos

You will visit a casino 40 times. The casino has two slot machines that stochastically return either 1 or -1 with different reward probabilities. You can only interact with one slot machine per visit. Half of the time you visit the casino, you can play, the other someone else is playing. During visits where you can play, you'll earn points from the chosen machine. During visits where someone else is playing, you'll learn what points are earned on the chosen machine. Your goal is to maximize the total amount of points you receive in all 20 visits you can play.

During your previous visits you have observed the following:

- On visit 1 someone else played Machine H and earned 1.0 point.
- On visit 2 you played Machine H and earned 1.0 point.
- On visit 3 you played Machine E and earned -1.0 point.

Q: You are now in visit 4. Which machine do you choose between Machine E and Machine H?
A: Machine [insert]

Prompt Task 2: full feedback casinos

You will visit a casino 40 times. The casino has two slot machines that stochastically return either 1 or -1 with different reward probabilities. You can only interact with one slot machine per visit. Half of the time you visit the casino, you can play, the other someone else is playing. During visits where you, can play, you'll earn points from the chosen machine. **You'll also learn what points would have been earned had the other machine been selected.** During visits where someone else is playing, you'll learn what points are earned on the chosen **and what points would have been earned had the other machine been selected. Nevertheless, you only accumulate points from the machine you choose to play.** Your goal is to maximize the total amount of points you receive in all 20 visits you can play.

During your previous visits you have observed the following:

- On visit 1 someone else played Machine H and earned 1.0 point.
  On Machine E the player would have earned -1.0 point.
- On visit 2 you played Machine H and earned 1.0 point.
  On Machine E you would have earned -1.0 point.
- On visit 3 you played Machine E and earned -1.0 point.
  On Machine H you would have earned 1.0 point.

Q: You are now in visit 4. Which machine do you choose between Machine E and Machine H?
A: Machine [insert]

C.2. Model comparison

We simplified the analysis of the original by fitting separate learning rates only for the chosen and unchosen option, but not separate learning rates for free and forced-choice trials. Furthermore, we used only the simulated behavior from full feedback casinos for the fitting of two cognitive models – a 2α and a 4α model.

The model comparison revealed that the 2α model provided a better fit to the behavior. The 2α model contained only two learning rates – combining the learning rates for the chosen and unchosen options, which either confirmed or disconfirmed prior beliefs. As can be seen in Figure 8, all agents show a clear tendency to overweight information that confirms their choices.

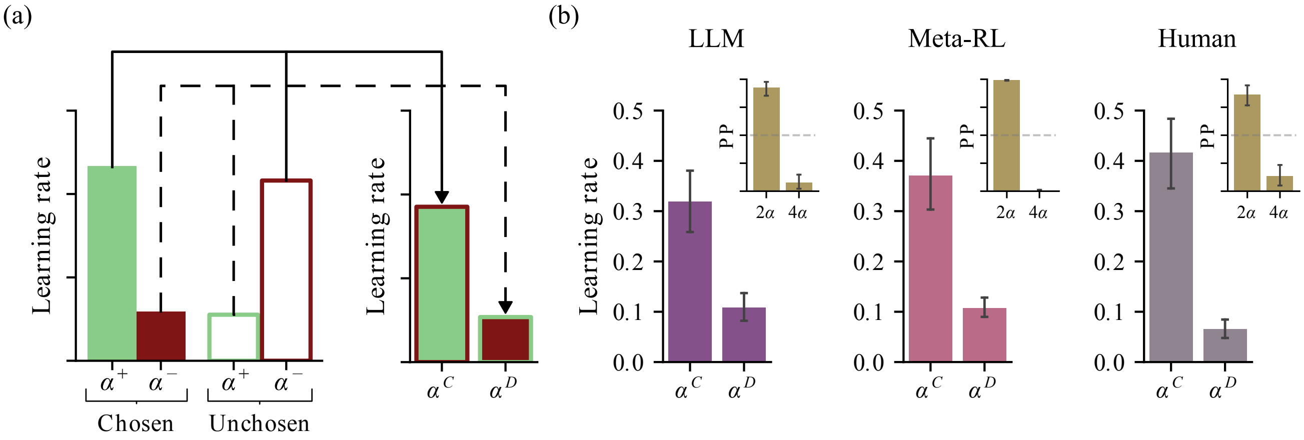

Figure 8. Confirmation bias in the full feedback task. (a) Schematic showing how the learning rates of the full model (i.e. a model for a different learning rate for each possible combination of outcome types and prediction error types) relate to those of the confirmation bias model (2α), which bundles together the learning rates for positive chosen and negative unchosen (i.e. confirmatory) prediction errors (α^C) and the learning rates for negative chosen and positive unchosen (i.e. disconfirmatory) prediction errors (α^D). Adapted from (Palminteri & Lebreton, 2022) (b) Fitted confirmation bias model for the LLM, the Meta-RL agent and human participants. The average posterior probabilities (PP) indicate that the 2α model is a superior fit in all model comparisons. Error bars correspond to 95% CIs.

