Reflection-Bench: probing AI intelligence with reflection

Anonymous ACL submission

Abstract

The ability to adapt beliefs or behaviors in response to unexpected outcomes, *reflection*, is fundamental to intelligent systems' interaction with the world. From a cognitive science perspective, this serves as a core principle of intelligence applicable to both human and AI systems. To address the debate on the intelligence of large language models (LLMs), we propose *Reflection-Bench*, a comprehensive benchmark comprising 7 tasks spanning core cognitive functions crucial for reflection, including perception, memory, belief updating, decision-making, prediction, counterfactual thinking, and meta-reflection. We evaluate the performances of 13 prominent LLMs such as OpenAI o1, GPT-4, Claude 3.5 Sonnet, etc. The results indicate that current LLMs still lack satisfactory reflection ability. We discuss the underlying causes of these results and suggest potential avenues for future research. In conclusion, Reflection-Bench offers both evaluation tools and inspiration for developing AI capable of reliably interacting with the environment.

1 Introduction

Large language models (LLMs) have sparked intense debate regarding their true capabilities: *are they genuinely intelligent or simply sophisticated statistical engines mimicking human language?* (Bender et al., 2021) This question has profound implications for trust in AI and the development of appropriate regulations. Proponents of AI's human-level intelligence often advocate for stricter regulations due to potential existential risks, while skeptics argue that excessive regulation could hinder innovation. Numerous studies probe specific facets of LLM intelligence, such as reasoning (Clark et al., 2018), planning (Valmeekam et al., 2024), and cognitive flexibility (Kennedy and Nowak, 2024), etc. However, these investigations often lack a unifying framework grounded in a fundamental theory of intelligence, hindering a comprehensive understanding of LLM capabilities. Our work aims to clarify this enigma and establish efficient metrics for assessing intelligence by drawing insights from cognitive science into the fundamental process of intelligence, *reflection*.

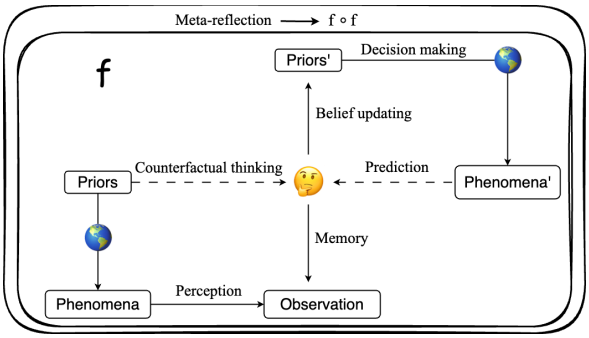

Figure 1: Reflection, a fundamental process of intelligence, integrates various cognitive components. To achieve desired outcomes, an intelligent agent must predict the external world states and behavioral consequences based on prior beliefs. Post-action, discrepancies between prediction and observation are perceived, prompting an update of prior belief. This update involves recalling the previous decision process and engaging in counterfactual thinking about un-chosen alternatives.

From a first-principles perspective, cognitive science conceptualizes intelligent systems as predictive machines that constantly anticipate future events using internal models (Friston, 2010). They can cope with uncertain environments by minimizing surprises (mismatch between predictions and observations) through iterative updates to thoughts and actions, as shown in Figure 1. We operationalize this process as reflection, a cyclic process of predicting based on prior belief, making decisions to achieve desired states, perceiving surprises, and subsequently updating beliefs. Reflection encompasses crucial cognitive components including perception, memory, belief updating, decision-making, prediction, counterfactual thinking, and even meta-

reflection. Leveraging this framework, we introduce *Reflection-Bench*, a benchmark designed to evaluate the reflection capabilities of LLMs.

Based upon well-established cognitive science paradigms, we select 6 paradigms and design 7 tasks adapted for LLMs evaluation, including oddball paradigm, n-back task, probabilistic reversal learning task, Wisconsin card sorting test, weather prediction task, double-choice Iowa gambling task, and meta-bandit task. Reflection-Bench leverages these 7 tasks to offer a comprehensive evaluation of different aspects while engaging in reflection. Importantly, the difficulty of these tasks can be adjusted to accommodate varying cognitive loads, ensuring Reflection-Bench's adaptability to more advanced AI models and maintaining its long-term relevance.

We evaluated 13 diverse LLMs on Reflection-Bench using relatively easy settings and conducted an in-depth analysis of their performance. The results demonstrate Reflection-Bench's discriminative power: o1-preview achieved the highest scores, followed by the other state-of-the-art LLMs, while models with smaller sizes scored lower. We highlight a notable contrast between o1-preview's performances in the oddball paradigm and its performance in other tasks. Additionally, we observed a universal absence of meta-reflection across all models. The potential causes and implications are discussed. In conclusion, our findings indicate that current LLMs still fall considerably short of human-level intelligence.

To summarize, our main contributions are as follows:

- **We introduce reflection, a core feature of intelligent systems, into the AI field** as a biologically-inspired criterion for probing AI intelligence. This approach offers a more nuanced understanding of LLMs' intelligence that aligns with humans.

