Brief Report

Language Models Do Not Possess the Uniquely Human Cognitive Ability of Relational Abstraction

Bradley Monk *, Timothy Meyer , Mary Ngo

Posted Date: 23 January 2024

doi: 10.20944/preprints202401.1681.v1

Keywords: Language models; LLM; human cognition; relational abstraction; comparative cognition; artificial intelligence; AI; AGI; benchmarks; natural language processing, generative AI

Brief Report

Language Models Do Not Possess the Uniquely Human Cognitive Ability of Relational Abstraction

Bradley Monk, Timothy Meyer and Mary Ngo

Director of Human-AI Integration, Pacific Science & Engineering, 5880 Oberlin Drive, San Diego, CA 92121
* Correspondence: bmonk@ucsd.edu

Abstract: Humans have a unique cognitive ability known as relational abstraction, which allows us to identify logical rules and patterns beyond basic perceptual characteristics. This ability represents a key difference between how humans and other animals learn and interact with the world. With current large language models rivaling human intelligence in many regards, we investigated whether relational abstraction was an emergent chat completion capacity in various models. We find that despite their impressive language processing skills, all tested language models failed the relational match-to-sample (RMTS) test, a benchmark for assessing relational abstraction. These results challenge the assumption that advanced language skills inherently confer the capacity for complex relational reasoning. The paper highlights the need for a broader evaluation of AI cognitive abilities, emphasizing that language proficiency alone may not be indicative of certain higher-order cognitive processes thought to be supported by language.

Keywords: Language models; LLM; human cognition; relational abstraction; comparative cognition; artificial intelligence; AI; AGI; benchmarks; natural language processing; generative AI

Introduction

The cognitive differences between humans and nonhuman animals has long intrigued researchers. This long standing fascination with the cognitive parallels and divergences between humans and animals has spurred extensive research in comparative cognition [1]. Since the age of Charles Darwin the prevailing tendency in psychological science has been to highlight the similarities between human and nonhuman minds, suggesting any differences are a matter of degree rather than fundamental nature [2–6]. Among the many cognitive capacities examined, considerable evidence suggests that humans have a unique cognitive ability known as *relational abstraction* - the capacity to perform abstract reasoning and identify abstract relationships between physical objects or symbols [1,7,8]. Relational abstraction allows us to identify logical rules and patterns that help us understand systemic behavior, causal relationships, and classification schemas beyond basic perceptual characteristics; this ability represents a key difference between how humans and other animals learn and interact with the world [1,9]. It has been described as the backbone of thinking [10], and the key executive function that unlocks higher cognition [11,12].

Another exciting frontier has emerged in comparative cognition tied to recent advancements in artificial intelligence (AI) systems like large language models (LMs). State-of-the-art LMs seem to rival human intelligence in many regards, performing remarkably on a variety of academic assessments. For example the GPT4 LM has scored in the 90th percentile on the Bar Exam, the 93rd percentile on the SAT Reading & Writing section, the 89th percentile on the SAT Math section, and received perfect marks on various high school advanced placement exams including biology, calculus, history, psychology, economics, and statistics [13]. This has led some to believe that we are nearing the eve of an artificial general intelligence (AGI) singularity, where AI surpasses human intelligence. Thus, there is growing urgency to formally characterize the cognitive abilities of advanced LM systems and highlight where they currently diverge from human abilities.

If indeed relational abstraction is unique to humans and is key to unlocking higher cognition, it seems prudent to measure AI performance on tests of relational abstraction. One such experimental assessment of relational abstraction abilities is the *relational match-to-sample* (RMTS) test [14]. RMTS

 doi:10.20944/preprints202401.1681.v1

is a well-established measure in comparative cognition research, for which humans easily pass and animals consistently fail [15–21]. In the last quarter century the RMTS test has been used as a benchmark for relational concept learning among comparative cognition researchers. RMTS tasks require subjects to discern and respond to the relational similarity among elements in a given set. This necessitates a level of cognitive abstraction that goes beyond perceptual processing [7]. In the study presented here, we aim to explore whether natural language processing (NLP) capabilities in LMs equip them with the cognitive capacity for analogical reasoning required to pass the RMTS test.

