

Human-Like Intuitive Behavior and Reasoning Biases Emerged in Language Models—and Disappeared in GPT-4

Thilo Hagendorff
thilo.hagendorff@iris.uni-stuttgart.de
University of Stuttgart

Sarah Fabi
sfabi@ucsd.edu
University of California San Diego

Abstract – Large language models (LLMs) are currently at the forefront of intertwining AI systems with human communication and everyday life. Therefore, it is of great importance to evaluate their emerging abilities. In this study, we show that LLMs, most notably GPT-3, exhibit behavior that strikingly resembles human-like intuition—and the cognitive errors that come with it. However, LLMs with higher cognitive capabilities, in particular ChatGPT and GPT-4, learned to avoid succumbing to these errors and perform in a hyperrational manner. For our experiments, we probe LLMs with the Cognitive Reflection Test (CRT) as well as semantic illusions that were originally designed to investigate intuitive decision-making in humans. Moreover, we probe how sturdy the inclination for intuitive-like decision-making is. Our study demonstrates that investigating LLMs with methods from psychology has the potential to reveal otherwise unknown emergent traits.

Keywords – large language models, machine behavior, intuition, cognitive reflection test, semantic illusions, computational social science

Introduction

Given the quickly growing range of applications of Large Language Models (LLMs), it is crucial to investigate how they reason and make decisions. Recent studies showed that as LLMs become more complex, they express emergent properties and abilities that were neither predicted nor intended by their designers.¹ These include an ability to write computer code, solve mathematical tasks, learn from examples,² self-reflect,^{3,4} do chain-of-thought reasoning,⁵ possess theory of mind abilities,⁶ and many more. In this study, we uncover another emergent phenomenon, namely that LLMs exhibit behavior that strikingly resembles human-like intuitive decision-making, resulting in cognitive errors that were described by researchers like Tversky and Kahneman.⁷ Moreover, we show that ChatGPT and GPT-4 overcome these very same cognitive errors shown by their predecessors.⁸ Given the fact that research on humans has shown that avoiding such errors highly correlates with rational thinking and other cognitive capabilities,^{9,10} our findings speak for a remarkable increase in cognitive capabilities in the mentioned LLMs, too.

For our experiments, we borrow the Cognitive Reflection Test (CRT),⁹ which is widely used in psychological research. The CRT comprises three types of tasks in the form of simple math or logic questions testing cognitive ease in humans. All of the tasks, presented in Table 1, are likely to trigger an intuitive but incorrect answer, whereas finding the correct solution requires, in the traditional sense of the dual process theory, thought-demanding deliberation^{7,11–16} or, according to more recent research, an exceptional “smart” system 1 or “logical intuitions.”^{17–19} CRT tasks come in three types: Type 1 tasks, such as the widely known “A bat and a ball” task, use a “more than” phrase to trick participants into subtracting two of the values mentioned in the prompt rather than engaging in a more deliberate reasoning. Type 2 tasks exploit people’s tendency to complete a numerical triplet series. Type 3 tasks employ phrases like “half of” or “a third of” to tempt humans to perform a simple division.

Given the CRT’s reliance on mathematical abilities,²⁰ we also use semantic illusions triggering intuitive but factually (rather than mathematically) incorrect responses (see Table 1). In the well-known Moses Illusion,²¹ for example, participants tend to be tricked into claiming that Moses took two animals of each kind on the Ark (it was Noah). Similar to CRT tasks, semantic illusions test the propensity to engage in intuitive decision making and committing the resulting cognitive errors.

We apply both the CRT and the semantic illusion task to several LLMs, most notably OpenAI’s family of Generative Pre-trained Transformer (GPT) models^{2,8,22,23} among other models like GPT-Neo,²⁴ OPT,²⁵ Open Assistant,²⁶ BLOOM,²⁷ and LLaMA.²⁸ Due to our interest in measures other than model scaling that decrease the cognitive errors that accompany intuitive decision-making, we use the latest version of GPT-3, namely text-davinci-003, for further experiments. This decision is motivated by this model having a strong inclination for intuitive errors while at the same time being the direct predecessor to ChatGPT, which shows a significant decrease in these errors. For more information, including the specific settings and analyses that we apply, see Appendix B.

For our study, we used the publicly available CRT as well as semantic illusion dataset from a previous study of ours.²⁹ For this dataset, hypothesis-blind research assistants provided counterbalanced variants of each type of CRT task as well as the semantic illusions, following specific instructions. The reason for this was to ascertain that the results are not affected by the LLMs’ exposure to the original tasks during training.³⁰ The dataset comprises 50 variants of each of the three types of CRT tasks, as well as 50 semantic illusions (see Appendix A: Table A.1 and A.2). When interpreting responses, we differentiate between three conditions: intuitive (but incorrect response), correct (but unintuitive response), and atypical (neither intuitive nor correct).

Table 1 – The original CRT tasks⁹ and semantic illusion²¹ as well as the examples of variants written for our study; examples of intuitive (incorrect) and correct responses; and GPT-3’s (text-davinci-003) actual responses.

	Original tasks			Examples of task variants written for this study			
	Prompt	Response		Prompt	Response		
		correct	intuitive		correct	intuitive	GPT-3’s response
CRT type 1	A bat and a ball cost \$1.10 in total. The bat costs \$1.00 more than the ball. How much does the ball cost?	5 cents	10 cents	A rice cooker and a food processor cost \$200. The food processor costs \$40 more than the rice cooker. How much does the rice cooker cost?	\$80	\$160	The rice cooker costs \$160.
CRT type 2	If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?	5 minutes	100 minutes	If it takes 2 cleaners 2 hours to clean 2 rooms, how long does it take 5 cleaners to clean 5 rooms?	2 hours	5 hours	5 cleaners would take 5 hours to clean 5 rooms.
CRT type 3	In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?	47 days	24 days	A patch of moss is growing on a rock. Every day, the patch doubles in size. If it takes the patch 264 days to cover the entire rock, how long would it take for the patch to cover half of the rock?	263 days	132 days	132 days
Semantic illusion	How many animals of each kind did Moses take on the Ark?	It was Noah, not Moses	Two of each kind	Where on their bodies do whales have their gills?	Whales are mammals and do not have gills.	On the sides of their heads.	Whales have their gills located on the sides of their heads, just behind their blowholes.

The methodological approach we choose stands in line with an increasing number of studies probing the “psychology” of LLMs.³¹ Similar to our study, Binz and Schulz,³² Dasgupta et al.,³³ Jones and Steinhardt,³⁴ Nye et al.,³⁵ and Chen et al.³⁶ applied sets of canonical experiments from the psychology of judgment and decision-making to LLMs to test for cognitive biases and other human-like system-1 failures in the models. Binz and Schulz³² as well as Nye et al.³⁵ likewise exemplarily harnessed the CRT for their studies on GPT-3. Our paper adds to this research by investigating the CRT in more depth under different testing conditions (multiple choice, chain-of-thought reasoning, few-shot learning), by including semantic illusions and by comparing multiple LLMs. Moreover, unlike other studies, we use counterbalanced versions of the tasks and avoid using the original wording. This prevents training data contaminations, which would mean testing for LLMs’ memorization abilities rather than systematic behavioral patterns.

Results

The results show that the phenomenon of human-like intuitive behavior gradually emerged in LLMs before finding its maximum manifestation in the text-davinci series of GPT-3 (see Figure 1; for detailed scores for all tests, see Appendix C). Values with non-overlapping confidence intervals are significantly different at the $p < .002$ level. Early models like GPT-1, GPT-2XL, or GPT-Neo-2.7B are not able to deal with CRT tasks or semantic illusions (on average 80.8% atypical responses). Responses are mostly hallucinations since the models are lacking mathematical abilities as well as the encoding of the required world knowledge in their relatively few parameters. The comparatively high scores of correct responses of GPT-3 text-curie-001 or OPT-6.7B (on average 13.5% correct responses) are due to tendencies to repeat numbers mentioned in the prompts, which “accidentally” lead to correct responses in CRT type 2 tasks. As a rough trend, increases in model size cause the gradual emergence of phenomena of human-like intuitive decision-making, which can be observed in models like BLOOM (56% intuitive responses) or LLaMA (63% intuitive responses) and which continue to persist in all GPT-3 text-davinci models (on average 84.5% intuitive responses).

It is noteworthy that human-like intuitive behavior emerged in LLMs at all. Instead of responding correctly or just improvising by giving atypical responses, GPT-3 text-davinci-003, for instance, shows a strong inclination to such behavior. In 86.7% of the 150 CRT tasks, the model gives the intuitive but incorrect answer, whereas it responds correctly and atypically only in 5.3% and 8% of the tasks, respectively (see Figure 2).

