Computational Linguistics in the Netherlands Journal 14 (2025) 451-472 Submitted 12/2024; Published 07/2025

The Riddle Experiment: two groups are trying to solve a Black Story behind a screen, only one group is alive.

Nikki S. Rademaker* N.S.RADEMAKER@UMAIL.LEIDENUNIV.NL
Linthe van Rooij* L.VAN.ROOIJ.3@UMAIL.LEIDENUNIV.NL
Yanna E. Smid* Y.E.SMID@UMAIL.LEIDENUNIV.NL
Tessa Verhoef* T.VERHOEF@LIACS.LEIDENUNIV.NL

**Leiden Institute of Advanced Computer Science, Leiden University, Leiden, The Netherlands*

Abstract

Investigating the cognitive abilities of large language models (LLMs) can inform theories about both artificial and human intelligence and highlight areas where AI may complement human cognition. This study explores GPT-4's logical reasoning abilities by comparing its performance in solving Black Story riddles to that of humans. Black Stories are riddles where players reconstruct a hidden narrative by asking yes-or-no questions to a player who knows the full story. These riddles test logical reasoning, creativity, and inference skills of the solvers in an interactive setting. The study utilized a set of 12 existing Black Stories, with deviations in details included. Each Black Story was tested twice in the human and GPT-4 group to minimize individual differences. The experiment was conducted via text messaging to align the testing set-up for the two groups and eliminate potential non-verbal advantages for the human test group. The primary performance indicator was the number of questions needed to solve the riddle, considering the number of given hints to come to the solution. This measure indicated no significant difference between the groups, where both groups managed to arrive at the correct answer eventually. Though GPT-4 was significantly more verbose in questioning than humans, and qualitative results showed that GPT-4 excelled in precise questioning and creativity, but often fixated too much on details. This led to missing the bigger picture and summarizing solutions prematurely. On the other hand, humans covered broader topics and adapted their focus quickly, but had more difficulty figuring out uncommon details. This research suggests that the performance of GPT-4 and humans in solving Black Stories is not significantly different, despite using alternative approaches to achieve results.

1. Introduction

Large Language Models (LLMs) are artificial intelligence (AI) programs designed to process and generate text. They achieve this by learning patterns and structures from extensive datasets containing large collections of texts (Brown et al. 2020). Recent advancements have demonstrated the remarkable success of these models across a variety of linguistic tasks, showcasing their capacity to generate coherent, contextually relevant, and nuanced responses (Wang et al. 2019). 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 understand and generate natural language. While LLMs can produce text that seems entirely plausible and logical, they sometimes generate incorrect or nonsensical answers, often referred to as 'hallucinations' (Bang et al. 2023, Augenstein et al. 2024). This suggests that some limitations in their reasoning processes can occur.

Learning more about the logical reasoning abilities of LLMs is beneficial for making advancements in artificial intelligence applications. The extent to which these models exhibit genuine reasoning abilities remains a subject of intense scholarly debate (Bender and Koller 2020, Mitchell and Krakauer 2023, Mahowald et al. 2024, Chang and Bergen 2024).

To explore the extent to which LLMs can understand logical reasoning problems, researchers have designed benchmarks that can be used to test whether these models can simulate human reasoning abilities through logical question answering (Cheng et al. 2025). Logical reasoning involves drawing conclusions or judgments based on evidence or prior experience. As reviewed by (Cheng et al. 2025), such tasks often involve presenting a natural language description of a premise and some constraints, leading to a question that can only be answered correctly through applying deductive (drawing conclusions based on formal logic), inductive (generalizing from past observations) or abductive (generating the most plausible hypothesis given some evidence) reasoning steps. 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 (Cheng et al. 2025).

Another task that requires both logic and creative thinking and is linguistic in nature, is solving riddles. Riddles are cognitive puzzles, often presented in the form of a question, statement, or phrase, designed to challenge the solver's ingenuity and out-of-the-box thinking. They are usually phrased ambiguously or metaphorically, requiring the solver to interpret clues creatively rather than literally. As reviewed below in Section 2, several benchmarks for riddle solving have also be tested on LLMs, where isolated riddles are presented as either open or multiple-choice questions. LLMs also struggle with solving such riddles (Lin et al. 2021, Jiang et al. 2023, Del and Fishel 2023).