In addition to our goal of determining LM capacity for relational abstraction, we aim to explore whether relational abstraction is an emergent property of advanced language processing abilities. LMs may allow us to determine whether relational abstraction emerges from the mastery of a physical symbol system (PSS) like language, or whether relational abstraction is an independent capacity. It could also be that language abilities are reliant on abstract reasoning capacity, and indeed there is some evidence to suggest relational abstraction is fundamental to interpreting complex linguistic structures [22]. Furthermore, if it were shown that relational abstraction supports language learning in humans, it would also explain why nonhuman animals do not develop complex language. The tests performed here focus on determining whether relational abstraction is an emergent property of a system that has advanced NLP abilities.

Methods

We examined the performance of state-of-the-art LMs in RMTS tests to determine whether current LMs have the cognitive capacity for relational abstraction. Six LMs were tested, including the latest 7b parameter versions of *mistral*, *llama2*, *vicuna*, *solar*, *orca2*, along with the latest OpenAI model *gpt-4-0613*. The 7b models were installed locally using the *ollama toolkit* and were interacted with through a bash shell chat completions API. The OpenAI models were interacted with through API requests. We also tested the *GPT3* and *GPT4* OpenAI models interactively through their web-based ChatGPT interface [1].

The rules of an RMTS test are fairly simple: If the pair of *sample* objects are the same (AA), the participant should choose from among the pairs of *test* objects a '*Same*' pair (BB), not a '*Different*' pair (CD). On the other hand, if sample objects are different (AB), they should choose a *Different* pair (CD), not a *Same* pair (EE). Each round, after the participant chooses a test pair, they are given feedback on whether the choice was correct or incorrect [7]. We used two versions of the RMTS test, one using letters as symbols (see Figure 1) and another using words.

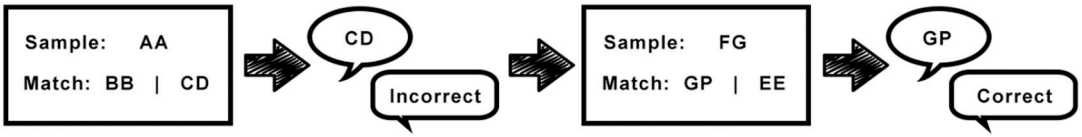

Figure 1. Depiction of two rounds of the letter-based RMTS test used in this report.

We provided the initial RMTS instructions followed by two completion examples (i.e., few-shot learning). The delivery of instructions and response feedback was provided by the system role; each round of sample and test objects were given by the user role (see API footnote).

{"role": "system", "content": "**We are going to play a game. In this game, I will give you a pair of letters called the sample (e.g. sample: FF), followed by two additional pairs of letters called the match (e.g. match: QQ | FH). Given the sample, your goal is to guess which pair of letters from the match sets are correct. After each round you will be told if your guess was correct or incorrect. Responses should be formatted like the following examples...**"},

{"role": "user", "content": "**sample: AA , match: AB | CC** "},

[1] https://chat.openai.com

 doi:10.20944/preprints202401.1681.v1

{"role": "assistant", "content": "**My guess is: CC** "},

{"role": "system", "content": "**Correct!** "},

{"role": "user", "content": "**sample: FJ , match: AB | CC** "},

{"role": "assistant", "content": "**My guess is: AB** "},

{"role": "system", "content": "**Correct!** "},

{"role": "user", "content": "**sample: QQ , match: WJ | KK** "}'

A word-based RMTS test was also performed, such that the pairs of symbols were two words instead of two letters.