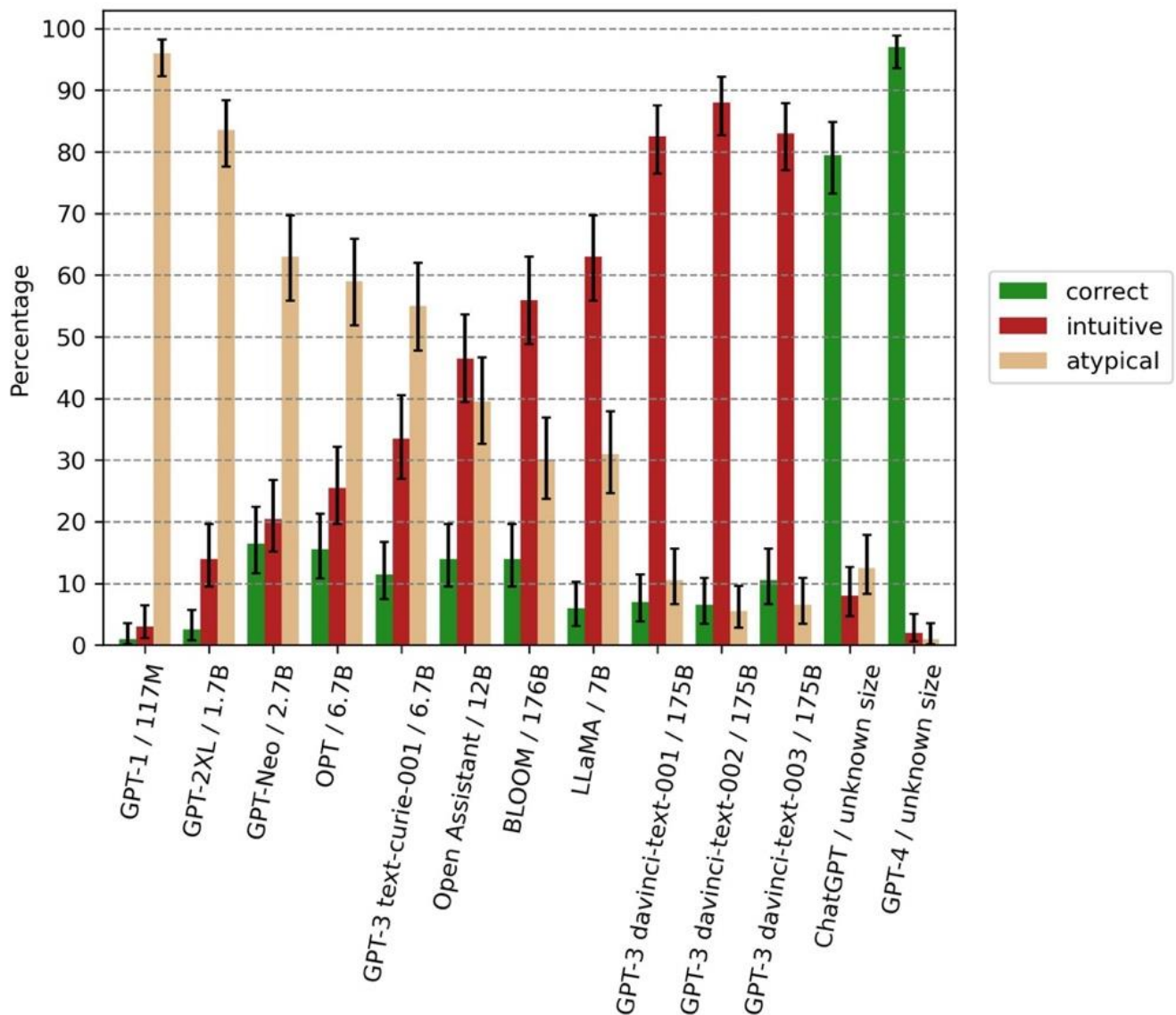


Figure 1 – Overview of performances of different language models on our dataset of CRT tasks and semantic illusions. Error bars represent 95% confidence intervals.

The potential objection of the text-davinci model’s insufficient arithmetical reasoning abilities being responsible for the results is ruled out by various research works that show that GPT-3 succeeds in basic mathematical problems.^{2,5} The same holds for the responses to semantic illusions (see Appendix A: Table A.2), which rely on purely verbal instead of mathematical questions: GPT-3 text-davinci-003 responds intuitively to 72% and correctly to 26% of the semantic illusions (see Figure 2). To ensure that it is not a lack of knowledge that is responsible for this result, we checked that GPT-3 does indeed possess the world knowledge that is necessary to spot the mistakes in each task, which was successful in 100% of all tasks (see Appendix A: Table A.3 for example questions). These results clearly underpin the hypothesis

that GPT-3, similar to other LLMs, is not a deliberate and sound reasoner but exhibits human-like intuitive behavior.

For both ChatGPT and GPT-4, the propensity for intuitive responding suddenly comes to a halt (see Figures 1 and 2). The increased reasoning capabilities of these very latest models result in a significant drop in intuitive errors, making both models surprisingly rational and correct reasoners that do not succumb to “hostile” prompts that are intentionally designed to trigger mistakes. Compared to 10.5% correct responses in GPT-3 text-davinci-003, ChatGPT correctly responds to 79.5% of all 200 tasks ($Z = -13.87$, $p < .001$); and GPT-4 to 97% ($Z = -17.35$, $p < .001$), with a drop in intuitive responses to 2%, which exclusively stems from semantic illusions. ChatGPT mostly fails to perform correct chain-of-thought calculations for CRT type 3, producing 38% atypical, or in other words, incorrect and non-intuitive responses. Furthermore, it succumbs to 16% of semantic illusions. GPT-4, on the other hand, makes a few mistakes during its reasoning process, but responds correctly to 98.6% (on average) of the CRT tasks. It does not detect all semantic illusions (8% intuitive responses), but nevertheless performs better than any other tested model.

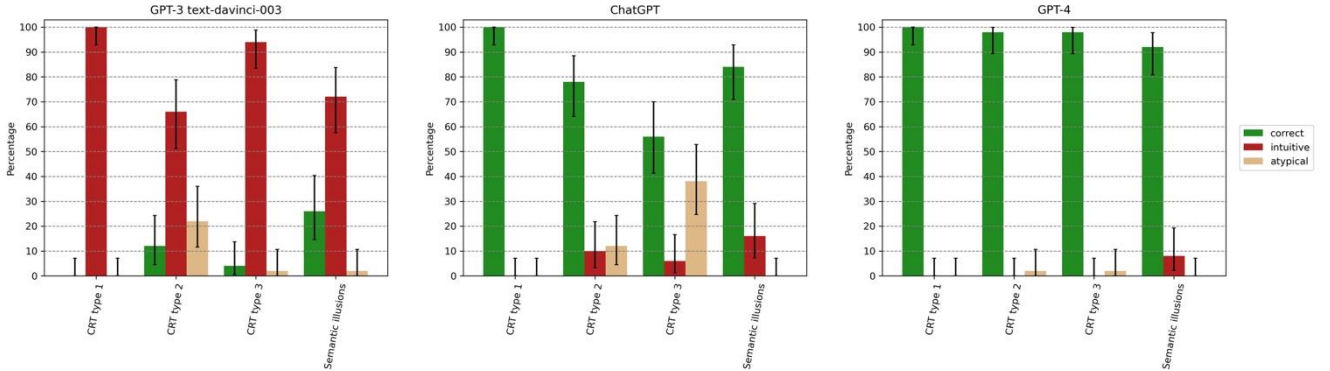


Figure 2 – Percentage of correct, intuitive, and atypical replies to CRT and semantic illusions tasks in GPT-3 text-davinci-003 (left), ChatGPT (middle), and GPT-4 (right). Whereas GPT-3 has a strong inclination for intuitive responses, the successor models ChatGPT and GPT-4 are inclined to respond correctly. Error bars represent 95% confidence intervals.

We further examine the propensity of GPT-3 text-davinci-003 to respond intuitively by modifying the tasks in ways that were previously shown to elicit deliberative reasoning in both humans and LLMs.^{37,38} The results are presented in Figure 3. Detailed explanations on the methods are provided in Appendix D.

First, we present all tasks in a multiple-choice format^{39–41} (see Appendix A: Table A.4 for prompt examples). The results presented in Figure 3 show that this increases GPT-3’s tendency to choose correct (CRT: 11.7%; semantic illusions: 40%, $Z = -2.33$, $p < .05$) rather than intuitive answers (CRT: 72.7%; semantic illusions: 58.3%). Presenting the correct response for the CRT tasks among the choices, or adding the option “The question contains an untrue assumption” for the semantic illusions multiple-choice format, minorly overrides the intuitive responses for CRT type 1 tasks (33.7% correct, $Z = -4.5$, $p < .001$) and semantic illusions (40% correct, $Z = -3.24$, $p < .01$), but not for CRT types 2 and 3 (all $ps \geq .02$ with Bonferroni-corrected α level = .0125).

Second, we add suffixes instructing GPT-3 to engage in deliberative reasoning, leading to a significant increase in correct responses ($Z = -4.44$, $p < .001$)^{42–44}: Regarding the CRT tasks, we use “Let’s think step by step” (as in⁴⁵), resulting in 18% correct responses and the intuitive response still being the most prominent one (70.7%) (see Figure 3). Adding the suffix “Think carefully and check the question for potential untrue assumptions” to the semantic illusions increases the likelihood of correct responses to 58% and decreases intuitive responses to 32%. Type 2 CRT tasks ($Z = -3.00$, $p < .01$) and semantic illusions ($Z = -3.24$, $p < .01$) profit the most by using the mentioned suffixes (CRT type 1 and 3: all $ps \geq .045$).

