The Black Stories Experiment: Two Groups are Trying to Solve a Riddle Game Behind a Screen, Only One Group Is Alive

Yanna E. Smid (y.e.smid@umail.leidenuniv.nl)
Nikki S. Rademaker (n.s.rademaker@umail.leidenuniv.nl)
Linthe van Rooij (l.van.rooij.3@umail.leidenuniv.nl)
Tessa Verhoef (t.verhoef@liacs.leidenuniv.nl)
Leiden Institute of Advanced Computer Science, Leiden University
Einsteinweg 55, 2333 CC, Leiden, The Netherlands

Abstract

Investigating how large language models (LLMs) perform on complex tasks can provide valuable insights into their strengths and limitations, while also highlighting ways in which they may complement human cognition. This study explores problem-solving capabilities of GPT-4 by comparing the performance of the model in solving Black Stories riddles to human performance. Players have to uncover a hidden story by asking yes-or-no questions, testing their logic, creativity, and inference skills. The study utilized a set of 12 existing Black Stories, with deviations in details and each tested twice within the human and GPT-4 group. The experiment was conducted through text messaging for both humans and GPT-4 to make the testing set-up maximally similar between groups. The primary measure of performance was the number of questions needed for the participant to solve the riddle, taking into account the number of hints used to come to the solution. Results indicated no significant difference between the groups. Qualitative results, however, showed that GPT-4 excelled in precise questioning and creativity but often fixated on details, missing the bigger picture and summarizing solutions prematurely. Humans covered broader topics and adapted the focus quickly but struggled with uncommon details. This research suggests that despite different approaches, GPT-4's performance was comparable to that of humans, demonstrating its potential as a capable participant in these types of problem-solving games.

Keywords: Problem-solving; Large Language Models; riddles; logical reasoning

Introduction

Large Language Models (LLMs) are artificial intelligence (AI) programs designed to process and generate text. Recent advancements have demonstrated remarkable success of these models across a variety of linguistic tasks (Wang et al., 2019), showcasing their capacity to generate coherent, contextually relevant, and nuanced responses. This success has been largely driven by the introduction of transformer-based architectures (Vaswani et al., 2017), which have significantly improved the ability of models to process and generate natural language. The extent to which these models exhibit genuine understanding and reasoning abilities remains a subject of intense scholarly debate (Bender & Koller, 2020; Chang & Bergen, 2024; Mitchell & Krakauer, 2023; Mahowald et al., 2024; Huang & Chang, 2023).

Benchmarks have been designed to test whether LLMs can simulate human reasoning abilities through logical question answering. As reviewed by Cheng et al. (2025), modern LLMs still often fail to solve such logical problems correctly, and tend to exhibit logical inconsistencies in their answers, directly contradicting their own previous answers.

While LLMs have been found to be able to generate human-like responses to some extent in domains such as mathematical reasoning (Wei et al., 2022; Liu et al., 2024), logical inference (Lampinen et al., 2024) and natural language inference (Bowman, Angeli, Potts, & Manning, 2015), errors occurring at the same time on simple deviations of the same tasks in many domains are inconsistent with the interpretation that the models show human-like understanding (McCoy, Yao, Friedman, Hardy, & Griffiths, 2024; Traylor, Feiman, & Pavlick, 2021; McCoy, Pavlick, & Linzen, 2019; Mitchell & Krakauer, 2023). Benchmarks that are based on cognitive tests designed for humans are not always suitable for testing abilities in LLMs, due to their ability to exploit spurious correlations, resulting in shortcut learning (Mitchell & Krakauer, 2023). Moreover, human intelligence is not confined to the isolated brain but can be seen as collaboratively constructed in social interaction and cultural contexts (Hutchins, 1995). Here, we focus on a novel way to assess reasoning abilities in LLMs, embedded in a *narrative* and *interactive* context, where the LLM *asks* instead of answers the questions, avoiding the use of typical benchmark-style isolated question-answer pairs.