Results

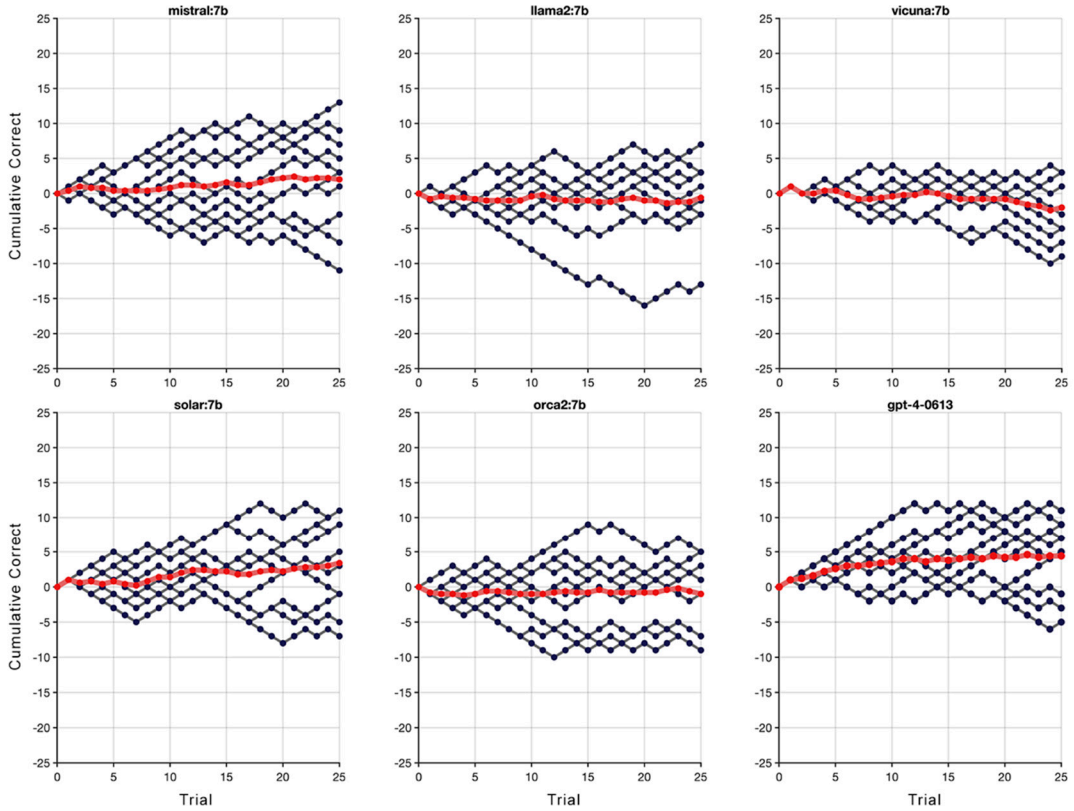

Figure 2. Black lines show cumulative scores for ten independent RMTS tests, where +1 was given for a correct response and -1 was given for an incorrect response. The red line is the mean across all ten tests.

Results showed that all the LMs *failed* the RMTS test (see Figure 2), suggesting they were unable to perform relational abstraction. These findings reveal a stark limitation in current LM abilities. Despite their sophisticated language processing capabilities, these models exhibit a significant deficit in RMTS tasks, as evidenced by their markedly poor performance. This outcome raises profound questions about the relationship between language proficiency and higher-order cognitive abilities, challenging the assumption that advanced language skills inherently confer the capacity for complex relational reasoning.

 doi:10.20944/preprints202401.1681.v1


```
sample: AA, match: AB | CC

sample: A A, match: A B | C C

sample: hat car, match: dog dog | sun cup
```

Figure 3. Tokenization of various letter and word combinations as performed by the OpenAI tokenizer tool[2].

We followed up the letters-based RMTS test with a word-based test. That is, instead of using two letters as the symbols we used two words. This was done for two reasons: (1) to ensure tokenization wasn't driving model performance in spurious ways (see Figure 3), and (2) to determine if the added semantic value of words would alter LM task performance in any regard. It was unclear if and how either of these things would effect LM ability to perform relational abstraction.