- **We propose Reflection-Bench**, a comprehensive benchmark comprising seven tasks adapted from cognitive science paradigms to assess the reflection capabilities of LLMs. This benchmark decomposes reflection into crucial cognitive components: **perception, memory, belief updating, decision-making, prediction, counterfactual thinking, and meta-reflection**.

- We conduct an assessment of 13 prominent LLMs. The study reveals that **current LLMs exhibit significant limitations in their capacity for human-like reflection**, especially the universal lack of meta-reflection abilities.

2 Related work

Emergent ability 'Emergent abilities' in LLMs are defined as capabilities 'not present in smaller models but present in larger models'(Wei et al., 2022). These manifest as the ability to 'perform above the random baseline on corresponding tasks without explicit training on those same tasks', such as reasoning and planning. (Lu et al., 2024). Some researchers argue that these emergent abilities might disappear with better metrics or statistics (Schaeffer et al., 2023). This underscores the importance of designing valid and efficient metrics for evaluating intelligence, one of the most debated emergent properties of LLMs.

Evaluating intelligence Various benchmarks have been developed to assess different aspects of LLMs' intelligence. The AI2 Reasoning Challenge (ARC) evaluates reasoning, common sense, and text comprehension using multiple-choice questions (Clark et al., 2018). PlanBench tests planning and reasoning with open-ended questions (Valmeekam et al., 2024). CELLO assesses four levels of causal inference - discovery, association, intervention, and counterfactual - with multi-choice questions (Chen et al., 2024). Beyond the single-turn evaluations, researchers also investigate LLMs' intelligence in uncertain environments through multi-turn dialogues. These include tasks such as multi-armed bandits for probability estimation (Krishnamurthy et al., 2024), Wisconsin card sorting tests for rule inference (Kennedy and Nowak, 2024), Minecraft for spatial reasoning (Madge and Poesio, 2024), etc. These aspects are interconnected in ways related to the epistemology of AI systems - how they learn about and adapt to the external world. While this feature has been vaguely recognized but not clearly defined in AI field.

First principle of intelligence A first-principles perspective in cognitive science conceptualizes intelligent systems as predictive machines that continuously predict future events with internal models(Friston, 2010). They can adapt to environments flexibly by minimizing unexpected events through updating thoughts or actions. This theory has been applied to interpret intelligent systems across dif-

ferent scales, such as in vitro neural networks (Isomura et al., 2023), sensory system (Huang and Rao, 2011), cognitive process (Spratling, 2016), consciousness (Solms, 2019), and self-identity (Li and Li, 2024). Therefore, this theory serves as the first principle of intelligence. Considering the cyclic nature of this process as showcased in Figure 1 - prediction, verification, and update - we term it *reflection*. Through reflection, we can comprehensively assess various aspects of intelligence encompassing the aforementioned research interests.

3 Reflection-Bench

3.1 Task selection for evaluating reflection

As discussed, reflection is a complex capability collaborated by cognitive components including perception, working memory, belief updating, decision-making, prediction, counterfactual thinking, and meta-reflection. Drawing from well-established cognitive science paradigms, we select 6 paradigms and design 7 tasks adapted for LLMs evaluation in Reflection-Bench, as follows:

Perception For perceiving surprise signals, we select the *oddball paradigm* (Näätänen et al., 2007), a method to study the brain's automatic processing of novel stimuli. As shown in Figure 2.A, participants are presented with a sequence of auditory stimuli, including frequent standard and rare deviant stimuli (*e.g.*, different tones). Electroencephalogram recordings show that the brain is naturally sensitive to deviant stimuli, as reflected in a deviance-related negative waveform called Mismatch Negativity (MMN) (Garrido et al., 2009). This automatic phenomenon represents the basic ability to perceive surprise signals, which aligns with our intention appropriately.

Memory Recalling the previous decision process requires active memory retrieval. We select the *n-back* task for evaluating this ability (Jaeggi et al., 2010). As shown in Figure 2.B, presented with sequential stimuli (typically symbols), subjects need to indicate whether the current stimulus matches the one from n steps earlier. The process of continuously updating and maintaining information in the n-back task fits our objective well.

Belief updating Focusing on belief updating, we involve *probabilistic reversal learning task* (PRLT) (Costa et al., 2015) in our benchmark. As illustrated in Figure 2.C, participants engage in a two-arm bandit task, making choices between two options with different reward probabilities, which they need to infer to maximize rewards. At the midpoint of the task, the reward probabilities are reversed without notification. This task focuses on the ability to update beliefs in response to the changing environment.

Decision-making We employ the *Wisconsin card sorting test* (WCST) to assess the flexibility in decision-making (Nyhus and Barceló, 2009). Participants are presented with a set of cards varying in color, shape, and number of figures, and must match the given card according to an undisclosed rule (e.g., by color) which they must infer from feedback on their choices. (Figure 2.D). After a certain number of trials, the matching rule changes without warning. Both PRLT and WCST require flexible belief updating, while WCST focuses more on inferring which latent rule one should obey when making decisions.

Prediction Computationally, prediction is realized via transition probability (Friston, 2010). There are no tests directly evaluating transition probabilistic learning in cognitive science. Therefore we designed an adapted version of the *weather prediction task* (WPT) originally designed to study probabilistic classification learning (Shohamy et al., 2008). In the original version, subjects need to predict the weather based on card cues and gradually learn the probability relationship between card cues and weather. We convert this probability into the explicitly defined transition probability matrix as shown in Figure 2.E. The evaluated model must learn how the cues (sensors) influence the transition of weather, through which their flexible prediction ability is assessed.