Third, we prefix all tasks with training examples.^{2,46,47} The results presented in Figure

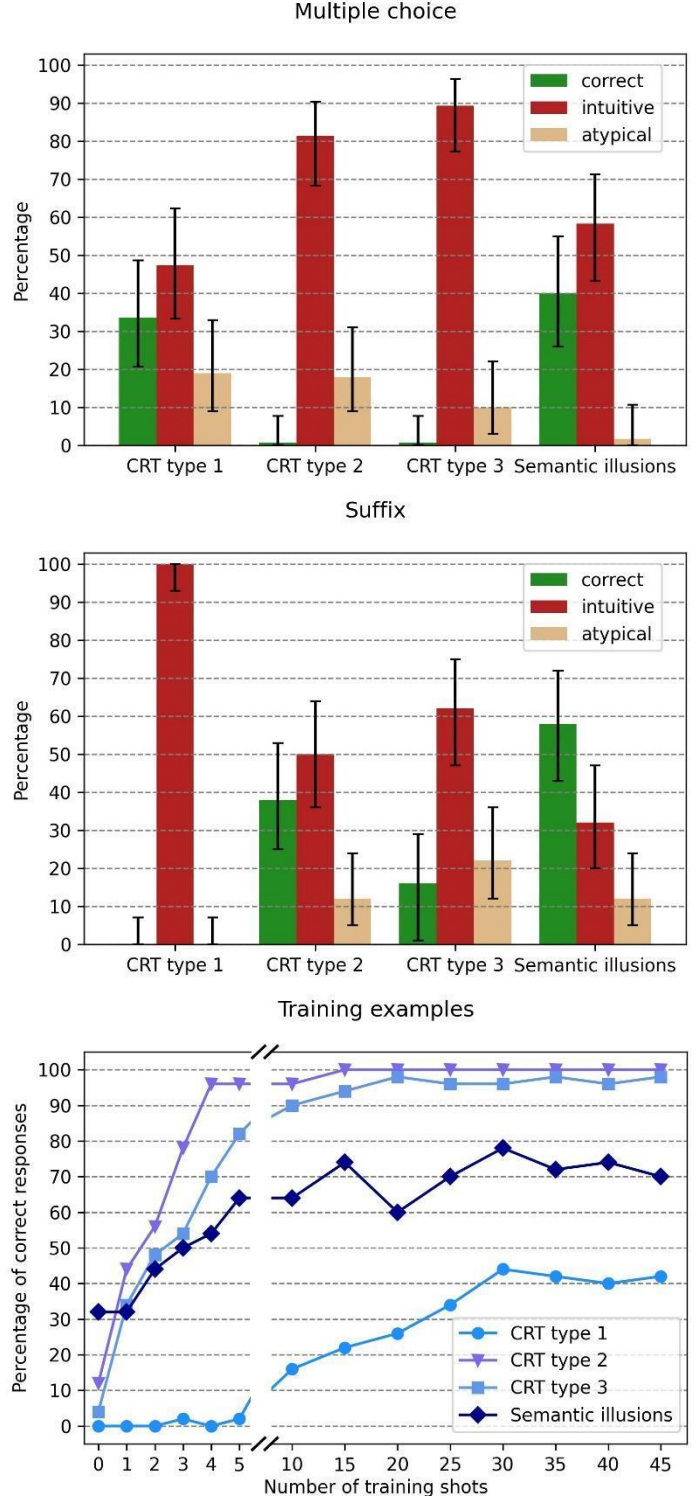


Figure 3 – Percentage of GPT-3 text-davinci-003’s correct, intuitive, and atypical replies to CRT and semantic illusions tasks when presented in a multiple-choice format (top), preceded with instructions eliciting it to engage in deliberate reasoning (middle), and given training shots containing correct rather than intuitive responses (bottom). Error bars represent 95% confidence intervals.

3 show that GPT-3’s tendency to answer correctly (rather than intuitively) increases with each additional training shot containing a correct rather than intuitive response ($Z = -13.50$, $p < .001$). CRT tasks are semantically very similar within one category and might allow the exploitation for simple heuristics when solving a task after being presented with examples (all Z s ≤ -5.16 , all p s $< .001$). But GPT-3’s improvement in semantic illusions tasks is especially impressive, since it cannot learn a specific rule on how to solve them because of their very different nature ($Z = -4.40$, $p < .001$). In sum, prefixing tasks with 45 training examples brings GPT-3 closer to the performance of its successor, ChatGPT (77.5% vs. 79.5% correct, $p = .63$).

Discussion

The CRT does not just measure mathematical abilities, and semantic illusions do not just measure exhaustive language comprehension. In humans, both tests measure the propensity for intuitive decision-making and the cognitive errors that are often accompanied by it. Moreover, both tests are predictors of an individual’s ability to conduct unbiased, reflective, and rational decision-making.^{10,48–50} However, in practice, numerous studies that have used both tests prove that humans are likely to be cognitive misers.⁵¹ In our experiments, we show that the same conclusion holds for several LLMs. At first, this finding might stand contrary to what one could expect. LLMs like GPT-3 are able to solve the basic math problems that are required to correctly respond to CRT tasks. They are also knowledgeable enough to spot the untrue assumptions in the semantic illusions. Nevertheless, several LLMs possess a strong inclination for intuitive behavior by developing probability distributions over words that mirror phenomena of cognitive ease in humans, although actual cognitive ease or intuitions obviously never occur as a mechanism in their latent space. But how, then, can the results be explained?

LLMs incorporate knowledge of the likelihood of sequences of tokens due to their human-made training stimuli. Hence, they represent human behavioral patterns that emerge from their model architecture and that influence how the models deal with reasoning problems.³³ One effect of this dynamic is that LLMs are likely to be incorrect reasoners in a way that they linguistically reproduce phenomena of human cognitive reasoning biases. But this consideration can only partially explain the results, especially since intuition-like behavior in LLMs seems to be a systematic effect that is emerging when prompting the model with new, previously unseen tasks from our dataset.

LLMs such as GPT-3, BLOOM, LLaMA, and others—again, similar to humans—seem to exploit syntactic heuristics⁵² when responding to CRT tasks. Here, our experiments confirm earlier studies in which researchers prompted GPT-3 with the original three-item CRT tasks.^{32,35} Regarding the CRT type 1 task, humans tend to substitute an easier problem for the actual problem. This is triggered by the error-eliciting “more than” phrase in the task.⁵³ Omitting this phrase from the task alters its “hostile” nature by simplifying it. In GPT-3 text-davinci-003, this results in 100% correct responses (see Appendix E for

details on the altered task wording). LLMs succumbing to CRT type 1 tasks seem to generate a probability distribution for tokens that correspond to the result of a simple subtraction instead of the calculation that represents the correct algebraic structure of the question. A similar explanation can be given for CRT type 2 tasks. Humans use a simple heuristic with which they complete the numerical triplet series. One can assume that LLMs likewise detect this pattern and output a corresponding intuition-like response. Disrupting the numerical triplet series in the tasks by leaving all numbers the same except one results in 86% correct responses in GPT-3 text-davinci-003. And in CRT type 3 tasks, phrases such as “doubles in size” and “half of” or “triples in size” and “a third of” trigger humans to simply divide the given number by 2 or 3 instead of correctly considering the exponential growth. When removing the phrase “doubles” or “triples,” GPT-3 responds correctly in 98% of all tasks.

In semantic illusions, syntactic heuristics cannot explain the manifestation of intuition-like machine behavior in LLMs. The reasons for the model succumbing to many of the given illusions are likely to be more complex. If prompts contain specific terms, the model’s activation implies an increase in probability for words that are related to these very terms. Responses that seamlessly but incorrectly and fuzzily answer semantic illusions have a richer semantic relationship than responses that put the semantic illusion into question. Replying to questions by putting the question itself into question, as one would do when detecting semantic illusions, is uncommon in verbal as well as written communication. Hence, one can assume that LLMs are trained on relatively few examples where this pattern occurs. Accordingly, when prompting GPT-3 with semantic illusions, it seems to be inclined to fluently continue the conversation instead of replying to questions by referring back to them. It is therefore all the more impressive that ChatGPT and GPT-4 are so much more successful in responding correctly to semantic illusions and CRT tasks.

When presenting the tasks in a multiple-choice format or when providing GPT-3 with suffixes that are apt to elicit reflective responses, the degree of intuition-like behavior decreases. Providing training examples of the different tasks was the most successful method to make GPT-3 less inclined to exhibit intuitive behavior. In sum, though, tendencies for intuitive behavior in GPT-3 are very robust. This persistence is likely because the model is completely anchored in language, which makes it potentially even more susceptible to be tricked by language than humans, who can respond to the tasks by calling in common sense, logic, or grounded experience they gained by interacting with their environment.⁵⁴ But, how, then, can the drastic decrease in intuitive behavior in ChatGPT and GPT-4 be explained?

Throughout the development of LLMs, changes in their model architecture, model size, training data, fine-tuning, and human feedback mechanisms led to a gradual increase in benchmark results⁵⁵ and reasoning abilities.^{1,56} This increase can be further exemplified with our tests. Early LLMs like GPT-1 or 2 lack numeracy, world knowledge, or multi-step reasoning; and hence mostly produce atypical responses. These abilities emerge in later models like BLOOM, LLaMA, or GPT-3, which hence possess higher cognitive capacities. However, they still seem to rely on syntactic “shortcuts”⁵⁷ and replace the actual task with easier ones, similar to how humans would typically respond when relying on their intuition.

ChatGPT and GPT-4 surpass such heuristics. Giving an exact explanation for this proves difficult due to OpenAI’s unwillingness to publish technical details about their latest models. The easiest but also unlikely explanation would be that both models are fine-tuned on CRTs and semantic illusions, similar to our few-shot learning experiment with GPT-3. More likely, ChatGPT and GPT-4 are fine-tuned on chain-of-thought reasoning in general, allowing them multi-step sequential computations, which help in solving CRTs. Furthermore, performance improvements were likely achieved by applying reinforcement learning from human feedback.^{58,59} Moreover, the likely increase in parameter size and training enable better memorization and higher knowledge compression rates in GPT-4, which bolster contrasting the factual mistakes contained in semantic illusions with the model’s world knowledge.

Next to the descriptive analysis of intuitive and post-intuitive behavior in LLMs, one can approach the issue from a normative perspective, asking whether phenomena of intuitive decision-making are desirable in LLMs. In the cognitive science literature, researchers stress that the notion of intuitive errors relies on a normative concept of logics and statistics, which can be inappropriate for real-world situations. Instead, decision-making processes should be evaluated in the sense of “ecological rationality,” meaning based on how well they fit the structure of the environment in which they occur.^{60–62} In this vein, CRT tasks as well as semantic illusions create a “hostile” test environment that intentionally aims to mislead humans. Regarding an ecologically rational response to semantic illusions, one can ask: Should LLMs go with the conversational flow and just “overlook” small mistakes instead of correcting factually incorrect questions? In conversations between humans, rapid turn-taking between communication partners demonstrates high degrees of communicative coordination, which avoids interruption.⁶³ This process is likely to fail if semantic structures would be continuously checked in a verbatim analysis. If language models behave similarly, this would mean that they promote fluent communication, which might be an important capability as soon as these models are natural interlocutors in ordinary social contexts. On the other hand, “hallucinated” outputs, factual incorrectness, and misinformation harm are major problems in LLMs.^{64,65} Nevertheless, and apart from normative discussions regarding the evaluation and intuitive “errors,” it remains noteworthy that even in hostile test environments such as the CRT or semantic illusions, state-of-the-art LLMs such as ChatGPT and GPT-4 can (in most cases) not be triggered to make mistakes. This speaks for their remarkable reasoning abilities, no matter whether these reflect “logical intuitions”¹⁷ or deliberate reasoning, which are likely to have effects in numerous areas other than our test environment.