Our method is based on the game "Black Stories". Black Stories are riddles that describe mysterious and often dark scenarios. The game involves multiple players, where one player reads a brief cryptic description of the ending of the story out loud and knows the full story. The goal of the other player(s) is to uncover the full story by asking yes-or-no questions, which will be answered solemnly with 'yes', 'no', 'false assumption', or 'not relevant'. Looking into how LLMs handle such riddles can provide insights into their reasoning capabilities in context, maintaining a dialogue while piecing together a logical and coherent narrative by asking relevant questions.

Since GPT-4 often outperforms other LLMs in riddle-solving tasks, which will be further discussed in the Background, we focus on this state-of-the-art model to answer the following main research question: *How does the performance of GPT-4 compare with human performance when solving Black Stories and what do the questions used reveal about their problem-solving strategies?*.

Given previously found challenges in the reasoning of LLMs, both in general (Chang & Bergen, 2024; Bang et al., 2023) and in riddles specifically (Lin, Wu, Yang, Lee, & Ren,

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

2021; Jiang, Ilievski, Ma, & Sourati, 2023; Del & Fishel, 2023), combined with the open-ended nature of our novel task, we may expect that GPT-4 will have difficulties solving Black Stories and may differ from humans in their approach to solving these riddles. On the other hand, the rich context of the story and the interactive nature of the task may positively affect the performance. Comparing the performance of GPT-4 and humans might show the strengths and weaknesses of both the model and humans. This can improve our understanding of the extent to which LLMs can mimic or complement human-like cognition in complex reasoning tasks.

Background

Riddles are cognitive puzzles, often presented in the form of a question, statement, or phrase, designed to challenge the solver's ingenuity and creativity. They are usually phrased ambiguously or metaphorically, requiring the solver to interpret clues creatively rather than literally, and are therefore often considered a great tool for testing flexibility in reasoning (Bar-Hillel, Noah, & Frederick, 2018). Bar-Hillel et al. (2018) showed how specific riddles called "stumpers" use common biases to mislead solvers. For example, gender stereotypes or assumptions about the time of day can block people from finding the correct answers. These riddles force solvers to rethink their assumptions and to consider alternative solutions. In some countries, riddles are used to teach reasoning and observation. Gwaravanda and Masaka (2008) studied Shona riddles in Zimbabwe. They found that these riddles help children develop skills in logic, memory, and quick thinking. Shona riddles often involve analogies, which require solvers to connect abstract ideas with concrete objects. Absattarovna (2021) explored the role of riddles in developing logical thinking. Riddles teach solvers to analyze clues, find patterns, and eliminate incorrect answers. In addition to that, they also encourage creativity. In sum, riddles are used to both test and teach cognitive skills in humans. Below we review previous work on riddle-solving in LLMs.

Riddle-Solving in LLMs

Lin et al. (2021) created the RiddleSense dataset to evaluate creativity and commonsense reasoning in models. Skills like understanding metaphors, and interpreting creative language are needed to interpret "I have five fingers but am not alive" as "glove". Humans had a much higher accuracy (91.3%) for solving riddles than the best model (68.8% for UnifiedQA by Khashabi et al. (2020)).

In addition, Jiang et al. (2023) introduced BrainTeaser, a novel benchmark designed to evaluate lateral thinking in LLMs. Unlike conventional vertical thinking tasks that depend on commonsense reasoning, lateral thinking puzzles challenge default assumptions and require creative, divergent reasoning. An example of this is "How could a cowboy ride into town on Friday, stay two days, and ride out on Wednesday?", where the right answer would be "His horse is named Wednesday". The best-performing model, ChatGPT, still performs significantly worse than humans, achieving only 53-63% accuracy compared to almost 92% for human scores.

Slightly more similar to our method, the True Detective task assesses abductive reasoning with stories as input (Del & Fishel, 2023). Here, models identify the most justified explanation for a set of clues in complex detective puzzles, sourced from the "5 Minute Mystery" platform. Models are tested through a multiple-choice question, which is hard to answer even for humans, who typically get about 47% of them right. The model with the highest score is again from the GPT family (GPT-4), and scores 38% correct, which is halfway between random guessing and the average human baseline.