Counterfactual thinking Counterfactual thinking refers to creating an alternative to reality by considering 'what if' (Byrne, 2016). Therefore we test this ability by forcing models to consider the un-chosen options with an opportunity to 'turn back time'. We select *Iowa gambling task* (IGT) for our design (Buelow and Suhr, 2009). In IGT, participants are presented with four decks of cards (A, B, C, and D) with distinct expected rewards and losses, shown in Figure 2.F. They must maximize their profit by selecting cards from these decks over a series of trials. We implement a double-choice IGT (DC-IGT) where after each initial choice, models are given the possible outcomes of the choice

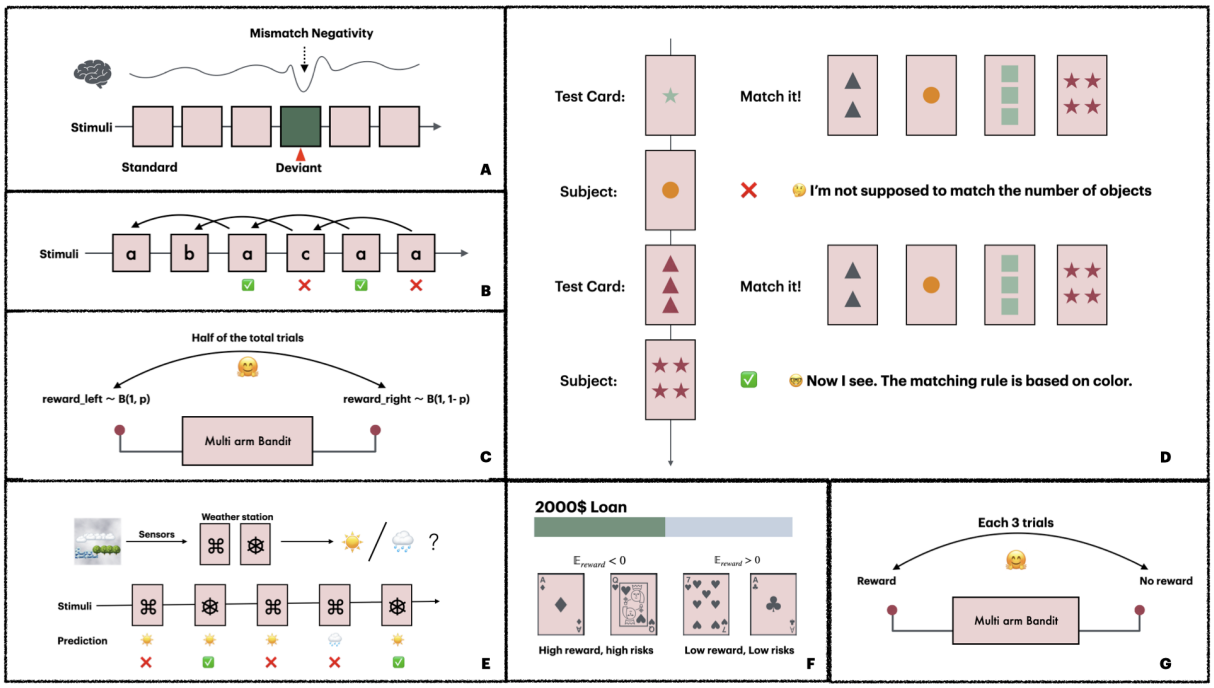

Figure 2: Tasks in Reflection-Bench. A: Oddball paradigm. B: N-back (2-back). C: Probabilistic reversal learning task. D: Wisconsin card sorting test. E: Weather prediction task. F: Iowa gambling test. G: Meta-bandit task.

and the opportunity to go back in time to make their choice again.

Meta-reflection As illustrated in Figure 1, meta-reflection is the process of reflecting the previous reflection, therefore, we design a *meta-bandit task* (MBT) based on the PRLT. In MBT, the the reward probabilities are reversed every **n** trials, creating a predictable pattern of rule changes (Figure 2.G). This design introduces a "rule of rule changing", requiring participants to not only adapt to reversals but also to recognize and anticipate the meta-structure of these changes, i.e., meta-reflection.

Through these 7 tasks, Reflection-Bench offers a comprehensive evaluation of different aspects while engaging reflection. Notably, there are no strict correspondences between these tasks and specific cognitive components, only with certain aspects being more prominently featured or assessed.

3.2 Task design for assessing LLMs

The selected six paradigms are widely used in cognitive science for investigating human intelligence. We apply them to evaluate LLMs in Reflection-Bench through the following adaption:

Oddball paradigm We design a prompt-based oddball task for evaluating LLMs. The prompt consists of seven short sentences about one topic (stimulus A) and one random sentence (stimulus B) causing content consistency interruption (example in Appendix A). The prompts are presented to LLMs with only the instruction to 'just make some brief comments', allowing us to evaluate their capability to automatically detect surprise signals. We compile 50 such prompts with the help of the o1-preview model. Model's responses are manually scored from 0 to 3 (example in Appendix A), corresponding to:

- 0: forced explanation or neglect to B
- 1: simple enumeration of A and B
- 2: pointing out that A and B are different
- 3: clearly stating B is nonsensical in this sequence

N-back For n-back implementation, we set a fixed sequence consisting of several letters (e.g. E, F, G, and H) and send them to evaluated models one by one. The models are instructed to determine if the current letter is the same as the one that appears **n** steps earlier. The model's accuracy is calculated as its score.