We propose to evaluate the reasoning and problem-solving abilities of LLMs by using the game "Black Stories". Black Stories are riddles that describe mysterious and often dark scenarios. At least two players are required to play this game. One player reads a brief cryptic description of the ending of the story out loud and knows the full story. The other players ask questions to piece together the story. The goal of the game is to uncover the full story by asking yes-or-no questions, which will be answered solely with 'yes' 'no', 'false assumption', or 'not relevant'. An example of a Black Stories riddle is the following, titled "The deadly punch": "A man went into a party and drank some of the punch. He then left early. Everyone at the party who drank the punch later died of poisoning. Why did the man not die?". Solving this riddle requires complex reasoning steps. For example, abductive reasoning is necessary to generate possible explanations based on the available evidence, such as considering that the man left early, so maybe something happened after he left and that the poison might have been added later or activated over time. Inductive reasoning is needed to recognize common patterns and use known facts, like the fact that poisonings usually occur due to contaminated drinks/food, so it's logical to assume the punch contained the poison. Finally, deductive reasoning helps to eliminate unlikely possibilities and arrive at the conclusion, like: "If the poison had been in the punch from the start, he would have died too. Therefore, the poison must have been introduced after he left. Since everyone who drank the punch later died, the poison must have been slow-acting or added after his departure.". Through asking the right questions and making such inferences, the solver will finally figure out that the poison was inside the ice cubes. The man drank the punch when the ice was still frozen, so he didn't ingest the poison. Everyone else who drank the punch later consumed the fully melted, poisoned liquid and died.

Looking into how LLMs handle such riddles can provide insights into their reasoning abilities in an *interactive setting*. It allows for assessing whether such a model can effectively mimic human reasoning strategies in complex tasks with unusual contexts. Furthermore, it makes it possible to compare the strategies of LLMs with those of humans. Since GPT-4 often outperforms other LLMs, which will be further discussed in the related work (Section 2), we focus on this state-of-the-art model, to answer the following main research question: *How do the logical reasoning abilities of GPT-4 compare with those of humans when solving Black Stories?*.

Given previously found challenges in the reasoning of LLMs (e.g. (Cheng et al. 2025); and see Section 2) and 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.

This paper provides an overview of relevant work related to the cognitive abilities of LLMs and related to reasoning for solving riddles (Section 2). The methodology (3) and experiment (4) sections describe the experimental setup, including the use of Black Stories, the evaluation criteria, and how GPT-4's performance will be compared to that of human participants. Results will present the findings of the study in Section 5, which will then be further interpreted in the discussion Section 6. A conclusion will be drawn 7, and the limitations of the study will be addressed. Finally, suggestions for further research will be provided in Section 8.

2. Related work

This section reviews existing research on the cognitive abilities of LLMs and their comparison to human reasoning. Furthermore, it discusses logical reasoning and riddle-solving in general, and what led to the idea of this current study.

2.1 Cognitive Abilities of LLMs

Cognitive abilites have been studied in LLMs within a wide variety of different domains. For example, research has explored the extent to which LLMs can perform Theory of Mind (ToM) tasks (van Duijn et al. 2023). ToM refers to the capacity to ascribe mental states to oneself and others. It is the cognitive ability to understand that others have their own thoughts, beliefs, desires, and emotions, which may differ from one's own. In their research, van Duijn et al. (2023) evaluated 11 state-of-the-art models against children aged 7-10 on advanced ToM tests. The study found that most current LLMs perform below the level of children aged 7-10 on standardized ToM tasks. However, the largest models with extensive instruction tuning, such as GPT-4, outperform children and surpass the other models (van Duijn et al. 2023).

Furthermore, Binz and Schulz (2023) explored how fine-tuning LLMs with data from psychological experiments enables these models to accurately predict human behavior in decision-making tasks (Binz and Schulz 2023). This suggests their potential to represent and predict human behavior.

Yax et al. (2024) explore how LLMs apply cognitive insights in reasoning and decision-making processes. This study compares reasoning in humans and LLMS using cognitive tests designed to measure biases in decision-making processes. They analyzed how humans and models differ in their reasoning patterns when solving problems that are designed to elicit responses that are intuitive for humans but happen to be incorrect. While people rely on a mix of intuitive and analytical reasoning, LLMs process information through statistical pattern recognition. Despite this fact, the errors LLMs made resembled those made by humans. It was also found that newer LLMs, such as GPT-4, outperformed humans in accuracy and reduced biases. However, LLMs are more responsive to explicit prompting, whereas humans adapt their reasoning more flexibly based on context (Yax et al. 2024).

Additionally, studies have investigated the problem-solving abilities of LLMs. Orrù et al. (2023) concluded that models like ChatGPT can match average human performance in verbal insight tasks (Orrù et al. 2023).