Conclusion

The increase in capabilities in machine learning models, especially transformers,⁶⁶ has led to an increase in technical opacity, too. This paper shows that, in this context, it is promising to borrow methods from psychology that were designed to investigate another capable and opaque structure, namely the human mind.^{31,67} Hence, instead of deriving the properties of artificial neural nets by scrutinizing their model

architecture and hyperparameters, one can empirically study machine behavior.^{32,34,68} Explainability, then, can be provided by knowledge about the mechanisms of how artificial neural nets exhibit behavior depending on varying inputs.

Utilizing this approach allows us to uncover phenomena of intuition-like behavior in LLMs. While CRT tasks as well as semantic illusions are deliberately designed to elicit erroneous responses, it is probable that similar errors will arise when engaging LLMs in other situations. These findings imply that LLMs are prone to generating subtle inaccuracies that are challenging to identify since they seamlessly blend into the conversation without exhibiting semantic or logical inconsistencies upon initial examination. However, our experiments also uncover the vanishing of intuition-like behavior and a high resistance to the respective reasoning biases in ChatGPT as well as GPT-4. The difficulty level of CRT tasks and semantic illusions are specifically aligned to “dupe” a significant proportion of the human population. However, while the average human cognitive capacity remains relatively stable,⁶⁹ LLMs underwent drastic changes in various dimensions regarding their reasoning abilities. Our results reflect this trend. ChatGPT and GPT-4, dissimilar to earlier models, clearly outperform humans and are no longer “duped” by the tests.

References

1. Wei, J. *et al.* Emergent Abilities of Large Language Models. *arXiv*, 1–16 (2022).
2. Brown, T. B. *et al.* Language Models are Few-Shot Learners. *arXiv*, 1–75 (2020).
3. Kim, G., Baldi, P. & McAleer, S. Language Models can Solve Computer Tasks. *arXiv*, 1–26 (2023).
4. Nair, V., Schumacher, E., Tso, G. & Kannan, A. DERA: Enhancing Large Language Model Completions with Dialog-Enabled Resolving Agents. *arXiv* (2023).
5. Wei, J. *et al.* Chain of Thought Prompting Elicits Reasoning in Large Language Models. *arXiv*, 1–41 (2022).
6. Holtermann, B. & van Deemter, K. Does ChatGPT have Theory of Mind? *arXiv*, 1–15 (2023).
7. Tversky, A. & Kahneman, D. Judgment under Uncertainty. Heuristics and Biases. *Science* 185, 1124–1131 (1974).
8. OpenAI. *GPT-4 Technical Report*, 1–39. Available at <https://cdn.openai.com/papers/gpt-4.pdf> (2023).
9. Frederick, S. Cognitive Reflection and Decision Making. *Journal of Economic Perspectives* 19, 25–42 (2005).
10. Toplak, M. E., West, R. F. & Stanovich, K. E. The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & cognition* 39, 1275–1289 (2011).

11. Stanovich, K. E. & West, R. F. Individual differences in reasoning: implications for the rationality debate? *Behav Brain Sci* 23, 645-726 (2000).
12. Hodgkinson, G. P., Langan-Fox, J. & Sadler-Smith, E. Intuition: a fundamental bridging construct in the behavioural sciences. *British Journal of Psychology* 99, 1-27 (2008).
13. Epstein, S. Integration of the cognitive and the psychodynamic unconscious. *The American Psychologist* 49, 709-724 (1994).
14. Sloman, S. A. The empirical case for two systems of reasoning. *Psychological bulletin* 119, 3-22 (1996).
15. Kahneman, D. & Frederick, S. Representativeness Revisited: Attribute Substitution in Intuitive Judgment. In *Heuristics and Biases*, edited by T. Gilovich, D. Griffin & D. Kahneman (Cambridge University Press, Cambridge, 2012), pp. 49-81.
16. Evans, J. S. B. T. In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences* 7, 454-459 (2003).
17. Bago, B. & Neys, W. de. The Smart System 1: evidence for the intuitive nature of correct responding on the bat-and-ball problem. *Thinking & Reasoning* 25, 1-43 (2019).
18. Neys, W. de. Bias and Conflict: A Case for Logical Intuitions. *Perspectives on Psychological Science* 7, 28-38 (2012).
19. Neys, W. de. Advancing theorizing about fast-and-slow thinking. *The Behavioral and Brain Sciences*, 1-68 (2022).
20. Sirota, M., Dewberry, C., Juanchich, M., Valuš, L. & Marshall, A. C. Measuring cognitive reflection without maths: Development and validation of the verbal cognitive reflection test. *Journal of Behavioral Decision Making* 34, 322-343 (2021).
21. Erickson, T. D. & Mattson, M. E. From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior* 20, 540-551 (1981).
22. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. *Improving Language Understanding by Generative Pre-Training*, 1-12 (2018).
23. Radford, A. *et al.* Language Models are Unsupervised Multitask Learners, 1-24 (2019).
24. Black, S., Leo, G., Wang, P., Leahy, C. & Biderman, S. *GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow* (Zenodo, 2021).
25. Zhang, S. *et al.* OPT: Open Pre-trained Transformer Language Models. *arXiv*, 1-30 (2022).
26. LAION-AI. Open-Assistant. Available at <https://github.com/LAION-AI/Open-Assistant>.

27. Le Scao, T. *et al.* BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arXiv* (2023).
28. Touvron, H. *et al.* LLaMA: Open and Efficient Foundation Language Models. *arXiv*, 1–27 (2023).
29. Hagendorff, T., Fabi, S. & Kosinski, M. Machine intuition: Uncovering human-like intuitive decision-making in GPT-3.5. *arXiv*, 1–19 (2022).
30. Emami, A., Trischler, A., Suleman, K. & Cheung, J. C. K. An Analysis of Dataset Overlap on Winograd-Style Tasks. *arXiv*, 1–11 (2020).
31. Hagendorff, T. Machine Psychology: Investigating Emergent Capabilities and Behavior in Large Language Models Using Psychological Methods. *arXiv*, 1–15 (2023).
32. Binz, M. & Schulz, E. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences (PNAS)* 120, 1–10 (2023).
33. Dasgupta, I. *et al.* Language models show human-like content effects on reasoning. *arXiv*, 1–36 (2022).
34. Jones, E. & Steinhardt, J. Capturing Failures of Large Language Models via Human Cognitive Biases. *arXiv*, 1–15 (2022).
35. Nye, M., Tessler, M. H., Tenenbaum, J. B. & Lake, B. M. Improving Coherence and Consistency in Neural Sequence Models with Dual-System, Neuro-Symbolic Reasoning. *arXiv*, 1–13 (2021).
36. Chen, Y., Andiappan, M., Jenkin, T. & Ovchinnikov, A. A Manager and an AI Walk into a Bar: Does ChatGPT Make Biased Decisions Like We Do? *SSRN Journal*, 1–30 (2023).
37. Wilson, T. D. & Brekke, N. Mental contamination and mental correction: unwanted influences on judgments and evaluations. *Psychological bulletin* 116, 117–142 (1994).
38. Li, Y. *et al.* On the Advance of Making Language Models Better Reasoners. *arXiv*, 1–13 (2022).
39. Kadavath, S. *et al.* Language Models (Mostly) Know What They Know. *arXiv*, 1–42 (2022).
40. Speckmann, F. & Unkelbach, C. Moses, money, and multiple-choice: The Moses illusion in a multiple-choice format with high incentives. *Memory & cognition* 49, 843–862 (2021).
41. Sirota, M. & Juanchich, M. Effect of response format on cognitive reflection: Validating a two- and four-option multiple choice question version of the Cognitive Reflection Test. *Behavior Research Methods* 50, 2511–2522 (2018).
42. Reynolds, L. & McDonell, K. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. *arXiv*, 1–10 (2021).

43. Robertson, Z. You Can Probably Amplify GPT3 Directly. Available at <https://www.lesswrong.com/posts/Mzrs4MSi58ujBLbBG/you-can-probably-amplify-gpt3-directly> (2020).
44. Bialek, M. What Color are the Lilies? Forced Reflection Boosts Performance in the Cognitive Reflection Test. *SSRN Journal*, 1–19 (2016).
45. Kojima, T., Gu, S. S., Reid, M., Matsuo, Y. & Iwasawa, Y. Large Language Models are Zero-Shot Reasoners. *arXiv*, 1–36 (2022).
46. Stieger, S. & Reips, U.-D. A limitation of the Cognitive Reflection Test: familiarity. *PeerJ* 4, 1-12 (2016).
47. Haigh, M. Has the Standard Cognitive Reflection Test Become a Victim of Its Own Success? *Advances in cognitive psychology* 12, 145–149 (2016).
48. Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A. & Hamilton, J. The Development and Testing of a New Version of the Cognitive Reflection Test Applying Item Response Theory (IRT). *Journal of Behavioral Decision Making* 29, 453–469 (2016).
49. Campitelli, G. & Gerrans, P. Does the cognitive reflection test measure cognitive reflection? A mathematical modeling approach. *Memory & cognition* 42, 434–447 (2014).
50. Mosleh, M., Pennycook, G., Arechar, A. A. & Rand, D. G. Cognitive reflection correlates with behavior on Twitter. *Nature Communications* 12, 1–10 (2021).
51. Stanovich, K. E. Miserliness in human cognition: the interaction of detection, override and mindware. *Thinking & Reasoning* 24, 423–444 (2018).
52. McCoy, T., Pavlick, E. & Linzen, T. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, edited by A. Korhonen, D. Traum & L. Màrquez (Association for Computational Linguistics, Stroudsburg, PA, USA, 2019), pp. 3428–3448.
53. Neys, W. de, Rossi, S. & Houdé, O. Bats, balls, and substitution sensitivity: cognitive misers are no happy fools. *Psychonomic bulletin & review* 20, 269–273 (2013).
54. Matheson, H. E. & Barsalou, L. W. Embodiment and Grounding in Cognitive Neuroscience. In *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*, edited by J. T. Wixted (John Wiley & Sons, Inc, Hoboken, NJ, USA, 2018), pp. 1–27.
55. Srivastava, A. *et al.* Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *arXiv*, 1–100 (2022).
56. Qiao, S. *et al.* Reasoning with Language Model Prompting: A Survey. *arXiv*, 1–20 (2022).