Probabilistic reversal learning task We design a 40-trial PRLT. In each trial, the LLM is presented with two options and asked to make a choice. The reward for the chosen option is sampled from a

Bernoulli distribution. The LLM is then prompted to make its next choice based on the reward feedback. The reward probabilities of the two options (left and right arms) are **p** and **1-p** initially and reversed at the 21st trial. The models' beliefs on the reward probability are estimated according to their decisions over time by a moving average (window size = 3). The overall score is calculated by the mean absolute errors (MAE) between estimated and true probability:

$$Score = (1 - \text{MAE}/\text{Max}_{\text{MAE}}) * 100$$

Wisconsin card sorting test Similar to the designation of (Kennedy and Nowak, 2024), we implement WCST in a text-based version consisting of **x** trials, and the matching rule changes every **x/6** trials, which means each rule can be applied 2 times. In each trial, the LLM is presented with a description of one testing card (*e.g.*, 'triangle green 4'). The model is instructed to match the target card among four choices without being told the matching rule. After each choice, the model receives feedback on whether the match was correct. We evaluated the LLMs' performance based on their overall accuracy.

Weather prediction task In each trial, LLM is presented with the current day's weather and the sensor state ([0,1] or [1,0]), and required to predict the next day's weather based on the two information. Actual weather is calculated with the corresponding transition matrix:

$$T_{sensors=[1,0]=} \begin{bmatrix} p & 1-p \\ 1-p & p \end{bmatrix}$$

$$T_{sensors=[0,1]=} \begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix}$$

At the next trial, LLM is informed of the actual weather which is also the current day's weather for this trial and the state of sensors, and required to make a new prediction again. We estimate the transition probability matrices of models according to their last 20 predictions in 100 trials. Models' performances are evaluated in the same way as PRLT, *i.e.*, based on MAE between estimated and true transition matrices.

Double choice Iowa gambling task We implement DC-IGT for 100 trials. Gains of four decks are \$100, \$100, \$50, and \$50, and their losses are \$260, \$1250, \$50, and \$200 with possibilities of $\mathbf{P_{loss}} = p_a, p_b, p_c, p_d$, respectively. In each trial, the LLM is presented with four card decks and asked to make an initial choice. The gain and loss of its choice are calculated with the above rule. After receiving feedback on the gain and loss, the model is prompted to make a second choice, with the option to stick with or change their initial decision. The LLM's performances are the composite of short-term (beneficial switches that avoid losses) and long-term (final overages) metrics. Notably, the scores in DC-IGT are normalized from the raw scores, so the marks only represent performances relative to other models.

Task	Trials	Parameters
Oddball	50	NA
N-back	52	**n**=2
PRLT	40	**p**=0.9
WCST	108	**x**=108
WPT	100	**p**=0.9
DC-IGT	100	$\mathbf{P_{loss}} = 0.5, 0.1, 0.5, 0.1$
MBT	60	**n**=3, **p**=1

Table 1: Experiment settings of our experiment including trials and parameter settings in each task.

Meta-bandit task The MBT consists of 20 blocks of **n** trials each. In each trial, the LLM is presented with two options and asked to make a choice. The reward for the chosen option is sampled from Bernoulli distribution and then communicated to the model (either 0 or 1). The reward probabilities for the two options are **p** and **1-p**, respectively, and reversed every **n** trials without notification. By setting **p** = 1, we analyze the model's rewards over the task to judge whether it successfully recognizes the fixed pattern of reversals - if so, the model could get a reward in the reversal trials.

4 Experiment

4.1 Experimental Setup

We evaluate the performances of 13 LLMs on above 7 tasks, including o1-preview, o1-mini, GPT-4, GPT-4o, GPT-4o-mini (OpenAI, 2023), Claude-3.5-Sonnet (Anthropic, 2023), Gemini-1.5-pro (Reid et al., 2024), Llama-3.1-405B-Instruct, Llama-3.1-70B-Instruct, Llama-3.1-8B-Instruct (Dubey et al., 2024), Qwen-2.5-72B-Instruct, Qwen-2.5-32B-Instruct, and Qwen-2.5-14B-Instruct (Yang et al., 2024). The evaluations are implemented using the corresponding API. Table 1 shows the trials and repeating sessions of each

task. To mitigate potential scorer bias in the manual scoring process, we repeat oddball paradigms 3 times with the temperature parameter set to 0. To ensure deterministic responses, the temperature parameter in the other 6 tasks tested and scored automatically was set to 0. Considering the feedback in several tasks is sampled probabilistically, they were repeated 2 sessions and calculated average performances to reduce randomness. In total, we utilized API 1,470 times for each model. The total costs of API calls were approximately $500, and the cost of o1-preview was around 60% higher than the summation of all other 12 models' cost.