Unlike these previous studies, our method does not involve asking the LLM a series of isolated complex questions, but invites the model to generate relevant questions to solve a riddle embedded in a larger story. The context of a story, or narrative, is proposed to be a fundamental structure of human meaning-making (Bruner, 1987) and stories have been demonstrated to help humans in problem-solving tasks (Hernandez-Serrano & Jonassen, 2003), as well as in other areas of cognition such as memorization, decision-making, planning and improvising (Schank & Berman, 2003). A recent study suggested that stories can in fact help LLMs with reasoning as well (Javadi, Trippas, Lal, & Flek, 2024) by providing a structured context for the problem domain.

Black Stories riddles require solvers to rebuild narratives by asking a series of yes-or-no questions, therefore also providing an interactive reasoning setting in contrast to previous studies. Complex cognitive skills are needed to solve Black Stories such as creative, out-of-the-box thinking to imagine unusual scenarios or solutions, deductive reasoning to recognize common patterns, use known facts and rule out possible options, and attention to detail for interpreting subtle hints in the initial riddle and clues revealed during questioning. To our knowledge, this game has not been previously explored within the context of reasoning in LLMs. As described in the introduction, we directly compare the performance of GPT-4 to that of humans who are asked to play the game in a comparable text-only setting.

Methodology

Black Stories Dataset

Twelve Black Stories were selected from the English version of the board game Black Stories (Bösch & Andersen, 2007). We selected these stories based on their comparable level of difficulty and length. Each Black Story was adjusted to a deviation of the story to prevent it from being recognized by the LLM. Table 1 shows an example of a deviation of one selected story. Key components to guess were extracted from each solution. For the story shown in Table 1 these include for example: "Hot air balloon", "Run out of fluid", "To reduce the demand of fluid, removed all their clothing (was not enough)", "Threw dice to decide who is going to jump", "But love decided - so jumped together".

Table 1: Example of an original and deviated version (given to participants and LLM) of a Black Story Situation and Solution.

Story	Situation	Solution
Original	A stark naked man was found dead at the foot of a mountain - with a matchstick in his hand.	A hot-air balloon carrying four passengers had gone off course and threatened to smash into a mountain. To gain height, the passengers threw all the ballast, including their clothing, overboard. It wasn't enough: one of them would have to jump. They drew slots - and the dead man drew the shortest match.
Deviation	A naked couple was found dead in a forest - with a dice in their hands.	A hot-air balloon carrying four passengers was almost out of combustion fluid. To reduce their demand for the fluid, preventing they land in the middle of a forest and allowing them to land safely, the passengers threw all the ballast, including their clothing, overboard. It wasn't enough: one of them would have to jump. They threw dice - and one of the dead couple lost. But since they were hopelessly in love and could not live without one another, the other one jumped along.

Analysis Methods

The analysis of the collected data included several factors:

- **Question Count**: We determined the question count as our main pillar of success for solving riddles. Fewer questions suggest that the solver is asking precise and targeted questions, while systematically narrowing down possibilities.
- **Hints**: Due to the complexity of Black Stories, we provided hints to guide solvers, following the original game mechanics. Hints were either requested by participants or offered by experimenters when repetition was noticed, solutions were insufficient, or questions led nowhere. Since LLMs don't have inherent mechanisms for independent hint asking, experimenters only intervened directly in this experimental group. Each hint carried a weighted value added to a corrected question count, defined as the variable *score* (see Equation 2). The weight was calculated as

$$w = \frac{\#Q - \overline{\#QH}}{\overline{\#H}} \tag{1}$$

  In this equation, $\#Q$ is the number of questions without hints, $\overline{\#QH}$ is the average number of questions with hints, and $\overline{\#H}$ is the average number of hints. The equation determines how many questions each hint effectively saves. The final score was then calculated as

$$score = \#Q + (\#H \times w) \tag{2}$$

  where the number of hints $\#H$ is multiplied by the calculated weight, adding up to the total question count $\#Q$.
- **Word Count**: The average word count per question was calculated to assess how detailed and elaborate the questions were. This average count was further averaged over all the questions in a single story for one participant or one round of the LLM.
- **Qualitative Analysis**: Additional qualitative analyses were done to investigate patterns related to strategies for solving the game, such as focus switching between various areas of the solution and the revelation of details to derive the final solution.