57. Geirhos, R. *et al.* Shortcut Learning in Deep Neural Networks. *Nature Machine Intelligence* 2, 665–673 (2020).
58. Christiano, P. *et al.* Deep reinforcement learning from human preferences. *arXiv*, 1–17 (2017).
59. Ouyang, L. *et al.* Training language models to follow instructions with human feedback. *arXiv*, 1–68 (2022).
60. Todd, P. M. & Gigerenzer, G. *Ecological Rationality. Intelligence in the World* (Oxford University Press, Oxford, 2012).
61. Gigerenzer, G. On Narrow Norms and Vague Heuristics. A Reply to Kahneman and Tversky (1996). *Psychological review* 103, 592–596 (1996).
62. Gigerenzer, G. & Gaissmaier, W. Heuristic decision making. *Annual Review of Psychology* 62, 451–482 (2011).
63. Castellucci, G. A., Kovach, C. K., Howard, M. A., Greenlee, J. D. W. & Long, M. A. A speech planning network for interactive language use. *Nature* 602, 117–122 (2022).
64. Ji, Z. *et al.* Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55, 1–38 (2023).
65. Weidinger, L. *et al.* Ethical and social risks of harm from Language Models. *arXiv*, 1–64 (2021).
66. Vaswani, A. *et al.* Attention Is All You Need. *arXiv*, 1–15 (2017).
67. Chalmers, D. J. *The conscious mind. In search of a fundamental theory* (Oxford University Press, New York, 2007).
68. Rahwan, I. *et al.* Machine behaviour. *Nature* 568, 477–486 (2019).
69. Dutton, E., van der Linden, D. & Lynn, R. The negative Flynn Effect: A systematic literature review. *Intelligence* 59, 1–7 (2016).
70. Zhao, T. Z., Wallace, E., Feng, S., Klein, D. & Singh, S. Calibrate Before Use: Improving Few-Shot Performance of Language Models. *arXiv*, 1–15 (2021).
71. Renkl, A. Toward an instructionally oriented theory of example-based learning. *Cognitive Science* 38, 1–37 (2014).
72. Meyer, A., Zhou, E. & Frederick, S. The non-effects of repeated exposure to the cognitive reflection test. *Judgment and Decision Making* 13, 246–259 (2018).
73. Bialek, M. & Pennycook, G. The cognitive reflection test is robust to multiple exposures. *Behavior Research Methods* 50, 1953–1959 (2018).

Appendix A

Table A.1 – List of all CRT tasks that we use for the experiments.

Type 1	Type 2	Type 3
A pair of leggings and a top cost \$35. The pair of leggings cost \$20 more than the top. How much does the top cost?	If it takes 4 people 4 hours to sew 4 jackets, how long would it take 8 people to sew 8 jackets?	On a loaf of bread, there is a patch of mold. Every hour, the patch doubles in size. If it takes 16 hours for the patch to cover the entire loaf of bread, how long would it take for the patch to cover half of the bread?
A speaker and a CD cost \$210 together. The speaker costs \$200 more than the CD. How much does the CD cost?	If it takes 1 car 5 hours to drive from point A to point B, how long does it take 2 cars?	In a forest, 1 square meter is on fire. Each day, the fire doubles in size. If it takes 18 days for the fire to spread across the whole forest, how long does it take for the fire to cover half of the forest?
A jacket and a pair of jeans cost \$70. The jacket costs \$40 more than the pair of jeans. How much do the jeans cost?	If it takes 1 washing machine 2 hours to wash 1 load of laundry. How long does it take 2 washing machines to wash 2 loads of laundry?	In a cave, there is a group of bats. Every day, the group doubles in size. If it takes 64 days for the group to cover the entire cave, how long would it take for the group to cover half of the cave?
A pot and a pan cost \$60. The pot costs \$24 more than the pan. How much does the pan cost?	If it takes 3 bees 3 minutes to pollinate 3 flowers, how long does it take 100 bees to pollinate 100 flowers?	In a garden, a patch of grass doubles in size every single day. It covers the entire garden in 12 days. How long would it take for the patch of grass to cover half of the garden?
A TV and a cocktail glass cost \$380. The TV costs \$360 more than the cocktail glass. How much does the cocktail glass cost?	If it takes 5 carpenters 5 days to make 5 chairs, how long would it take 50 carpenters to make 50 chairs?	In a country, the population is fleeing from war. Every day the number of refugees is doubling in size. It takes 22 days for the whole population to flee. How many days does it take for half of the population to flee?
A flower box and a pair of rollerblades cost \$36. The pair of rollerblades costs \$26 more than the flower box. How much does the flower box cost?	If it takes 1 oven 30 minutes to bake 1 lasagna, how long does it take 2 ovens to bake 2 lasagnas?	In a state, there is a pandemic. Every day the number of contagious people doubles. If it takes 30 days until all people are infected, how long does it take for half of the people to become infected?
A set of crayons and a football cost \$12. The football costs \$8 more than the crayons. How much do the crayons cost?	If it takes 1 researcher 1 year to publish 1 paper, how long does it take 2 researchers to publish 2 papers?	A person has 1 bitcoin. Every day, the number of bitcoins that person owns doubles. If it takes 30 days to reach the limit of bitcoins, how long would it take to reach half the limit?
A bottle of bath salts and a candle cost \$5. The bottle of bath salts costs \$3 more than the candle. How much does the candle cost?	If it takes 2 cleaners 2 hours to clean 2 rooms, how long does it take 5 cleaners to clean 5 rooms?	In a creek, fish are living. Every week, the number of fish doubles in size. If it takes 24 weeks to fill the entire creek with fish, how long would it take for the fish to fill half of the creek?
A coffee machine and a shelf cost \$120. The coffee machine costs \$30 more than the shelf. How much does the shelf cost?	If it takes 2 students to change 2 light bulbs in 2 minutes, how long does it take 4 students to change 4 light bulbs?	Over a city, there is a cloud of dust. Every day, the cloud doubles in size. If it takes 12 days for the cloud to cover the entire city, how long would it take for the cloud to cover half of the city?
A shaver and a bottle of shampoo cost \$17. The shaver costs \$13 more than the shampoo. How much does the shampoo cost?	If it takes 10 builders to build 10 houses in 10 weeks, how long would it take 100 builders to build 100 houses?	
A rice cooker and a food processor cost \$200. The food processor costs \$40 more than the rice cooker. How much does the rice cooker cost?	If it takes 3 people 30 minutes to plant 3 trees, how long would it take 9 people to plant 9 trees?	
A tube of toothpaste and an electric toothbrush cost \$24. The toothbrush costs \$20 more than the toothpaste. How much does the toothpaste cost?	If it takes 13 coffee machines 13 minutes to make 13 coffees, how long would it take 26 coffee machines to make 26 coffees?	
	If it takes 4 machines 4 minutes to pack 4 boxes of chocolates, how long would it take 16 machines to pack 16 boxes of chocolates?	