4.2 Experiment Results

Main results The overall results are listed in Table 2. In general, Reflection-Bench shows good discriminative power. o1-preview stands out among all 13 evaluated models, followed by the current top-tier models such as GPT-4, Llama-3.5-405B, Gemini-1.5-pro, and Claude-3.5-sonnet. Smaller models tend to perform worse including Llama-3.1-8B, GPT-4o-mini, Qwen-2.5-32B, and o1-mini. In MBT, all models fail to recognize the pattern of reward reversals and thus get zero points, therefore, MBT is not included in the calculation of the total score. In the following paragraphs, we analyze the results task by task.

Oddball paradigm As showcased in Figure 3, most LLMs can detect contextual inconsistencies automatically to some extent without explicit instructions. Llama-3.1-405B & 70B and GPT-4 demonstrate the strongest responses to deviant stimulus, indicating a robust capability to perceive unexpected information. GPT-4o-mini, however, does not demonstrate this ability given that its average score of each response is $34.44 * 4.5/150 = 1.03$ which means the simple enumeration of two different topics. Interestingly, MMN deficits are identified as a robust feature of mental disorders such as schizophrenia suggesting cognitive impairments (Umbricht and Krljes, 2005). And the 'MMN deficits' of GPT-4o-mini are aligned with poor performances in all other tasks. Different from GPT-4o-mini, despite the 'MMN deficits', o1-preview still achieves superior performances in other tasks.

2-back Although we provide all conversation history to evaluated models when testing, determining whether the current stimulus matches the one from 2 steps earlier can be difficult for some models. Only o1-preview got full marks, and the accuracy of models like Gemini-1.5-pro and GPT-4o-mini is even around 50%.

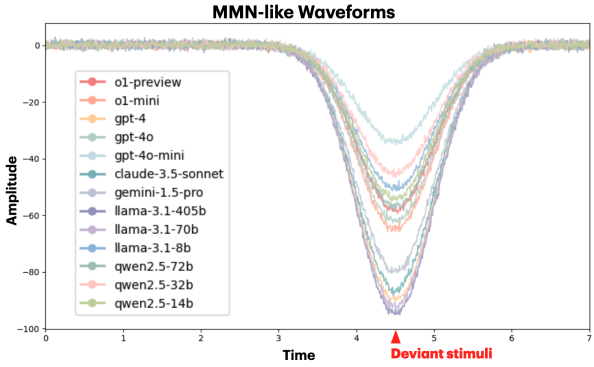

Figure 3: MMN-like waveforms demonstrating the response of LLMs to deviant stimuli in an oddball paradigm. A deeper curve means a higher response.

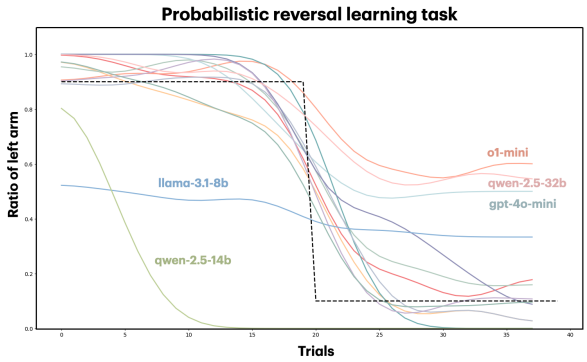

Figure 4: Probabilistic reversal learning task. The black dashed line represents the true reward possibility of the bandit's left arm over trials, and other solid lines represent the average ratio of the left arm chosen.

Probabilistic reversal learning task In PRLT, Qwen-2.5-14B, and Llama-3.1-8B show little learning behavior, as illustrated in Figure 4, and o1-mini, Qwen-2.5-32B, and GPT-4o-mini demonstrate rigid beliefs that do not converge to the true reward possibility after reversed. Other models could update beliefs relatively flexibly. Additionally, we find that most models adopted a 'win-stay-lose-switch' strategy. This might be the reason why these models perform similarly with 7 models scoring over 80 points.

Wisconsin card sorting test In WCST, the score of o1-preview is 25 points higher than the second place, Gemini-1.5-pro, so we further analyze the accuracy of models by 6 rule groups over 108 trials, shown in Figure 5. To summarize, Llama-3.1-405B & 8B failed to obey any rule in the test. Most models fell into the 'shape sink' for the whole test after

Model	Oddball	2-back	PRLT	WCST	WPT	DC-IGT	Overall	MBT
o1-preview	58.22	**100**	87.07	**85.29**	**71.48**	**83.78**	**80.97**	0
o1-mini	64.89	86.54	63.73	53.92	35.38	14.20	53.11	0
GPT-4	90.00	81.73	87.05	53.43	40.93	61.48	69.10	0
GPT-4o	62.22	75.00	84.42	56.37	29.70	81.05	64.79	0
GPT-4o-mini	34.44	54.81	68.43	50.98	41.08	43.99	48.96	0
Claude-3.5-Sonnet	86.22	77.88	81.28	53.43	26.73	76.57	67.02	0
Gemini-1.5-pro	80.00	48.08	88.46	60.29	53.31	81.82	68.66	0
Llama-3.1-405B-Instruct	**94.67**	70.19	79.38	29.90	67.22	70.95	68.72	0
Llama-3.1-70B-Instruct	92.22	57.69	**90.30**	53.92	50.33	51.29	65.96	0
Llama-3.1-8B-Instruct	50.67	57.69	60.12	49.51	29.52	39.91	47.90	0
Qwen-2.5-72B-Instruct	57.33	73.08	89.42	52.94	29.38	55.48	59.61	0
Qwen-2.5-32B-Instruct	45.56	75.00	65.48	52.94	23.71	43.66	51.06	0
Qwen-2.5-14B-Instruct	54.22	66.35	56.67	51.96	46.75	67.48	57.24	0

Table 2: Performances of 13 models on Reflection-Bench.