Experiment Part A (Alive)

Participants 16 volunteers aged 18-35 (M: 23.3, SD: 2.5) participated in the experiment. The participants were recruited through the network of the experimenters. To ensure a fair comparison of problem-solving abilities, prior knowledge of Black Stories and English fluency were required to match the knowledge and capabilities of the LLM as much as possible, rather than assessing humans' ability to figure out how the game works when playing it for the very first time.

Procedures The experiment was conducted via a private WhatsApp conversation with one of the experimenters and the participant to mimic GPT-4's text-only conditions and minimize human advantages, such as non-verbal communication and intonation. The participants were initially instructed on the rules of the game before the experiment officially started (Table 2). Additionally, the instructions included information about the average time length of the experiment, the language of the experiment, the format of asking one question at a time, and the availability of hints. Participants were informed that they could request a hint or that hints were provided by the experimenter to guide them in the appropriate direction.

Table 2: Instructions given to both humans and GPT-4

> "You have to solve a Black Stories riddle. I'll tell you a Black Story and you have to solve it by asking me questions. You have to solve the riddle with as few questions as possible. The riddle tells you the end of a story and you have to find out what lead to this end. When I tell you the riddle, you have to try to solve the riddle by asking questions that I can only answer by yes, no or not relevant to the story. You will use my answers to solve the riddle and find the story that lead to the end. I'll tell you when you have solved the riddle, then you give me a summary of the story of the riddle. Note that giving a summary to guess the answer is also counting as one question."

The experiment started with a deviated Black Stories' situation sketch as shown in Table 1. If the participant was already familiar with a particular story, an alternative story from the 12 available stories was selected. Participants proceeded with questioning until they solved the riddle based on the key components of the solution. Depending on their availability, one or two Black Story Riddles were completed. Each story took approximately 30-45 minutes to solve. Upon completion, the WhatsApp conversation was exported as a .txt file for analysis in Python. We aimed to test each story twice to minimize individual biases and reduce variability. However, one story was only tested only once due to time constraints. In total, 23 data points were collected across 12 different deviated stories.

Experiment Part B (Bot)

LLM GPT-4 was selected as the pre-trained model for solving Black Story Riddles. This model (`gpt-4`) was accessed via the OpenAI API. Standard temperature and top_p settings were used.

Procedures The experiment was conducted via Python's terminal. System content was given to GPT-4 as shown in Table 2. The experiment started with a deviated Black Stories' situation sketch as shown in Table 1. The model proceeded with questioning until they solved the riddle based on the key components of the solution. Hints were provided to GPT-4 by the experimenter's judgment to guide the model in the right direction. This judgment was based on factors, such as incomplete preliminary summaries, repetitive questioning around specific details, or moments when humans would typically request hints, as GPT-4 did not request hints independently. When a sufficient solution was given, the conversation was ended and saved to a data file. One story selected was recognized by GPT-4 and therefore replaced with another story. One story was conducted three times, resulting in a total of 25 data points across 12 different stories.

Results

Quantitative Analysis

Question Count The difference in question count without hint correction between GPT-4 and humans across all Black Stories was analyzed with an independent t-test. Assumptions of normality and variance were met. GPT-4 had an average question count of 46 (SD=17.1) similar to humans with an average question count of 45 (SD=17.0). A student's t-test showed no significant difference between group means $t(46) = 0.253$, $p = 0.801$.