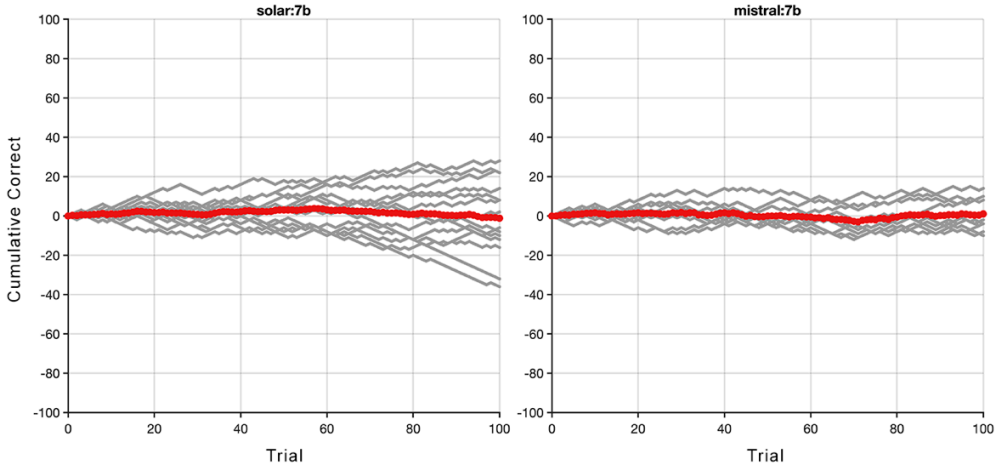

Figure 4. Performance of the solar and mistral models on the word-based version of the RMTS test.

In this word-based test the number of trials was increased from 25 to 100 to determine whether the model needed more examples to make the relational abstraction leap. This was done to determine whether some LMs could improve their performance more slowly than would be detected using only 25 trials. If performance did indeed slowly increase, it would qualitatively differ from typical human performance on RMTS tests; it would be as if LMs improved through a mechanism similar to non-declarative conditioning. For this longer test we examined just two models, including *solar* and *mistral*. Again however, we find that LM performance was at chance guessing levels (see Figure 4).