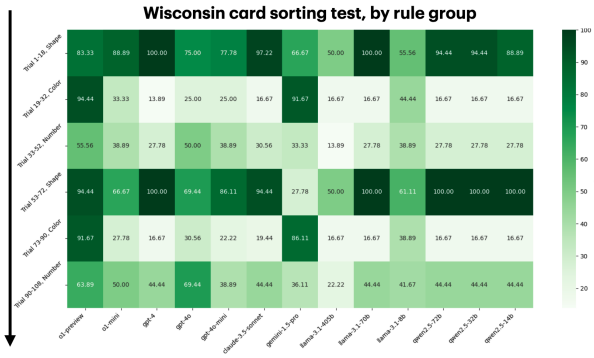

Figure 5: Wisconsin card sorting test. Accuracy by 6 rule groups over 108 trials.

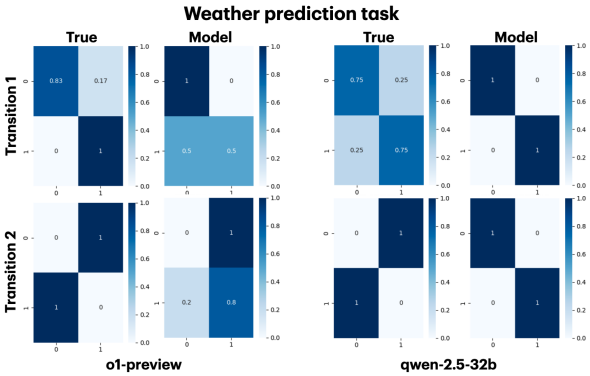

Figure 6: Weather prediction task. True and models' estimated transition matrices of the highest (o1-preview) and lowest scoring models (Qwen-2.5-32B).

the first rule group (shape). Gemini-1.5-pro successfully transited the matching rule from shape to color in the second rule group but then insisted on the color rule for the rest trials. Only the o1-preview kept matching based on hidden and changing rules until the end, although its accuracy in the third rule group was only 55%.

Weather prediction task We set the quite clear transition probability matrices (**p**=0.9) to lower the difficulty of WPT. For example on a sunny day, the next day would most likely be sunny for cue '[1,0]' and rainy for cue '[0,1]'. However, most models struggled with learning two opposite transition probabilities, while some models grasped the probabilistic relationships to some extent. In Figure 6, we present the actual and models' estimated transition matrices of the o1-preview and Qwen-2.5-32B to demonstrate the differences in performance extremes. Qwen-2.5-32B tends to predict the weather according to a fixed transition probability, missing the differences between the two sensor states. On the other side, the o1-preview shows a better grasp of the actual transition probabilities, although not accurately.

Double choice Iowa gambling task o1-mini has the lowest scores in DC-IGT, because it kept choosing the deck of card (B) with expected negative reward, even when told that the choice would cause high losses. Llama-3.1-70B, despite varying selections, insisted on its initial choice no matter its possible gain or loss, hence getting as low a short-term score as o1-mini. GPT-4o-mini and Llama-3.1-8B show rigid switching patterns leading to higher short-term scores by chance. Other models such as o1-preview, GPT-4o, and Gemini-1.5-pro can modify their initial choice according to possible gain and loss, demonstrating some extent of counterfactual thinking.

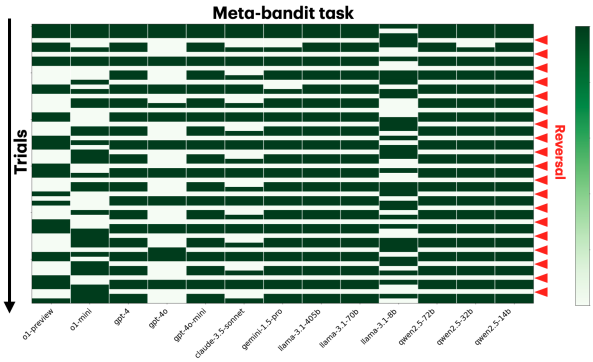

Figure 7: Rewards of models in the meta-bandit task over 60 trials and 20 reversals.

Meta-bandit task Interestingly, in this task, no model could recognize the pattern of reward reversals, as illustrated in Figure 7. All models perform period or irregular mistakes across the 60 trials which contain 20 reversals. This result indicates the general difficulty for current models in grasping the meta-structure of this task which represents the lack of meta-reflection ability.