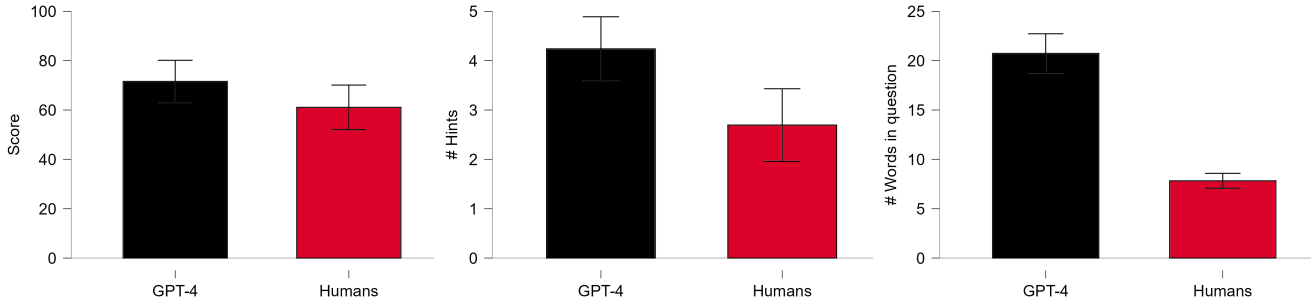
Score Using Equation 1, the hint weight was set to 6 based on GPT-4's 59 questions without hints for one story, compared to an average of 35 questions with 4 hints for the same story. This weighting was applied to analyze score differences between GPT-4 and humans across all stories using an independent t-test. A student's t-test showed no significant difference between group means $t(46) = 1.450$, $p = 0.154$, despite humans (M=61.1, SD=25.2) gaining a slightly lower average score than GPT-4 (M=71.6, SD=25.0), see Figure 1a.

Hint Count The difference in score displayed in Figure 1a shows a slightly better, although not significant, score for humans. However, when comparing these results with the average question count, hints seem to impact the score. The difference in the number of hints between GPT-4 and humans across all Black Stories was analyzed with an independent t-test. Assumptions of normality and variance were met. A student's t-test showed a significant difference between group means $t(46) = 2.706$, $p < 0.05$, with GPT-4 (M=4.2, SD=1.9) using more hints on average than humans (M=2.7, SD=2.1), see Figure 1b.

Additionally, an ANOVA test was conducted to test

Table 3: Excerpts of questions from GPT-4 (top) and humans (bottom) for the icicle murder story set in a sauna, showing that GPT-4's questions involved detailed scenarios and possible causes of death, while humans focused on specific physical causes.

GPT-4	
did the other person use the water bottle to cause the woman's death	did the other person lock the woman in the sauna causing her death by overheating or dehydration
did the death result from a physical altercation with the other person	did the other person tamper with the water bottle in a way that led to the woman's death
was the cause of death related to the sauna's heat or steam	was the woman's death caused by something unrelated to the sauna environment, the water bottle, and not directly physical, such as a scare or shock
did the other person poison the woman	did the woman die from natural causes
did the other person do something to the sauna equipment to cause her death	did the woman die from an allergic reaction
did the woman's death result from a preexisting medical condition that was triggered in the sauna	did the death involve electrocution
Humans	
did she suffocate	was she shot
did she drown	was she hit with an item
was she trapped	was she stabbed
was she poisoned	



(a) Average score across all Black Stories (b) Average number of hints needed (c) Average word count of a sentence

Figure 1: Quantitative comparison of human and GPT-4 performance

whether the number of hints given to GPT-4 differentiated between the three experimenters due to its relative subjective procedure. A post hoc Tukey HSD showed no significant differences between experimenter 1 (M=4.375, SD=2.560), 2 (M=5.125, SD=1.553), and 3 (M=3.333, SD=1.118) in the number of hints that were given to GPT-4.

Word Count The difference in word count between GPT-4 and humans across all Black Stories was analyzed with an independent t-test. Assumptions of normality and variance were not met, therefore a Welch's test was performed showing a significant difference in word count ($t(30.7) = 12.556$, $p < 001$). As shown in Figure 1c, GPT-4 was using longer sentences (M=20.7, SD=4.8) than humans (M=7.8, SD=1.8), including more details as discussed in the next section.

Qualitative Analysis

In addition to the quantitative findings, the dialogues were analysed for more insights into the strategies used by both participant types.