[2] https://platform.openai.com/tokenizer

 doi:10.20944/preprints202401.1681.v1

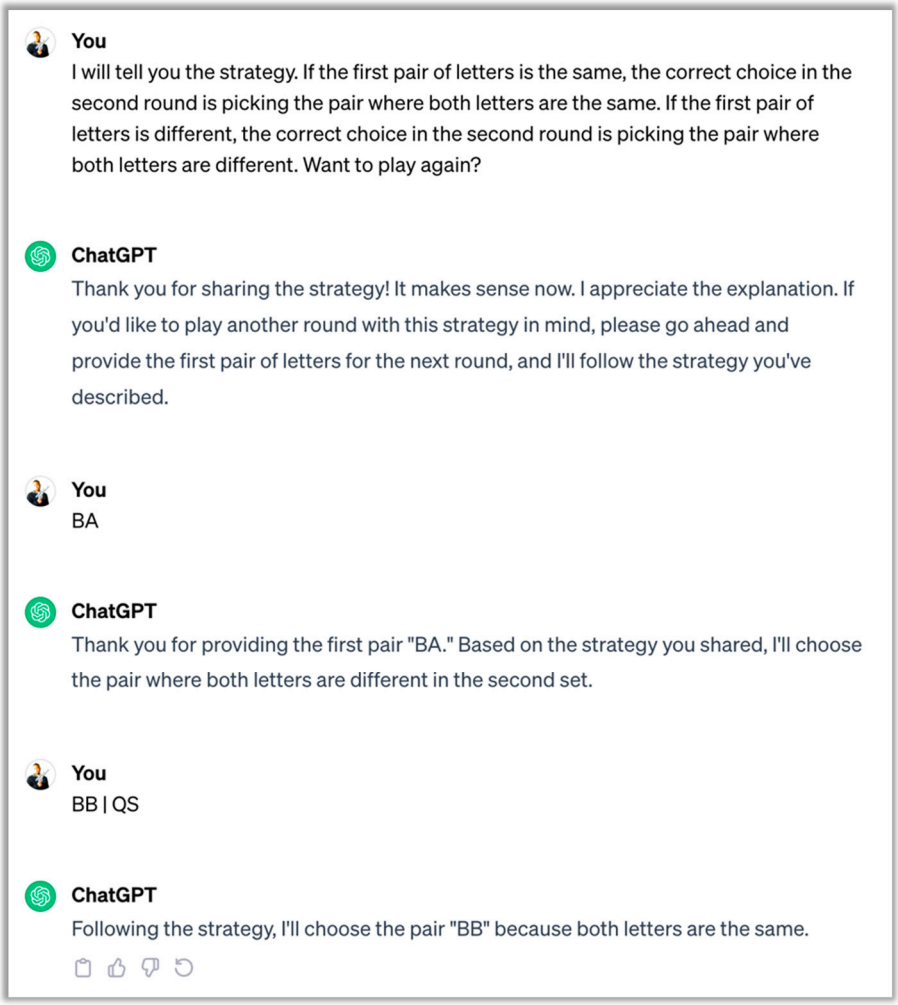

Figure 5. Both GPT3.5 and GPT4 fail the RMTS test even when given the rules ahead of time.

One reason a cognitive system would fail the RMTS test is the inability to glean the logical relationship between the sample and the match. To determine if LMs were unable to abstract the logical rules of the RMTS test, or whether their poor performance was due to something else, we explained the RMTS rules to the LMs and gave a brief assessment. We found that even after explaining the rules, still none of the models performed well. Interestingly, both GPT3.5 and GPT4 correctly repeated and clearly articulated the rules during each step, *yet still failed* to make the correct choice (see Figure 5) [3].

Discussion

What could explain this behavior? How could an AI that aces the Bar Exam and SAT fail a test that is mastered by kindergarteners and some preschoolers? The RMTS test is a simple logical test, so there's no doubt LMs could be trained to pass specific versions of the test through fine-tuning. Our aim here though was to determine if relational abstraction was an emergent phenomena of LMs like "one-shot" or "few-shot" learning.

[3] Link to full GPT4 conversation: https://chat.openai.com/share/a99df312-31f4-4cd9-beec-2cd95be58179

Link to full GPT3.5 conversation: https://chat.openai.com/share/790b425f-368e-4acf-a206-717019c6736d

 doi:10.20944/preprints202401.1681.v1

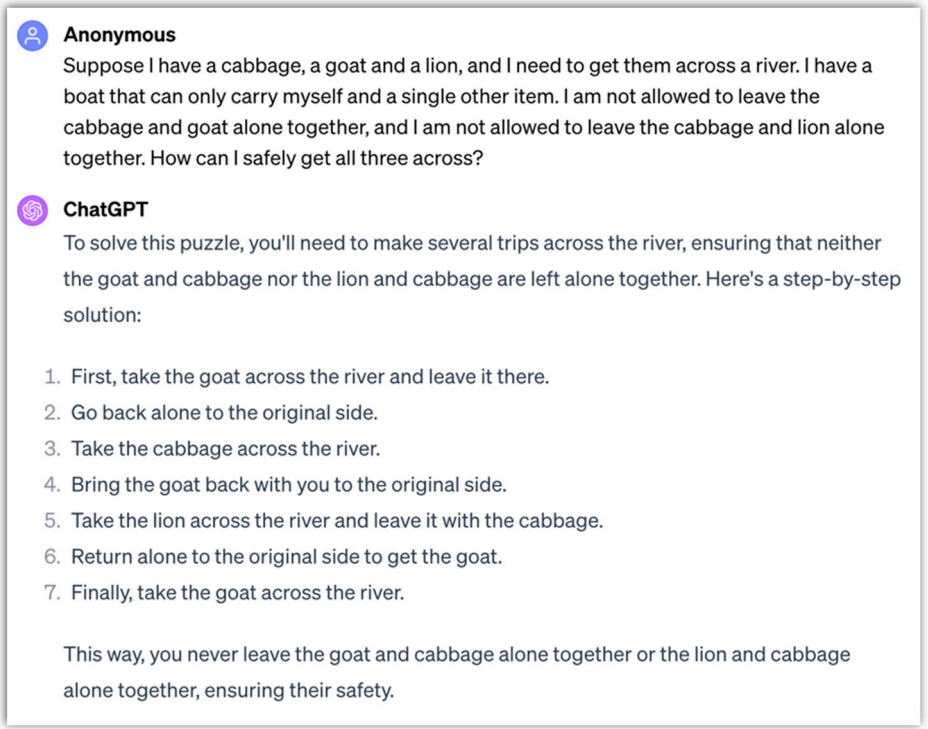

Figure 6. GPT4 provides a response to a modified version of the river crossing problem.