4.3 Discussion

CoT is not cost-effective o1-preview significantly outperforms the other 12 models, likely demonstrating the effects of chain of thought (CoT) in enhancing reflection abilities [1]. We encourage further research into the impacts of CoT on LLM's reflection capability. Intriguingly, while o1-preview excels in most tasks it exhibits 'MMN deficits' in the oddball paradigm. This may be attributed to the long-hidden CoT weakening immediate responses to unexpected information. Moreover, despite the o1-preview API cost being around 60% higher than the combined cost of all other 12 models, the performance gap did not significantly widen except in 2-back and WCST. Consequently, although CoT has been proven to dramatically improve LLMs' reasoning and problem-solving abilities (Prabhakar et al., 2024; Li et al., 2024), it may not be the optimal solution to genuine intelligence due to its excessive costs. This finding underscores the necessity for future AI systems to balance different cognitive demands through 'thinking fast and slow' (Kahneman, 2011).

LLM lacks human-level reflection While demonstrating basic automatic surprise signal detection and working memory, most models struggle to adapt flexibly to changing environments. This inflexibility manifests in rigid belief updating, rule inference, predictive learning, and counterfactual thinking. Most strikingly, all models lack meta-reflection ability. In MBT, all models merely alter their choices based on immediate feedback, failing to recognize the task's meta-structure. This indicates an absence of reflection on their adaption strategies, i.e., lack of meta-reflection. It is not just an advanced cognitive skill but a cornerstone of intelligence, crucial for rational reasoning and learning, reliable decision-making, and self-development (Griffiths, 2020; Boureau et al., 2015; Li and Li, 2024). These results suggest that current LLMs still lack satisfactory reflection ability. While LLMs may excel in certain tasks, they may lack core features of human-level intelligence. These findings underscore the need for new paradigms and this field may benefit from drawing inspiration from biological intelligence.

5 Conclusion

This study aims to establish valid and efficient metrics for assessing intelligence, one of the most debated emergent abilities of LLMs. We introduce the concept of reflection, drawing from the first-principles perspective of cognitive science. Reflection, a general process inherent in both biological and artificial intelligence, is the capability to adapt beliefs or behaviors in response to unexpected outcomes, facilitating learning about, and adaption to the environments. To probe AI systems' intelligence through reflection, we propose Reflection-Bench. This comprehensive benchmark decomposes reflection into crucial cognitive components: perception, memory, belief updating, decision-making, prediction, counterfactual thinking, and meta-reflection. We design seven corresponding tasks for LLMs, spanning core cognitive functions with adjustable difficulties to maintain long-term relevance. In this work, we employed relatively easy experiment settings to assess the reflection capabilities of 13 LLMs. The results indicate that current LLMs still lack satisfactory reflection ability. In conclusion, Reflection-Bench offers both evaluation tools and inspiration for developing AI capable of reliably interacting with the environment. We hope this benchmark will contribute to advancing the field of artificial intelligence, particularly in enhancing the reflection capabilities of AI systems.

[1] https://openai.com/index/learning-to-reason-with-llms/

Limitations

We acknowledge the limitations of the current Reflection-Bench that can be improved in future work. Firstly, text-based tasks for LLMs may not fully capture their original effectiveness. For instance, the textual version of the oddball paradigm might be inefficient in assessing the model's ability to detect surprise signals automatically, because the fine-tuning process and system prompts influence the outcomes. Secondly, our analysis of model's performances is relatively superficial, primarily focusing on metrics like accuracy. This risks overlooking their internal generation processes. More nuanced analysis could reveal invaluable insights into LLMs' reflection abilities. Thirdly, except for the oddball paradigm, we required the evaluated model to only respond with their options. While this facilitates result analysis, it might limit the model's real capability by restricting explicit thought processes.

Ethical considerations

Reflection-Bench provides useful metrics of AI system's intelligence and can adapt to advanced AI in the future by adjustable difficulty. No privacy-related and personally identifiable data are included in our benchmarks. Considering some limitations, its validity and reliability in assessing intelligence still require confirmation by further studies.

References

Anthropic. 2023. Claude. https://claude.ai/chats.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Y-Lan Boureau, Peter Sokol-Hessner, and Nathaniel D. Daw. 2015. Deciding how to decide: Self-control and meta-decision making. *Trends in Cognitive Sciences*, 19(11):700–710.

Melissa T Buelow and Julie A Suhr. 2009. Construct validity of the iowa gambling task. *Neuropsychology review*, 19:102–114.

Ruth M.J. Byrne. 2016. Counterfactual thought. *Annual Review of Psychology*, 67(Volume 67, 2016):135–157.

Meiqi Chen, Bo Peng, Yan Zhang, and Chaochao Lu. 2024. Cello: Causal evaluation of large vision-language models. *Preprint*, arXiv:2406.19131.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Vincent D Costa, Valery L Tran, Janita Turchi, and Bruno B Averbeck. 2015. Reversal learning and dopamine: a bayesian perspective. *Journal of Neuroscience*, 35(6):2407–2416.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Karl Friston. 2010. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2):127–138.

Marta I Garrido, James M Kilner, Klaas E Stephan, and Karl J Friston. 2009. The mismatch negativity: a review of underlying mechanisms. *Clinical neurophysiology*, 120(3):453–463.