Table 4: Excerpts of questions by GPT-4 (left) and a human (right), showing that humans switch focus faster and cover more topics.

GPT-4	Humans
are they playing russian roulette with each move in the checkers game	are they on a secret mission
are they playing checkers to determine who will get to use the pistol	are there other people
does the man want to use the pistol on the woman or himself	is it sinking
does the man want to use the pistol on himself	is there a way out of the submarine
does the woman want to use the pistol on him as well	do they play checkers as a distraction from dying
	is there a reason they are playing checkers

Table 5: Excerpts of quick summaries from GPT-4

GPT-4
so to summarize ... concept for her
did the woman check if she could see while she was inside the dark tunnel
in summary the ... inside the tunnel
did she jump out of the bus in panic of the unknown fearing the possibility that her vision was still impaired despite the surgery
to sum up ... this tragic outcome

Detailed vs. Direct Questions A key finding was the contrast in detail between GPT-4 and human participants. GPT-4's responses were more elaborate, often suggesting multiple possible scenarios (e.g., "death by overheating or dehydration" in Table 3), while humans asked targeted, single-focus questions to uncover precise items. Moreover, the detailed questions given by GPT-4 often guessed unusual and specific settings, whereas humans often had more difficulty identifying the story's creative settings.

Topic Coverage and Switching The dialogue analysis further shows that humans were more capable of switching their focus in the stories and covering more topics faster. In contrast to the LLM, which often stuck to further specific aspects of the same detail in the storyline (see Table 4).

Premature summarization While human participants typically ask multiple questions and then provide a summary only at the end, GPT-4 has a tendency to give premature summaries, showing a pattern of asking one question followed by another summary immediately (see Table 5).

Emotions and Affirmation in Humans. Human participants often express emotions when frustrated or motivated by the game. They included phrases such as *"am I even close I am so bad at this"*, or *"I just named all boat types haha do I have to go in a different direction"* and *"what am I missing"* when seeking affirmation to go in the right direction. As expected, such phrases were absent in the LLM responses.

Discussion

Using the game Black Stories, we introduced a novel approach to explore complex problem solving in LLMs, in which participants ask instead of answer the questions, to solve a riddle interactively in the context of a larger story. When comparing human and GPT-4 performance, both groups were able to eventually get to the right solution for every riddle and the average score did not differ significantly between the two types of participants. This does not mean, however, that GPT-4 solves this game in a human-like way and the model may exploit a different strategy to arrive at the same outcome (Mitchell & Krakauer, 2023). Indeed, when analyzing the dialogues and questions asked, we see clear differences. While GPT-4's questions were generally very detailed, this level of detail sometimes hindered its ability to pinpoint the correct answer efficiently. For example, the model occasionally posed questions that included multiple possibilities, like "...overheating or dehydration" (see Table 3) and proposed elaborate but unlikely scenarios. In such cases, hints were provided to redirect GPT-4 toward the correct path, as it was getting close to the answer but not entirely accurate. Human participants, on the other hand, were more likely to focus on larger storylines first before going into detail and resisted the use of hints, persisting until they were unable to deduce the answer independently. While emotions play no role in the decision-making of LLMs, humans tended to react more emotionally regarding pride or self-reliance, motivating them to solve the riddle without any help.

On average, GPT-4 used more than twice the number of words as human participants when formulating questions. This finding resonates with previous observations of verbosity bias in LLMs (Saito, Wachi, Wataoka, & Akimoto, 2023), where models tend to prefer more lengthy answers even if they are not more informative. In addition, when neural network or LLM agents are trained to play repeated signaling games, they often develop longer words than humans would, and may not naturally adopt the efficiency patterns observed in human communication (Chaabouni, Kharitonov, Dupoux, & Baroni, 2019; Kouwenhoven, Peeperkorn, & Verhoef, 2025). Interestingly, this verbosity sometimes gave GPT-4 an advantage during the experiments, as its questions covered a broader range of facts from the riddles and were generally more detailed. In some cases, this helped GPT-4 to be quicker to identify unusual aspects or settings in the storyline. Such unusual settings were harder for human participants to detect, as their focus was often on more conventional aspects, such as identifying the cause of death (see Table 5). This difference could be attributed to GPT-4's design, which is extensively instruction-tuned to be conversational and increase user comprehension. In contrast, humans tended to focus on formulating concise and to-the-point questions in a more efficient communication style.