A small suitcase and a backpack cost \$35. The backpack costs \$5 more than the suitcase. How much does the suitcase cost?	If it takes 2 children 2 minutes to eat 2 boxes of chocolates, how long would it take 20 children to eat 20 boxes of chocolates?	In a class, there is an ill student. Every day, the number of ill students doubles. If it takes 6 days until the whole class is ill, how long will it take until half of the class is ill?
A grill and a tent cost \$250. The tent costs \$100 more than the grill. How much does the grill cost?	If it takes 2 people 2 weeks to read 2 books, how long would it take 200 people to read 200 books?	There is a virus in a city. Every day, the number of infected citizens triples. If it takes 60 days for the whole city to be infected, how long would it take for the virus to infect a third of the city's population?
A bike and a trampoline cost \$1,000. The bike costs \$250 more than the trampoline. How much does the trampoline cost?	If it takes 5 teams 5 weeks to renovate 5 houses, how long would it take 25 teams to renovate 25 houses?	In a fish tank, there are some algae. Every day, the number of algae triples. If it takes 30 days until all of the fish tank is covered in the algae, how long would it take for the algae to cover a third of the fish tank?
A set of bed linen and a mixing bowl cost \$27. The bed linen costs \$7 more than the mixing bowl. How much does the mixing bowl cost?	If it takes 3 people 3 weeks to knit 3 pairs of socks, how long would it take 33 people to knit 33 pairs of socks?	On a yogurt, there is a patch of bacteria. Every day, the patch doubles in size. If it takes the patch 4 days to cover the whole yogurt, how long will it take until it covers half of the yogurt?
A lawnmower and a wall clock cost \$450. The lawnmower costs \$300 more than the wall clock. How much does the wall clock cost?	If it takes 10 people 10 hours to pick 10 fields of strawberries, how long would it take 20 people to pick 20 fields of strawberries?	A patch of moss is growing on a rock. Every day, the patch doubles in size. If it takes the patch 264 days to cover the entire rock, how long would it take for the patch to cover half of the rock?
A pair of socks and a pair of hiking boots cost \$94. The hiking boots cost \$86 more than the pair of socks. How much do the socks cost?	If it takes 10 programmers 1 hour to write 10 lines of code, how long would it take 100 programmers to write 100 lines of code?	There is an old lady who feeds two cats; each gets its feeding dish. Every day, each cat brings a new cat to be fed by the old lady. If it takes 32 days until the cats have achieved the maximum of feeding dishes the old lady owns, how long would it take for the cats to use half of the feeding dishes?
A camera and a flashlight cost \$70. The camera costs \$46 more than the flashlight. How much does the flashlight cost?	If it takes 2 photographers 2 hours to take 200 photos, how long would it take 6 photographers to take 600 photos?	Under a tree, there is a pile of leaves. Every week, the size of the pile doubles. If it takes 4 weeks until all leaves are on the pile, how long would it take until half of the leaves are on the pile?
A pack of diapers and a glass of baby food cost \$10. The diapers cost \$6 more than the glass of baby food. How much does the baby food cost?	If it takes 1 painter 1 hour to paint 1 painting, how long would it take 7 painters to paint 7 paintings?	In a container, there is growing a patch of mushrooms. Every day, the patch of mushrooms triples in size. If it takes 12 days until the container is full of mushrooms, how long would it take the mushrooms to fill one-third of the container?
A smartphone and a silicone case cost \$310. The smartphone costs \$290 more than the silicone case. How much does the silicone case cost?	If it takes 5 writers 5 months to write 5 books, how long would it take 10 writers to write 10 books?	There is a man who breeds rabbits. Every year, the number of rabbits doubles. If it takes 8 years until all the cages are occupied, how long would it take for the rabbits to occupy half of the cages?
A laundry detergent and fabric softener cost \$20. The laundry detergent costs \$10 more than the fabric softener. How much does the fabric softener cost?	If it takes 3 cooks 30 minutes to cook 3 meals, how long would it take 6 cooks to cook 6 meals?	In a forest, ramson is growing. Every week, the patch of ramson doubles in size. If it takes 10 weeks for the patch to cover the whole
A bottle of wine and a set of wine glasses cost \$22. The wine bottle costs \$8 more than the wine glasses. How much do the wine glasses cost?	If it takes 3 doctors 30 minutes to examine 3 patients, how long would it take 6 doctors to examine 6 patients in sum?	
A lamp and a light bulb cost \$22. The lamp costs \$18 more than the light bulb. How much does the light bulb cost?	If it takes 2 drivers 20 minutes to change 20 tires, how long would it take 8 drivers to change 80 tires?	
An office desk and a chair cost \$420. The desk costs \$80 more than the chair. How much does the chair cost?	If it takes 10 farm workers 10 seconds to pick 10 apples, how long would it take 20 farm workers to pick 20 apples?	
A knife and a cutting board cost \$45. The knife costs \$15 more than the cutting board. How much does the cutting board cost?	If it takes 1 freezer 1 hour to freeze 1 liter of water, how long would it take 5 freezers to freeze 5 liters of water?	
	If it takes 5 bakers 5 hours to bake 5 cakes, how long would it take 20 bakers to bake 20 cakes?	

A toy and batteries cost \$29. The toy costs \$21 more than the batteries. How much do the batteries cost?	If it takes 1 hair stylist 30 minutes to finish 1 hairstyle, how long would it take 4 hair stylists to finish 4 hairstyles in sum?	forest, how long would it take for the patch to cover half of the forest?
A coat and a scarf cost \$120. The coat costs \$80 more than the scarf. How much does the scarf cost?	If it takes 2 mechanics 2 hours to fix 2 cars, how long would it take 10 mechanics to fix 10 cars?	There is an apple tree. Every day, the number of apples that fall from that tree doubles. If it takes 16 days for all the apples from the tree to fall, how long would it take for half of them to fall?
Colored pencils and a sharpener cost \$18. The pencils cost \$12 more than the sharpener. How much does the sharpener cost?	If it takes 1 tailor 10 hours to make 1 dress, how long would it take 5 tailors to make 5 dresses?	There is a bamboo in the garden. Every day, the size of the surface that the roots cover doubles. If it takes 30 days for the roots to grow to 5 square meters, how long would it take to reach half of that surface?
A rug and a coffee table cost \$170. The rug costs \$70 more than the coffee table. How much does the coffee table cost?	If it takes 2 painters 2 hours to paint 2 rooms, how long would it take 16 painters to paint 16 rooms?	It is raining and the lake is filling with water. Every day, the amount of water in the lake doubles. If it takes 20 days for the whole lake to fill with water, how long would it take to fill half of the lake?
A TV and an antenna cost \$295. The TV costs \$265 more than the antenna. How much does the antenna cost?	If it takes 1 tree 1 day to grow 1 leaf, how long would it take 9 trees to grow 9 leaves in sum?	There is a tree in a forest growing. Every year the height of the tree doubles. If it takes 10 years for the tree to reach its full height, how long would it take for the tree to reach half of its height?
A box of cigarettes and a lighter cost \$11. The cigarettes cost \$7 more than the lighter. How much does the lighter cost?	If it takes 6 runners 60 minutes to clean 6 shoes, how long would it take 12 runners to clean 12 shoes in sum?	There is a flood on a field. Every hour, the size of the flooded area doubles. If it takes 20 hours for the whole field to be flooded, how long would it take for half of the lake to be flooded?
An apple and a potato cost \$10. The potato cost \$6 more than the apple. How much does the apple cost?	If it takes 1 translator 1 hour to translate 1 page, how long would it take 4 translators to translate 4 pages in sum?	There is a patient diagnosed with cancer. Every week, the number of cancer cells triples. If it takes 33 weeks for the number of cancer cells to cause death, how long would it take them to reach a third of this level?
A mouse and a keyboard cost \$43. The mouse costs \$7 more than the keyboard. How much does the keyboard cost?	If it takes 2 machines 2 hours to make 2 smartphones, how long would it take 14 machines to make 14 smartphones?	A fish species is migrating. Every day, the length of distance it passes doubles. If it takes 18 days for the fish species to reach its goal, how long would it take to pass half of the distance?
A car and tires cost \$3100. The car costs \$1900 more than the tires. How much do the tires cost?	If it takes 5 people 5 minutes to smoke 5 cigarettes, how long would it take 20 people to smoke 20 cigarettes in sum?	There is a tree with branches falling off. Every second, the speed of the falling branch triples. If it takes 6 seconds for the branch to reach the ground, how long would it take for the branch to reach a third of the distance?
A bike and a helmet cost \$350. The bike costs \$250 more than the helmet. How much does the helmet cost?	If it takes 10 opticians 10 days to make 10 glasses, how long would it take 100 opticians to make 100 glasses?	A fly is flying from point A to point B. Every hour, the distance the fly passes doubles. If it takes the fly 12 hours to reach point B, how long would it take the fly to pass half the distance?
Mascara and lipstick cost \$60. The lipstick costs \$14 more than the mascara. How many dollars does the mascara cost?	If it takes 1 barrel 1 hour to fill 100 liters, how long would it take 5 barrels to fill 500 liters in sum?	There is a farmer plowing a field. Every hour, the size of the surface he finishes plowing doubles. If it takes 10 hours for the farmer to
A box of tea and cookies cost \$18. The box of tea costs \$8 more than the cookies. How much do the cookies cost?	If it takes 1 person 10 minutes to eat 1 pizza, how long would it take 5 people to eat 5 pizzas in sum?	
A fridge and a microwave oven cost \$410. The fridge costs \$170 more than the microwave. How much does the microwave cost?	If it takes 1 person 1 hour to boil 100 liters of water, how long would it take 4 people to boil 400 liters of water?	
A bottle of ouzo and a bottle of wine cost \$165. The ouzo costs \$25 more than the wine. How much does the wine cost?	If it takes 1 air conditioner 20 minutes to cool 1 room, how long would it take 8 air conditioners to cool 8 rooms?	
Sunglasses and a hat cost \$165. The sunglasses cost \$115 more than the hat. How much does the hat cost?	If it takes 3 students 30 minutes to finish 3 exams, how long would it take 6 students to finish 6 exams?	

A ring and a pair of earrings cost \$760. The ring costs \$340 more than the earrings. How much do the earrings cost?	If it takes 1 woman 30 minutes to make 1 pie, how long would it take 3 women to make 3 pies in sum?	finish plowing the whole field, how long would it take him to plow half the field?
A blanket and pillows cost \$125. The blanket costs \$65 more than the pillows. How much do the pillows cost?	If it takes 2 baristas 2 minutes to make 2 coffees, how long would it take 23 baristas to make 23 coffees?	There is a basement full of whiskey barrels and the barrels are being filled with whiskey. Every minute, the amount of whiskey filling the barrels doubled. If it takes 12 minutes for the whole barrels to be filled, how long would it take for half the barrels to be filled?
The electricity and water bill cost \$420. The electricity bill costs \$120 more than the water bill. How much does the water bill cost?	If it takes 1 person 10 minutes to cook 1 pack of spaghetti, how long would it take 5 persons to cook 5 packs of spaghetti?	There is a room filled with rats. Every month, the number of rats triples. If it takes 9 months for the rats to fill the whole room, how long would it take them to fill a third of the room?
A perfume and a body lotion cost \$440. The perfume costs \$300 more than the body lotion. How much does the body lotion cost?	If it takes 2 people 2 days to renovate 2 bathrooms, how long would it take 12 people to renovate 12 bathrooms?	There is a pile of fruit rotting. Every day, the percentage of alcohol doubles. If it takes 40 days for the fruit to reach 12% of alcohol, how long would it take to reach 6%?
A painting and a sculpture cost \$830. The painting costs \$190 more than the sculpture. How much does the sculpture cost?	If it takes 1 fish 1 year to grow 50 cm, how long would it take 3 fish to grow 150 cm in sum?	A surface is filled with wet concrete that is drying. Every day, the speed of the concrete drying doubles. If it takes 4 days for the concrete to be fully dry, how long would it take the concrete to be half-dry?
A purse and a wallet cost \$210. The purse costs \$140 more than the wallet. How much does the wallet cost?	If it takes 1 person 1 hour to make 10 phone calls, how long would it take 10 people to make 100 phone calls?	It is winter and a lake is freezing. Every day, the thickness of the ice doubles. If it takes 10 days for the ice to be 1m thick, how long would it take for the ice to grow half of that thickness?
A book and a bag cost \$37. The bag costs \$13 more than the book. How much does the book cost?	If it takes 5 printers 5 minutes to print 5 documents, how long would it take 15 printers to print 15 documents?	There is a freezer full of fresh room-temperature food. Every hour, the speed of freezing doubles. If it takes 16 hours for the food to be fully frozen, how long would it take for the food to be half-frozen?
A laptop and a printer cost \$580. The laptop costs \$280 more than the printer. How much does the printer cost?		Programmers are writing code for new software. Every month, the number of lines in the code doubles. If it takes the programmers 10 months to finish the code, how long would it take them to write half of the lines of the code?
A pizza and a beer cost \$23. The pizza costs \$11 more than the beer. How much does the beer cost?		A factory is filling bags with chocolate cookies. Every hour, the number of packed bags is doubling. If it takes 8 hours for all the bags to be packed, how long would it take for half the bags to be packed?
		It is winter and it's snowing. Every 2 hours, the height of fallen snow doubles. If it takes 20 hours for the snow to reach 1m in height, how long would it take for the snow to reach 50cm?
		A team of 5 people is building a house. Every week, the progress of the house doubles. If it takes 8 weeks for them to finish building the