D. Task 3: Additional details

D.1. Prompts

The agency condition consisted of 12 casinos and provided only partial feedback. Half of these casinos were free-choice casinos containing only free-choice trials, and the other half were a mixture of free-choice and forced-choice trials. The free-choice casinos consisted of 40 trials per casino. The mixed-choice casinos consisted of 80 trials with 40 free-choice trials and 40 forced-choice trials. Below is a sample prompt for both types of casinos. The differences between the free-choice and mixed-choice casinos are shown in bold.

Prompt Task 3: free-choice casinos

You will visit a casino 40 times. The casino has two slot machines that stochastically return either 1 or -1 with different reward probabilities. You can only interact with one slot machine per visit. Your goal is to maximize the total amount of points you receive in all 40 visits you can play.

During your previous visits you have observed the following:

- On visit 1 you played Machine H and earned 1.0 point.
- On visit 2 you played Machine N and earned -1.0 point.
- On visit 3 you played Machine H and earned -1.0 point.

Q: You are now in visit 4. Which machine do you choose between Machine N and Machine H?
A: Machine [insert]

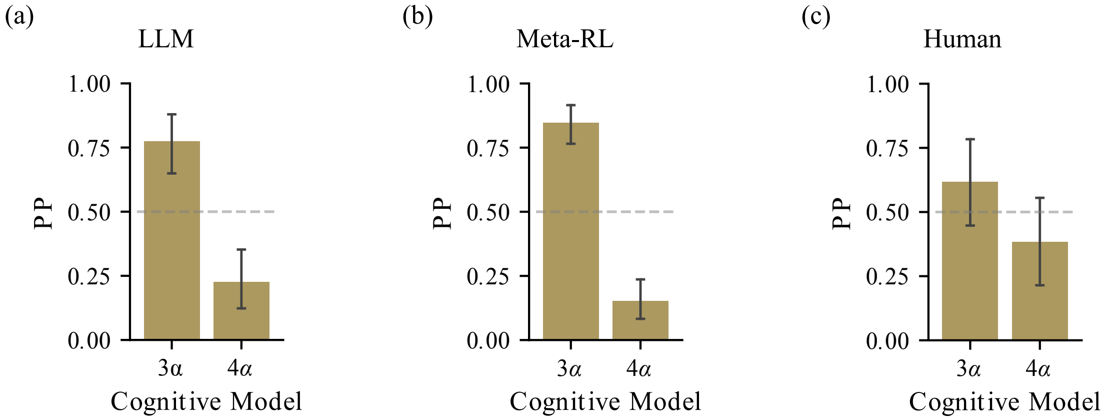

Figure 9. Model comparison for the agency condition for all agents. The 3α model has the highest average posterior probabilities (PP). This indicates that information from forced-choice trials where no agency is involved are weighted symmetrically whereas the free-choices are weighted asymmetrically. Error bars correspond to 95% CIs.

Prompt Task 3: mixed-choice casinos

You will visit a casino 80 times. The casino has two slot machines that stochastically return either 1 or -1 with different reward probabilities. You can only interact with one slot machine per visit. **Half of the time you visit the casino, you can play, the other half someone else is playing and you can only see the rewards for their chosen slot machine.** Your goal is to maximize the total amount of points you receive in all 40 visits you can play.

During your previous visits you have observed the following:

- On visit 1 you played Machine H and earned 1.0 point.
- On visit 2 someone else played Machine N and received -1.0 point.
- On visit 3 you played Machine H and earned -1.0 point.