However, while GPT-4's eye for detail sometimes gave it an edge, the model struggled with shifting focus. Over time, GPT-4 occasionally became fixated on specific details in the story and continued asking questions about these aspects. In some cases, even after posing the right question and receiving a correct answer, GPT-4 continued exploring the same topic instead of returning to broader aspects of the storyline. In contrast, human participants demonstrated a stronger ability to shift focus between topics. As seen in Table 4, humans explored a wider range of storyline aspects in consecutive questions, ultimately getting to the solution more efficiently.

Our work has a few notable limitations. Unlike humans, GPT-4 was not able to independently decide when to ask for hints. In our experiment, it was therefore up to the experimenters to decide when GPT-4 was given a hint, based on typical situations in which humans would ask for it. The differences in the number of hints needed to solve the game between humans and GPT-4 could have been influenced by this. However, this does not have any consequences for our overall findings since, even with the heightened use of hints, the score of the GPT-4 group still did not differ significantly from that of the human group. Moreover, the number of hints given to GPT-4 did not differ significantly between experimenters, showing that although there was possibly some subjectivity present, this likely did not affect the findings for the model. Nonetheless, future work could explore more advanced prompting methods that would possibly allow the model to ask for a hint when needed, providing a more objective method for giving hints.

Another potential disadvantage of the presented approach is the time needed to conduct these experiments. Most previous work that explored riddle-solving in LLMs evaluated the models automatically using benchmarks, but our approach is more interactive and requires experimenters to have an ongoing conversation with the model, costing significant data collection time. While this unusual method allows for a unique exploration of LLM reasoning abilities in an interactive setting, it also causes this method to be less easy to implement than using typical riddle task benchmarks.

Future studies could investigate the performance of hybrid teams of humans and LLMs in solving Black Stories. Given the slightly different approaches used by both groups, combining their strengths could potentially lead to advantages in solving this game and in problem-solving more generally.

Conclusion

This study explored the problem-solving abilities of GPT-4 in Black Stories riddles, comparing its approach to that of human participants. The findings align with the hypothesis that both participant types employ distinct strategies in solving these riddles. Specifically, the difference lies in the formulation of questions and focus areas. GPT-4 applies a detailed and extensive exploration, while humans are more concise and adaptable. Despite these differences, GPT-4's performance was comparable to that of humans, demonstrating its potential as a capable participant in these types of problem-solving games.

References

Absattarovna, T. K. (2021). The role of riddles in teaching logical thinking. *International Journal of Innovations in Engineering Research and Technology*, *8*(3), 24–25.

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., ... Fung, P. (2023). A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*. Retrieved from https://arxiv.org/abs/2302.04023

Bar-Hillel, M., Noah, T., & Frederick, S. (2018). Learning psychology from riddles: The case of stumpers. *Judgment and Decision Making*, *13*(1), 112–122. doi: 10.1017/S193029750000886X

Bender, E. M., & Koller, A. (2020, July). Climbing towards NLU: On meaning, form, and understanding in the age of data. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5185–5198). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.acl-main.463/ doi: 10.18653/v1/2020.acl-main.463

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In L. Màrquez, C. Callison-Burch, & J. Su (Eds.), *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 632–642). Lisbon, Portugal: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D15-1075/ doi: 10.18653/v1/D15-1075

Bruner, J. (1987). Life as narrative. *Social research*, 11–32.

Bösch, H., & Andersen, D. (2007). *Black stories : solve 50 creepy mysteries* (21th ed.). Moses.

Chaabouni, R., Kharitonov, E., Dupoux, E., & Baroni, M. (2019). Anti-efficient encoding in emergent communication. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)* (Vol. 32). Vancouver, Canada.

Chang, T. A., & Bergen, B. K. (2024). Language model behavior: A comprehensive survey. *Computational Linguistics*, *50*(1), 293–350.