It is known that GPT4 performance on other more difficult logical puzzles is due to rote memorization. For example it has been reported that GPT4 provides the correct answer for the classic *river crossing problem* [23], in which the goal is to carry items from one river bank to another in the fewest trips possible while ensuring the safety of the items. This popular puzzle comes in many variants that have been described online on sites like Wikipedia, which are part of the training corpus for GPT4 and other LLMs [24,25]. If the rules of this riddle are slightly modified, GPT4 fails (see Figure 5) and regurgitates the first step of the solution for the original version, thus immediately failing the riddle under a modified ruleset [4] [5].

These insights provide context for understanding the limitations of LMs in performing tasks that require higher-order relational reasoning. Despite their advanced linguistic capabilities, LMs may not possess the necessary representational processes for such complex cognitive tasks, as suggested by their poor performance on RMTS tests. Higher-order relational reasoning may be a key differentiator in human cognition that helps explain why advanced LMs, despite their apparent linguistic sophistication, fail on tasks requiring such complex cognitive processes. This aligns with our findings, suggesting that language ability alone is insufficient for tasks that demand pattern recognition relational abstraction abilities akin to the human cognitive processes.

Given that advanced LMs do extremely well on standardized tests like the SAT, and claims of AGI passing Turing tests, there is growing urgency to formally characterize the cognitive abilities of advanced LM systems and highlight where they currently diverge with human abilities. A variety of cognitive tests including RMTS have been central in comparative assessments of the mental abilities of humans and animals. As we have shown here, such tests could be used to probe for the emergence of AGI. The following are several tests that could be useful in characterizing the cognitive abilities of AI:

---

[4] Link to GPT4 conversation https://chat.openai.com/share/0b008c03-4d91-4a8c-8287-e1a84a265abb

[5] A solution to the wolf, goat, and cabbage riddle would be: (1) take the goat over, (2) return empty-handed, (3) take the cabbage over, (4) return with the goat, (5) take the wolf over, (6) return with nothing, and finally (7) take the goat over.

 doi:10.20944/preprints202401.1681.v1

1. **Simultaneous Same-Different (S/D) Task**: In this task, subjects are trained to respond differently based on whether two presented stimuli are the same or different, and includes the RMTS test.
2. **Analogical Reasoning Tests**: These assess the ability to draw parallels between different sets of relations, beyond mere perceptual similarity.
3. **Transitive Inference Tasks**: These tasks test the ability to infer relationships between elements based on their positions within a sequence.
4. **Hierarchical Reasoning**: Evaluates the ability to understand and reason about hierarchical structures.
5. **The Physical Causality Task**: This evaluates understanding of physical causation principles.
6. **The Social Learning and Imitation Task**: Assesses the capacity for learning through observation of others' actions.
7. **The Theory of Mind Tasks**: These are designed to test the understanding that others have beliefs, desires, and intentions different from one's own.

Humans are known to perform extremely well on these tests compared to other animals. Challenging LMs on such tests would provide a broad view of model abilities in logical processing, analogy, hierarchical understanding, and relational abstraction.

Overall, this preliminary study contributes to the ongoing discourse on the cognitive distinctions between humans and other entities, whether biological or artificial. This exploration not only enriches our understanding of human cognitive uniqueness but also sheds light on the current limitations of the emergent capacities from LM text completion output. Interesting next-steps might involve using more sophisticated strategies than raw text completion, like chain of though or tree of thought prompting methods.

Conflict of interest statement: The authors have declared that no conflict of interest exists.