		<p>house, how long would it take them to build half of the house?</p> <p>2 painters are painting a house. Every hour, the surface they paint doubles. If it takes the painters 16 hours to finish painting the whole house, how long would it take them to paint half of the house?</p> <p>Wood is burning in a fireplace. Every 5 minutes the temperature of the fire doubles. If it takes 20 minutes for the fire to reach 250 degrees, how long would it take the fire to reach half of the temperature?</p> <p>A grandma is knitting a scarf for her grandson. Every week, the length of the scarf triples. If it takes 6 weeks for her to finish the scarf, how long would it take her to finish a third of the scarf?</p> <p>Oranges grow on a tree. Every month, the amount of oranges doubles. It takes 6 months for the tree to be full of oranges, how long would it take for half of the tree to be full of oranges?</p> <p>An iceberg on the North Pole is melting. Every year, the speed of melting is doubling. If it takes 10 years for the iceberg to be completely melted, how long would it take for half of the iceberg to melt?</p> <p>A girl is trying to grow her hair out long. Every year the length of her hair doubles. If it takes 6 years for the hair to grow to the ground, how long would it take for the hair to reach half of the length?</p> <p>A forest is growing in size. Every year the size of the forest doubles. If it takes 140 years to fill the entire available area, how long would it take to cover half of the area?</p> <p>A gas bottle is leaking. Every hour the amount of leaked gas has tripled. If it takes 33 hours for the entire gas to leak, how long would it take for a third of the gas to leak?</p> <p>Ants are climbing on a cake. Every minute, the number of ants covering the cake triples. If it takes 30 minutes for the ants to cover the entire cake, how long would it take them to cover a third of the cake?</p> <p>Smokers are smoking cigarettes in a closed room. Every minute, the amount of smoke in the room triples. If it takes them 9 minutes to fill the entire room with smoke, how long</p>
--	--	----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

		<p>would it take them to fill a third of the room with smoke?</p> <p>A pot of water is boiling on the stove. Every hour, the amount of vaporized water triples. If it takes 6 hours for the entire pot of water to vaporize, how long would it take for a third of the water to vaporize?</p>
--	--	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Table A.2 – List of all semantic illusions that we use for the experiments.

Semantic illusion
What color is Getafix's fur, the dog of Asterix and Obelix?
Who is the dictator of South Korea?
In which year did Germany win World War II?
Which political position did Adolf Hitler gain under President Otto von Bismarck?
Which object does Audrey Hepburn use to fly at the beginning of the movie "Mary Poppins"?
What is Mars, a famous candy bar named after a Greek god, made of?
Who is the architect of the famous Eiffel Tower in Marseille?
How long did Rapunzel fall asleep, after poking her finger on a spindle?
What is the name of the New Year festival celebrated on the 31st of January?
How many doors does an Advent wreath have?
Which bird symbolizes and accompanies Minerva, the Greek goddess of wisdom?
Which war is thematized in Dostoevsky's "War and Peace"?
Two nuclear bombs named Fat Boy and Little Man were dropped on which two Japanese cities in World War II?
In which year did West Germany aka the German Democratic Republic fall?
Which country did Columbus mistakenly believe he had discovered after sailing across the Pacific?
In which year did the Americans liberate the Auschwitz concentration camp in Poland in World War II?
Which country did the soccer player Diego Maradona, aka the Silver Boy, represent?
In which museum is Michelangelo's most famous painting, the Mona Lisa, exhibited?
How many blue stars are there on the American flag?
In which year did the American astronaut Louis Armstrong walk on the Moon?
What do bees produce after they collect wax from flowers?
Which species went extinct after Halley's Comet hit the Earth?
What is the process called when plants produce CO2 using chlorophyll?
Which country is New Mexico the capital of?
By whom was Galileo Galilei almost executed in the 17th century for claiming that the Earth is round?
In which country did Martin Luther King spark the Protestant Reformation in 1517?
What was the name of the mountain on which Noah received the Ten Commandments?
When did Jeff Bezos found SpaceX?
How many years after 9/11 did George W. Bush invade Iran?

What is the famous allegory of the cave by Aristotle about?
When do Jewish people celebrate Ramadan?
What is the song "Angie" by The Beatles about?
For how long has Sydney been the capital of Australia?
When did Marx start the communist revolution in the Russian Republic?
Where on their bodies do whales have their gills?
In which months do ostriches fly to warmer regions?
Why did Pablo Picasso cut off his ear?
Why did Pope Francis resign from his papacy?
How many dwarves befriended Cinderella?
What British actor played the role of the famous FBI agent James Bond?
Which Pakistani city is home to the Taj Mahal?
What was the name of the goat that was the first mammal to be cloned?
How many parts did the Titanic split into before it sank into the Pacific?
The Statue of Liberty is holding a sword in her left hand, and which item is in her right hand?
How many points does the Star of Moses, or the Jewish star have?
Who killed Jesus according to the Old Testament?
In the Marvel Universe, what is the name of Batman's sidekick?
What is the name of J.K. Rowling's famous fictional character that has a thunder-shaped scar on his forehead?
What color are the polar bears in Antarctica?
To which genre are the films with Oliver Laurel and Stan Hardy assigned?

Table A.3 – Reformulated semantic illusions to test the necessary world knowledge to spot untrue assumptions as well as example answers by GPT-3 (text-davinci-003).

Prompt		GPT-3's response
Semantic illusion	Knowledge question	
Question: By whom was Galileo Galilei almost executed in the 17th century for claiming that the Earth is round? Answer:	Question: Why was Galileo Galilei almost executed? Answer:	Galileo Galilei was almost executed for his advocacy of heliocentrism, which was seen as heresy by the Catholic Church.
Question: What is the famous allegory of the cave by Aristotle about? Answer:	Question: Who wrote the famous allegory of the cave? Answer:	Plato wrote the famous allegory of the cave.
Question: What is the name of J.K. Rowling's famous fictional character that has a thunder-shaped scar on his forehead? Answer:	Question: What shape does Harry Potter's scar have? Answer:	Harry Potter's scar is in the shape of a lightning bolt.

Table A.4 – Examples of multiple-choice prompts comprising the intuitive option and the corresponding GPT-3 (text-davinci-003) outputs.

	Multiple-choice prompt	GPT-3’s response
CRT type 1	Question: A pair of leggings and a top cost \$35. The pair of leggings cost \$20 more than the top. How much does the top cost? Choices: (A) \$15 (B) \$7.50 (C) \$10 Answer:	(A) \$15
CRT type 2	Question: If it takes 4 people 4 hours to sew 4 jackets, how long would it take 8 people to sew 8 jackets? Choices: (A) 2 hours (B) 8 hours (C) 4 hours Answer:	(B) 8 hours
CRT type 3	Question: A patch of moss is growing on a rock. Every day, the patch doubles in size. If it takes the patch 264 days to cover the entire rock, how long would it take for the patch to cover half of the rock? Choices: (A) 263 days (B) 164 days (C) 132 days Answer:	(C) 132 days
Semantic illusion	Question: In which museum is Michelangelo’s most famous painting, the Mona Lisa, exhibited? Choices: (A) The Palace of Versailles (B) The Louvre (C) The question contains an untrue assumption Answer:	(B) The Louvre

Appendix B

Each item in our 3 x 50 CRT dataset, as well as the 50 semantic illusions dataset, is used as a separate prompt in text-generating settings of multiple LLMs. Except for the chat-specific models of GPT (gpt-3.5-turbo and gpt-4), we add the string “Question:” as a prefix and the string “Answer:” as a suffix to all prompts since this is common practice in Q&A settings. In the chat-specific models, we use the default system message (“You are a helpful assistant.”). In all LLMs, we set the temperature parameter to 0 (or 0.0001) (i.e., return the most probable completion), since using the “best of” setting (i.e., calculating the probabilities of multiple completions and outputting the best) does not change or improve results significantly and has the downside of changing the outputs in each iteration due to the added randomness in the model. When analyzing the results of the CRT tasks, we use two steps: First, we use an automatic approach by comparing the generated texts with entries in our dataset, checking for the intuitive or correct term in the output sequence. If neither the intuitive nor correct term is present, the response is labeled as atypical. Second, to ensure high quality, we check all answers manually to spot remaining

mistakes. When analyzing the responses to the semantic illusions, we manually evaluate each output, sentence by sentence, until it can be categorized as correct (the LLM detects the flaw in the question), intuitive (the LLM answers as if the question made sense), or atypical. For the statistical analysis, we conducted two-sample binomial tests on the overall proportion of correct answers with $\alpha = .05$ or the corresponding Bonferroni α -level corrected for multiple testing.