Q: You are now in visit 4. Which machine do you choose between Machine N and Machine H?
A: Machine [insert]

D.2. Model comparison

We only fit two cognitive models to the mixed-choice casinos. One model consisted of separate learning rates for positive and negative prediction errors for free- and forced-choice trials (4α). The second model consisted of two learning rates for free-choice trials and only one learning rate for forced-choice trials (3α). Model parameters were fit based on the free-choice trials of the mixed-choice casinos. Model comparison indicated that for all agents, the 3α model provided a better fit to the data based on the average PP (see Figure 9).

E. Meta-RL: Additional details

E.1. Agent

The agent consisted of a Transformer network with a model dimension of $8 \cdot \text{input\_size}$, two feedforward layers with a dimension of 128, and eight attention heads, followed by two linear layers that output a policy and a value estimate, respectively. The agent received the previously selected action (one-hot encoded) and the reward of that action, alongside the current context. This context included a normalized time index for all tasks. For Task 1, a one-hot encoded representation of the four casinos was added, resulting in an input size of eight. The context of Tasks 2 and 3 included a bit representation of the trial type (i.e. $00 =$ free-choice, $10 =$ forced-choice left option chosen, $01 =$ forced-choice right option chosen) for the current and previous trial. Task 2 additionally included the reward of the current unchosen option, resulting in an input size of eight for Task 2 and nine for Task 3. In the partial feedback casinos of Task 2, a placeholder of 0 for the missing reward signal of the unchosen option was propagated. To prevent learning from forced-choice trials, we masked the policy and

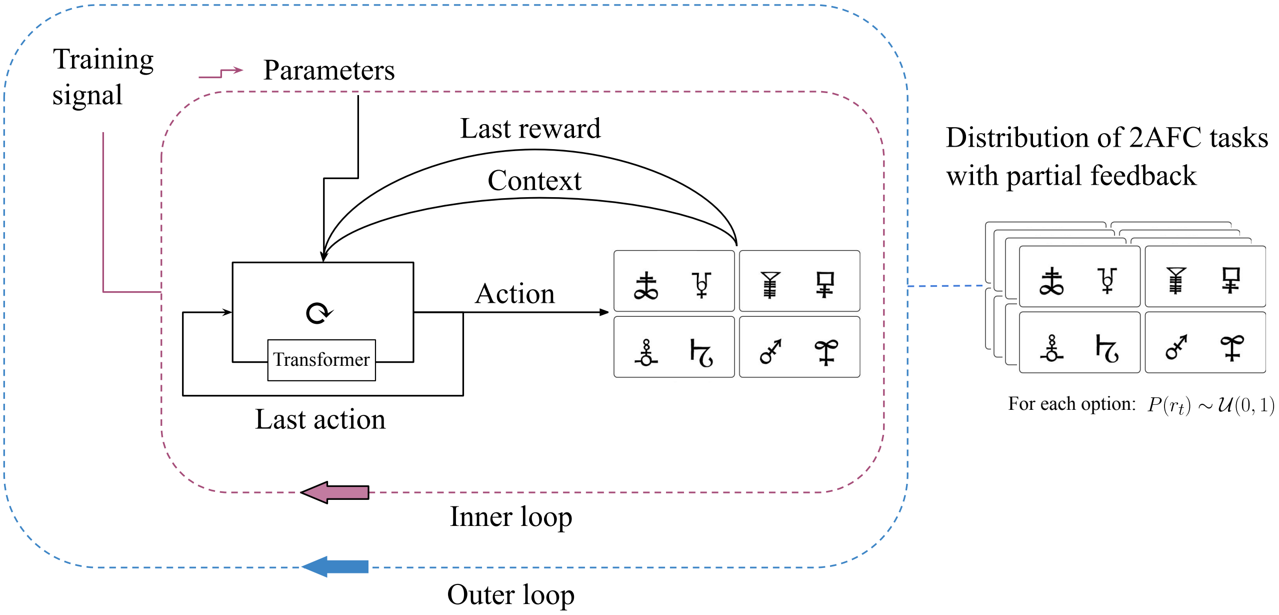
Training signal
Parameters
Last reward
Context
Action
Transformer
Last action
Inner loop
Outer loop
Distribution of 2AFC tasks with partial feedback
For each option: $P(r_t) \sim \mathcal{U}(0,1)$

Figure 10. Schematic of the Meta-RL training for Task 1 highlighting the inner and outer training loops. The outer loop adjusts the weights of the Transformer in response to the learning experience. These weights shape the behavior of the Transformer in the inner loop as it interacts with 2AFC tasks (here Task 1). After each cycle of the outer loop, a new task is sampled where the probabilities for each option are sampled from a uniform distribution. Adapted from Botvinick et al. (2019).

value loss for those trials. To ensure a consistent input dimension, we use placeholder values for the initial inputs and all the unseen trials.