References

1. D. C. Penn, K. J. Holyoak and D. J. Povinelli, "Darwin's mistake: Explaining the discontinuity between human and nonhuman minds," *Behavioral and Brain Sciences,* vol. 31, pp. 109-178, 2008.
2. C. Darwin, "The descent of man, and selection in relation to sex.," *John Murray,* 1871.
3. M. Bekoff, C. Allen and G. Burghardt, "The Cognitive Animal," *MIT Press,* 2002.
4. I. M. Pepperberg, "Intelligence and rationality in parrots. Rational Animals?," *Oxford University Press,* pp. 469-488, 2005.
5. J. D. Smith, W. E. Shields and D. A. Washburn, "The comparative psychology of uncertainty monitoring and metacognition," *Behavioral and Brain Sciences,* vol. 26, pp. 317-373, 2003.
6. . Tomasello, J. Call and B. Hare, "Chimpanzees understand psychological states – The question is which ones and to what extent.," *Trends in Cognitive Sciences,* vol. 7, no. 4, pp. 153-156, 2003.
7. D. J. Smith, B. N. Jackson and B. A. Church, "Breaking the Perceptual-Conceptual Barrier: Relational Matching and Working Memory," *Memory & Cognition,* vol. 47, no. 3, pp. 544-560, 2019.
8. E. A. Wasserman and M. Young, "Same-different discrimination: The keel and backbone of thought and reasoning," *Journal of Experimental Psychology,* vol. 36, pp. 3-22, 2010.
9. J. Locke, An essay concerning human understanding, Philadelphia, PA: Troutman & Hayes, 1690.
10. W. James, The principles of psychology, New York, NY: Dover, 1890.
11. G. S. Halford, W. H. Wilson and S. Phillips, "Relational knowledge: The foundation of higher cognition," *Trends in Cognitive Sciences,* vol. 14, pp. 497-505, 2010.
12. G. J. and I. Litvan, "Importance of deficits in executive functions," *Lancet,* vol. 354, pp. 1921-1923, 1999.
13. OpenAI, "GPT-4 Technical Report," *arXiv,* 2023.
14. D. Premack, "Animal cognition," *Annual Review of Psychology,* vol. 34, pp. 351-362, 1983.
15. A. Shivaram, R. Shao, N. Simms, S. Hespos and D. Gentner, "When do Children Pass the Relational-Match-To-Sample Task?," *Proceedings of the Annual Meeting of the Cognitive Science Society,* 2023.
16. D. E. Carter and T. J. Werner, "Complex learning and information processing by pigeons: A critical analysis," *Journal of the Experimental Analysis of Behavior,* no. 29, pp. 565-601, 1978.
17. P. W. Holmes, "Transfer of matching performance in pigeons," *Journal of the Experimental Analysis of Behavior,* vol. 31, pp. 103-114, 1979.
18. D. A. Washburn and D. M. Rumbaugh, "Rhesus monkey (Macaca mulatto) complex learning skills reassessed," *International Journal of Primatology,* vol. 12, pp. 377-388, 1991.

 doi:10.20944/preprints202401.1681.v1

19. D. M. R. and M. Columbo, "On the limits of the matching concept in monkeys (Cebus apella)," *Journal of the Experimental Analysis of Behavior,* no. 52, pp. 225-236, 1989.
20. J. S. Katz, A. A. Wright and J. Bachevalier, "Mechanisms of same/different abstract-concept learning by rhesus monkeys (Macaca mulatta)," *Journal of Experimental Psychology: Animal Behavior Processes,* vol. 28, pp. 358-368, 2002.
21. W. E. Shields, J. D. Smith and D. A. Washburn, "Uncertain responses by humans and rhesus monkeys (Macaca mulatta) in a psychophysical same-different task," *Journal of Experimental Psychology,* vol. 126, pp. 147-164, 1997.
22. Gentner, Shao, Simms and Hespos, "Learning Same and Different Relations: Cross-Species Comparisons," *Current Opinions in Behavioral Sciences,* vol. 37, pp. 84-89, 2001.
23. M. Ascher, "A River-Crossing Problem in Cross-Cultural Perspective," *Mathematics Magazine,* vol. 63, no. 1, pp. 26-29, 1990.
24. "River crossing puzzle," [Online]. Available: https://en.wikipedia.org/wiki/River_crossing_puzzle. [Accessed 18 December 2023].
25. "Wolf, goat, and cabbage problem," [Online]. Available: https://en.wikipedia.org/wiki/Wolf,_goat_and_cabbage_problem. [Accessed 18 December 2023].