Complex reasoning skills are also widely studied using programming tasks and mathematical problem-solving (Roumeliotis and Tselikas 2023, Huang and Chang 2023, Liu et al. 2024). LLMs have demonstrated to be able to answer questions on mathematical world problems and commonsense reasoning quite well when prompted to follow a series of reasoning steps, following given examples of chain-of-thought reasoning (Wei et al. 2022). These findings indicate that there is potential for creative problem-solving in LLMs and that there is potential for AI to emulate human cognitive processes when the model is appropriately trained.

Not all previous findings are as optimistic. A closer examination of semantic representations encoded by language models revealed no evidence that these models can distinguish basic logical symbols (e.g., AND vs. OR) (Traylor et al. 2021), for example. Moreover, McCoy et al. (2024) demonstrated that LLM's performance on a variety of tasks highly depends on the probability of the question and answer given the training data. Without any difference in the complexity of the question, LLMs will fail more often on rare questions or when the correct answer is rare than on frequent questions and answers (McCoy et al. 2024). Therefore, despite a large amount of work on cognitive abilities and reasoning problems in LLMs, evidence of real understanding of such reasoning tasks is lacking.

2.2 Reasoning and Riddle-Solving

Logical reasoning is important for solving riddles. 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, which suggests that riddles test flexibility in reasoning (Bar-Hillel et al. 2018).

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 (Gwaravanda and Masaka 2008). 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 (Absattarovna 2021).

Riddles can be used to test advanced cognitive and language skills in LLMs. Lin et al. (2021) created the RiddleSense dataset to evaluate creativity and commonsense reasoning in models. Solving riddles requires skills like thinking about "what if" scenarios, understanding metaphors, and interpreting creative language. For example, interpreting "I have five fingers but am not alive" as "glove" requires an understanding of metaphoric descriptions of everyday objects. Humans had a much higher accuracy (91.3%) for solving riddles than the best language model (68.8%) (Lin et al. 2021).

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 (Jiang et al. 2023).

Slightly more similar to our method, the True Detective task assesses abductive reasoning with stories as input (Del and 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.

Together, these studies show that riddles are effective sources to aid in testing complex reasoning skills. LLMs, however, have not been able to match humans in solving these puzzles.

2.3 Evaluating LLM Reasoning Abilities in an Interactive Story Game Context

While impressive performance has been demonstrated in various tasks, language models still perform worse than humans in tasks that combine logic and creativity. Black Stories provide a unique way to explore these abilities further.

The context of a story, or narrative, is proposed to be a fundamental structure of human meaning-making (Bruner 2004) and stories have been demonstrated to help humans in problem-solving tasks (Hernandez-Serrano and Jonassen 2003), as well as in other areas of cognition such as memorization, decision-making, planning and improvising (Schank and Berman 2003). Stories help in reasoning tasks because they provide structure, context, and meaning, making complex information easier to understand and remember. For example, children showed enhanced false belief reasoning when they engaged with a story version instead of a traditional false belief task, because they could reconstruct the sequence of events more effectively in the context of a story (Lewis et al. 1994). A recent study suggested that stories can in fact help LLMs with reasoning as well (Javadi et al. 2024): Complex physics, chemistry, math, and biology questions are solved with higher accuracy by LLMs when the prompt includes a narrative providing a structured context for the problem domain. We therefore suggest that the story context of the Black Stories game may provide a fruitful basis for testing reasoning abilities in LLMs.

Unlike most riddle-solving tasks, this context does not involve asking the LLM a series of unrelated complex questions but invites the model to generate relevant questions to solve a riddle embedded in a larger story. These riddles require solvers to rebuild narratives by asking yes-or-no questions, therefore 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, logical reasoning to rule out scenarios, and attention to detail for interpreting subtle hints in the initial riddle and clues revealed during questioning, building on the information already uncovered. It requires reasoning on all three levels: deductive, inductive and abductive reasoning. 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.

3. Methodology

3.1 Black Stories dataset

Twelve Black Stories were selected from the English version of the board game Black Stories (Bösch and Andersen 2007), see also Appendix A. We selected these stories based on the same level of difficulty and length. Each Black Story was adjusted to a deviation from the story to prevent it from being recognized by the LLM. This was preferable over creating entirely new stories, as this allowed for using unique stories while maintaining consistency from the original game and efficiency within the time constraints of the research. Table 1 shows an example of a deviation of one selected story. Appendix A provides a detailed overview of the adjustments of each story. Key components to guess were extracted from each solution as shown in Table 1.