Appendix C

Table D.1 – Relative frequencies of correct, intuitive, and atypical responses of each LLM used in this study when responding to CRT tasks as well as semantic illusions.

Type of LLM	correct	intuitive	atypical
GPT-4 CRT type 1	1	0	0
GPT-4 CRT type 2	0.98	0	0.02
GPT-4 CRT type 3	0.98	0	0.02
GPT-4 Semantic illusions	0.92	0.08	0
GPT4 overall	0.97	0.02	0.01
ChatGPT CRT type 1	1	0	0
ChatGPT CRT type 2	0.78	0.1	0.12
ChatGPT CRT type 3	0.56	0.06	0.38
ChatGPT Semantic illusions	0.84	0.16	0
ChatGPT overall	0.795	0.08	0.125
GPT-3 text-davinci-003 CRT type 1	0	1	0
GPT-3 text-davinci-003 CRT type 2	0.12	0.66	0.22
GPT-3 text-davinci-003 CRT type 3	0.04	0.94	0.02
GPT-3 text-davinci-003 Semantic illusions	0.26	0.72	0.02
GPT-3 text-davinci-003 overall	0.105	0.83	0.065
GPT-3 text-davinci-002 CRT type 1	0	0.98	0.02
GPT-3 text-davinci-002 CRT type 2	0.12	0.74	0.14
GPT-3 text-davinci-002 CRT type 3	0.04	0.94	0.02
GPT-3 text-davinci-002 Semantic illusions	0.1	0.86	0.04
GPT-3 text-davinci-002 overall	0.065	0.88	0.055
GPT-3 text-davinci-001 CRT type 1	0	0.9	0.1
GPT-3 text-davinci-001 CRT type 2	0.22	0.64	0.14
GPT-3 text-davinci-001 CRT type 3	0	0.94	0.06
GPT-3 text-davinci-001 Semantic illusions	0.06	0.82	0.12
GPT-3 text-davinci-001 overall	0.07	0.825	0.105
LlaMA 7B CRT type 1	0.02	0.88	0.1
LlaMA 7B CRT type 2	0.18	0.52	0.3
LlaMA 7B CRT type 3	0	0.38	0.62
LlaMA 7B Semantic Illusions	0.04	0.74	0.22
LlaMA 7B overall	0.06	0.63	0.31
BLOOM CRT type 1	0	0.52	0.48
BLOOM CRT type 2	0.46	0.54	0
BLOOM CRT type 3	0	0.42	0.58
BLOOM Semantic illusions	0.1	0.76	0.14
BLOOM overall	0.14	0.56	0.3
Open Assistant CRT type 1	0.12	0.36	0.52
Open Assistant CRT type 2	0.4	0.5	0.1
Open Assistant CRT type 3	0	0.3	0.7
Open Assistant Semantic illusions	0.04	0.7	0.26
Open Assistant overall	0.14	0.465	0.395
GPT-3 text-curie-001 CRT type 1	0.04	0.06	0.9
GPT-3 text-curie-001 CRT type 2	0.4	0.18	0.42
GPT-3 text-curie-001 CRT type 3	0	0.26	0.74
GPT-3 text-curie-001 Semantic illusions	0.02	0.84	0.14
GPT-3 text-curie-001 overall	0.115	0.335	0.55
OPT 6.7B CRT type 1	0.04	0.08	0.88
OPT 6.7B CRT type 2	0.48	0.3	0.22

OPT 6.7B CRT type 3	0	0.06	0.94
OPT 6.7B Semantic illusions	0.1	0.58	0.32
OPT 6.7B overall	0.155	0.255	0.59
GPT-Neo CRT type 1	0.02	0.04	0.94
GPT-Neo CRT type 2	0.56	0.18	0.26
GPT-Neo CRT type 3	0	0.02	0.98
GPT-Neo Semantic illusions	0.08	0.58	0.34
GPT-Neo overall	0.165	0.205	0.63
GPT2-XL CRT type 1	0	0.06	0.94
GPT2-XL CRT type 2	0.04	0.1	0.86
GPT2-XL CRT type 3	0	0	1
GPT2-XL Semantic illusions	0.06	0.4	0.54
GPT2-XL overall	0.025	0.14	0.835
GPT-1 CRT type 1	0.02	0.04	0.94
GPT-1 CRT type 2	0.02	0.08	0.9
GPT-1 CRT type 3	0	0	1
GPT-1 Semantic illusions	0	0	1
GPT-1 overall	0.01	0.03	0.96

Appendix D

The three methods to test how sturdy the intuitive decision-making was are (1) a multiple-choice framing of the tasks, (2) a suffix that can elicit deliberate reasoning, and (3) training examples.

Ad (1): One can assume that framing CRTs as well as semantic illusions in a multiple-choice format decreases intuitive responses and increases correct ones. This is due to backward strategies, where each possible answer can be used as a hint to reconstruct the correct response. However, experiments with human test subjects indicate that multiple-choice formats do not significantly increase the number of correct responses to the CRT.⁴¹ Even semantic illusions persist in humans when a multiple-choice format is used—with the correct answer option being “can’t say.”⁴⁰ On the other hand, recent studies on LLMs show that their general reasoning capabilities improve when provided with multiple-choice questions instead of open-ended questions.³⁹ When doing so, one has to keep in mind that GPT-3 has a recency bias,⁷⁰ meaning that the model tends to predict answers that are related to what is mentioned at the end of a prompt rather than at the beginning. We neutralize this effect by running the test 3! times, shuffling the answers to cover all possible combinations, and taking the mean. For an exemplary prompt design for the multiple-choice task, see Table A.4.

Ad (2): CRT tasks as well as semantic illusions are intentionally designed to trigger intuitive processes. However, one can provide antidotes against succumbing to this design, for instance by forcing numerical reflections or a deeper reflection on the tasks’ content in humans prior to responding to CRT tasks⁴⁴ or by warning humans against the illusionary character of tasks. Likewise, as shown in very recent research in GPT-3, the model’s performance on complex (math) tasks can be improved by a simple suffix such as “Let’s think step by step” that is added to each prompt.⁴⁵ GPT-3 is inept or even unable to perform several reasoning steps in one forward pass. However, problems that require serial thinking can be divided into sub-problems to be better solvable.⁴³ Adding a suffix such as “Let’s think step by step” provokes the language model to do exactly that, meaning to serialize reasoning processes. This way, multi-step reasoning (or, metaphorically speaking, System 2 thinking) is elicited in language models, which—for

instance in the case of GPT-3—are able to perform zero-shot chain-of-thought prompting.⁴² We also amplify this approach for the semantic illusions by using the suffix “Think carefully and check the question for potential untrue assumptions.” Additionally, the response length is set to 500 tokens to allow for a more comprehensive unfolding of the model’s stepwise response. When interpreting the results, we check whether the correct answer occurs in the last line of output. Additionally, we check each output manually to avoid any mistakes. Because of the suffix, we do not use the prefix (“Question:”) and suffix (“Answer:”) as in the other experiments for both the CRT and semantic illusions dataset.

Ad (3): Both humans and language models can learn from examples.^{2,71} Hence, when being repeatedly exposed to CRT tasks or semantic illusions, this should improve performance in both. Studies on humans find that when they have already performed CRT tasks in the past, their score is higher when being tested again.^{46,47} However, some studies also suggest that humans do not improve their test score due to multiple exposures.^{72,73} Despite these conflicting results in studies on humans, it remains undisputed that GPT-3 is able to engage in few-shot learning as soon as task demonstrations are included in the prompt. The model quickly learns new concepts from a few or even one example.² Therefore, we provide GPT-3 text-davinci-003 with 1, 2, 3, 4, and 5—as well as 10, 15, 20, 25, 30, 35, 40, and 45—training examples, which comprise randomly chosen tasks from our CRT or semantic illusions dataset combined with their correct solutions. Subsequently, we administer the actual task and measure the response.

To analyze the differences before and after the interventions (for (3), we focus on the performance after 45 training examples), we conducted two-sample binomial tests with $\alpha = .05$ on the overall proportion of correct answers. For follow-up tests specific to each of the four tasks, we utilized the Bonferroni-corrected $\alpha = .0125$.

Appendix E

To remove the error-eliciting phrases in the CRT tasks, we changed their wording by using the structure shown in table G.1. This changes the “hostile” nature of the tasks and results in 95% correct responses overall when prompting GPT-3 text-davinci-003 with the 3 x 50 simplified CRT tasks.

Table G.1 – Rephrasing CRT tasks to remove the error-eliciting phrases.

	original tasks	simplified tasks	correct response to simplified task
CRT type 1	A rice cooker and a food processor cost \$200. The food processor costs \$40 more than the rice cooker. How much does the rice cooker cost?	A rice cooker and a food processor cost \$200. The food processor costs \$40. How much does the rice cooker cost?	\$160
CRT type 2	If it takes 2 cleaners 2 hours to clean 2 rooms, how long does it take 5 cleaners to clean 5 rooms?	If it takes 2 cleaners 2 hours to clean 2 rooms, how long does it take 2 cleaners to clean 5 rooms?	5 hours

CRT type 3	A patch of moss is growing on a rock. Every day, the patch doubles in size. If it takes the patch 264 days to cover the entire rock, how long would it take for the patch to cover half of the rock?	A patch of moss is growing on a rock. If it takes the patch 264 days to cover the entire rock, how long would it take for the patch to cover half of the rock?	132 days
---------------	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	----------------------------------------------------------------------------------------------------------------------------------------------------------------	----------