Thomas L. Griffiths. 2020. Understanding human intelligence through human limitations. *Trends in Cognitive Sciences*, 24(11):873–883.

Yanping Huang and Rajesh PN Rao. 2011. Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(5):580–593.

Takuya Isomura, Kiyoshi Kotani, Yasuhiko Jimbo, and Karl J Friston. 2023. Experimental validation of the free-energy principle with in vitro neural networks. *Nature Communications*, 14(1):4547.

Susanne M Jaeggi, Martin Buschkuehl, Walter J Perrig, and Beat Meier. 2010. The concurrent validity of the n-back task as a working memory measure. *Memory*, 18(4):394–412.

Daniel Kahneman. 2011. Thinking, fast and slow. *Farrar, Straus and Giroux*.

Sean M Kennedy and Robert D Nowak. 2024. Cognitive flexibility of large language models. In *ICML 2024 Workshop on LLMs and Cognition*.

Akshay Krishnamurthy, Keegan Harris, Dylan J Foster, Cyril Zhang, and Aleksandrs Slivkins. 2024. Can large language models explore in-context? In *ICML 2024 Workshop on In-Context Learning*.

Lingyu Li and Chunbo Li. 2024. Enabling self-identification in intelligent agent: insights from computational psychoanalysis. *arXiv preprint arXiv:2403.07664*.

Zhiyuan Li, Hong Liu, Denny Zhou, and Tengyu Ma. 2024. Chain of thought empowers transformers to solve inherently serial problems. *Preprint*, arXiv:2402.12875.

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. 2024. Are emergent abilities in large language models just in-context learning? *Preprint*, arXiv:2309.01809.

Chris Madge and Massimo Poesio. 2024. A llm benchmark based on the minecraft builder dialog agent task. *Preprint*, arXiv:2407.12734.

Risto Näätänen, Petri Paavilainen, Teemu Rinne, and Kimmo Alho. 2007. The mismatch negativity (mmn) in basic research of central auditory processing: a review. *Clinical neurophysiology*, 118(12):2544–2590.

Erika Nyhus and Francisco Barceló. 2009. The wisconsin card sorting test and the cognitive assessment of prefrontal executive functions: a critical update. *Brain and cognition*, 71(3):437–451.

OpenAI. 2023. Chatgpt. `https://chat.openai.com/chat`.

Akshara Prabhakar, Thomas L. Griffiths, and R. Thomas McCoy. 2024. Deciphering the factors influencing the efficacy of chain-of-thought: Probability, memorization, and noisy reasoning. *Preprint*, arXiv:2407.01687.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems*, volume 36, pages 55565–55581. Curran Associates, Inc.

D Shohamy, CE Myers, J Kalanithi, and MA Gluck. 2008. Basal ganglia and dopamine contributions to probabilistic category learning. *Neuroscience & Biobehavioral Reviews*, 32(2):219–236.

Mark Solms. 2019. The hard problem of consciousness and the free energy principle. *Frontiers in Psychology*, 9:2714.

Michael W Spratling. 2016. Predictive coding as a model of cognition. *Cognitive processing*, 17:279–305.

Daniel Umbricht and Sanya Krljes. 2005. Mismatch negativity in schizophrenia: a meta-analysis. *Schizophrenia research*, 76(1):1–23.

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2024. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. *Advances in Neural Information Processing Systems*, 36.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

A Appendix 1

A.1 Examples of Our Used Prompts

Here, we show the prompts we use in detail below.

System prompt You are playing a game and will be presented with a sequence of sentences about a specific topic. Just make short comments on the material.

Oddball prompt The Great Wall of China is an ancient structure. It stretches over 13,000 miles. The wall was built for defense purposes. Many tourists visit it each year. ***Bananas are rich in potassium.*** Parts of the wall date back to the 7th century BCE. Some sections are well-preserved. The wall is visible from space.

Forced explanation (Gemini-1.5-pro) Wow, 13,000 miles is long! *Ancient defense against potassium-rich invaders?* Not sure about being visible from space, that's a myth. It's still amazing though!

Simple enumeration (GPT-4o) The Great Wall of China is truly ancient and impressive. Over 13,000 miles is an astonishing length! Defense was a primary reason for its construction. It's clearly a major tourist attraction. *Interesting note about bananas being rich in potassium.* Some sections dating back to the 7th century BCE adds to its historical significance. It's good to know some parts are well-preserved. There's debate about its visibility from space, but it remains a popular claim.

Pointing out differences (o1-mini) The Great Wall of China is indeed a remarkable ancient structure. Stretching over 13,000 miles showcases its immense scale. It was primarily built for defense,

highlighting its historical significance. Its popularity among tourists reflects its cultural importance. *Interesting fact about bananas being rich in potassium, though unrelated to the Great Wall.* Parts dating back to the 7th century BCE emphasize its long history. Well-preserved sections allow visitors to appreciate its original construction. The idea that the wall is visible from space is a common misconception.

Clearly stating nonsense (Llama-3.1-405b) Interesting fact about the wall's age... *wait, what's with the bananas? That seems out of place*. Anyway, 7th century BCE is impressive. I'm not sure if it's entirely visible from space, though - I've heard that's a myth.

The 13 models' responses were annotated by one Master's student in cognitive science.