Cheng, F., Li, H., Liu, F., van Rooij, R., Zhang, K., & Lin, Z. (2025). *Empowering llms with logical reasoning: A comprehensive survey.* Retrieved from https://arxiv.org/abs/2502.15652

Del, M., & Fishel, M. (2023). True detective: A deep abductive reasoning benchmark undoable for GPT-3 and challenging for GPT-4. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (* SEM 2023)* (pp. 314–322).

Gwaravanda, E. T., & Masaka, D. (2008). Shona reasoning skills in zimbabwe: the importance of riddles. *Journal of Pan African Studies*, *2*(4).

Hernandez-Serrano, J., & Jonassen, D. H. (2003). The effects of case libraries on problem solving. *Journal of Computer Assisted Learning*, *19*(1), 103–114.

Huang, J., & Chang, K. C.-C. (2023). Towards reasoning in large language models: A survey. In *61st annual meeting of the association for computational linguistics, acl 2023* (pp. 1049–1065).

Hutchins, E. (1995). *Cognition in the wild*. Cambridge, MA: MIT Press.

Javadi, V. S., Trippas, J. R., Lal, Y. K., & Flek, L. (2024). Can stories help LLMs reason? Curating information space through narrative. In *The first workshop on system-2 reasoning at scale, neurips'24*.

Jiang, Y., Ilievski, F., Ma, K., & Sourati, Z. (2023). BRAINTEASER: Lateral thinking puzzles for large language models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 14317–14332). Singapore: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.emnlp-main.885 doi: 10.18653/v1/2023.emnlp-main.885

Khashabi, D., Min, S., Khot, T., Sabharwal, A., Tafjord, O., Clark, P., & Hajishirzi, H. (2020). UNIFIEDQA: Crossing format boundaries with a single QA system. In T. Cohn, Y. He, & Y. Liu (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 1896–1907). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.findings-emnlp.171/ doi: 10.18653/v1/2020.findings-emnlp.171

Kouwenhoven, T., Peeperkorn, M., & Verhoef, T. (2025). Searching for structure: Investigating emergent communication with large language models. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)* (pp. 9977–9991).

Lampinen, A. K., Dasgupta, I., Chan, S. C., Sheahan, H. R., Creswell, A., Kumaran, D., ... Hill, F. (2024). Language models, like humans, show content effects on reasoning tasks. *PNAS nexus*, *3*(7), pgae233.

Lin, B. Y., Wu, Z., Yang, Y., Lee, D.-H., & Ren, X. (2021). RiddleSense: Reasoning about riddle questions featuring linguistic creativity and commonsense knowledge. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 1504–1515).

Liu, W., Hu, H., Zhou, J., Ding, Y., Li, J., Zeng, J., ... others (2024). Mathematical language models: A survey. *arXiv preprint arXiv:2312.07622*.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, *28*, 517-540.

McCoy, R. T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3428–3448). Florence, Italy:

Association for Computational Linguistics. Retrieved from https://aclanthology.org/P19-1334/ doi: 10.18653/v1/P19-1334

McCoy, R. T., Yao, S., Friedman, D., Hardy, M. D., & Griffiths, T. L. (2024). Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, *121*(41), e2322420121.

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, *120*(13), e2215907120.

Saito, K., Wachi, A., Wataoka, K., & Akimoto, Y. (2023). Verbosity bias in preference labeling by large language models. In *Neurips 2023 workshop on instruction tuning and instruction following.*

Schank, R. C., & Berman, T. R. (2003). The pervasive role of stories in knowledge and action. In *Narrative impact* (pp. 287–313). Psychology Press.

Traylor, A., Feiman, R., & Pavlick, E. (2021). AND does not mean OR: Using formal languages to study language models' representations. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (pp. 158–167). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.acl-short.21/ doi: 10.18653/v1/2021.acl-short.21

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In I. Guyon et al. (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30, pp. 6000–6010). Curran Associates, Inc.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., ... Bowman, S. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems*, *32*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... others (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837.