Network weights were adjusted by gradient-based optimization using ADAM (Kingma & Ba, 2014) on a standard actor-critic loss (Mnih et al., 2016) at the end of each training episode. The initial learning rate for ADAM was 0.0003. For the actor-critic loss, we used a discount factor of 0.8 and weighted the critic loss with 0.5. In addition, we used entropy regularization to discourage premature convergence to sub-optimal policies (Williams & Peng, 1991). Starting with an entropy coefficient of 1, we linearly decayed the influence of the entropy term to 0 after half of the $5,000$ episodes. We used a batch size of 64 during training.

E.2. Training

The training process of the Meta-RL agent is graphically depicted in Figure 10. In the process, there are two optimization loops – an outer and an inner loop. At the beginning of each training episode (outer loop), we sample a new task b_i from the prior distribution of 2AFC tasks. As stated earlier, we assumed a uniform distribution for the win probabilities for each option in each task. The goal of the agent is to find a history-dependent policy $\pi_{\mathbf{\Theta}}$ that maximizes the expected total discounted reward accumulated during an episode. For that, the parameters $\mathbf{\Theta}$ are adjusted at the end of each episode. Since the decision strategy changes across training episodes, the agent must act differently according to its prior belief over which part of the task distribution it is currently in. As the optimization maximizes the expected total rewards across tasks, the policy starts to generalize the underlying principles that help reach this objective.

The agent interacts in the inner loop with the specific sampled task b_i, aiming to maximize its rewards across all steps with the help of its policy. At the beginning of each step, a context c_t is drawn from a uniform distribution. Upon receiving an action a_t, the environment computes a reward r_t and samples the next context c_{t+1} to which the agent steps forward. The next context c_{t+1}, the action a_t, and the reward r_t are concatenated and added to the input to the Transformer. As demonstrated by Wang et al. (2017), this input design is crucial for an agent to learn an association between choices that have been made in particular states and their subsequent rewards.

After training to convergence, we tested the Meta-RL agent on the three experimental task under the same conditions as with the LLM. This idealized in-context learning agent outperformed the human as well as the LLM (see Figure 11).

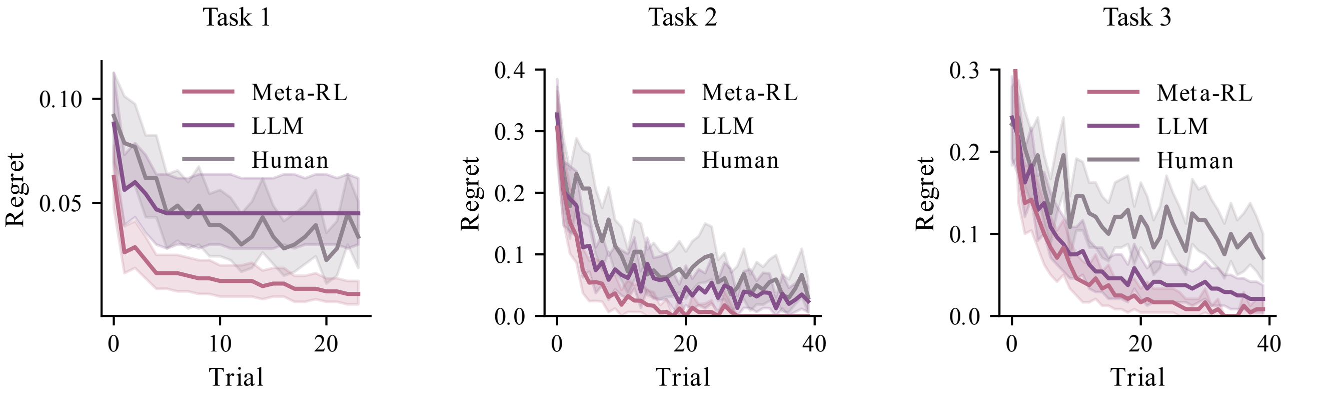

Figure 11. Average performance comparison of all three agents in terms of regret. In all tasks, the Meta-RL agent improves its performance over time fastest. Shaded areas correspond to 95% CIs.

F. Data availability

Our code is publicly available at `https://github.com/jschbrt/InContext-Learning-Dynamics`.