3.2 Analysis Methods

The analysis of the collected data included several factors:

- **Question count**: We determined the question count as our main pillar of success in solving riddles. Fewer questions needed to get to the correct answer suggest that the solver is asking precise and targeted questions, while systematically narrowing down possibilities. The fewer questions asked, the more likely the solver is thinking strategically, indicating a strong ability for logical reasoning (Bang et al. 2023).

- **Hints**: Solutions of Black Story Riddles are often complicated. Similar to the original Black Story game, the experimenter could give hints to guide the solvers in the right direction. Therefore, the hint count was included in our analysis. The solvers could either ask for a hint themselves, or the experimenters would intervene with a hint if the solution given was insufficient, repetition was noticed, or if the questions asked would not lead to the solution. This also resembles the way the original Black Stories game is designed (Quoting the instructions of the game: "If nobody in the group is getting even close to the solution, and if he/she (Riddle Master) so decides, the Riddle Master may offer a hint or two"). Since LLMs don't have inherent mechanisms for independently asking for hints, the experimenters only intervened themselves to guide the model in the right direction. Additionally, a calculated weight was multiplied by

Type	Original story	Deviation
The description given to participant & LLM	A stark naked man was found dead at the foot of a mountain - with a matchstick in his hand.	A naked couple was found dead in a forest - with a dice in their hands.
Solution of Black Story	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.	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.
Components to guess		* 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 deviation of Black Story description and solution.

each given hint and added to a corrected question count, defined by the variable *score*. The weight was calculated as

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

In this equation, $\#Q$ represents the number of questions asked without any hints, $\#QH$ represents the average number of questions asked when hints are provided, and $\#H$ is the average number of hints given when hints are provided. The equation is used to determine the equivalent value of each hint in terms of the number of questions, showing how much each hint effectively reduces the number of questions needed. Subsequently, the final score was calculated as

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

where each given hint is multiplied by the calculated weight, adding up to the total question count.

- **Word count**: The average number of words in a single question was calculated to investigate the detailedness of questions and elaboration on a specific setting. This average count was further averaged over all the questions in a single story for one participant/ one round of the LLM.

- **Qualitative analysis**: Additional qualitative analyses were done to investigate other interesting patterns related to logical reasoning, such as focus switching between various areas of the solution and the revelation of details to derive the final solution.

4. Experiment

4.1 Experiment Part A (Alive)

4.1.1 Participants

16 human 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 commitment to the research. Our main criteria included fluency in English and prior knowledge of Black Stories. Fluency in English was required since we used the English version of Black Stories. Prior knowledge of Black Stories was needed to simulate the prior knowledge of our used LLM and ensure a fair comparison of logical reasoning, rather than comparing with figuring out how the game works in the first place.

4.1.2 Procedures

The experiment was conducted via a private WhatsApp conversation with one of the experimenters and the participant. The choice to do this experiment via text messaging rather than an actual conversation was made to mimic the conditions of GPT as much as possible and eliminate the advantages humans might have over LLMs as much as possible, such as non-verbal communication and intonation. The participants were initially instructed with a text message before the experiment (Table 2). The instructions also specified the average duration of the experiment, the language used, the format of asking one question at a time, and the fact that participants could explicitly request hints. Additionally, it was stated that hints were sometimes provided in the experimenter's judgment to help guide participants in the right direction. This was given, following the original instructions of the Black Stories game (see section 3.2), when participants were repeatedly asking questions concertedly from the solution, or repetition in the details of the questions.

The experiment started with the description of one of the deviated Black Stories as shown in Table 1. Random stories were assigned to experimenters and participants to eliminate systematic biases. If the participant was familiar with a story, a different story from the 12 available stories was chosen. This was done by asking the participant to indicate if the story's description was familiar to the participant. Subsequently, the participants started with questioning until they solved the riddle according to the main extracted components of the solution. Depending on the participant's availability, one or two Black Story Riddles were asked. Completing one story of the experiment took approximately 30-45 minutes. After completing the experiment, the WhatsApp conversation was exported to a .txt file for further analysis in Python. We aimed to test each story 2 times to minimize individual biases and to level out as much variation as possible. One story was only tested once due to time constraints. In total, 23 data points of 12 different stories were collected.

4.2 Experiment Part B (Bot)

4.2.1 LLM

GPT-4 was selected as the pre-trained model for solving Black Story Riddles. This model was accessed via the OpenAI API [1]. Standard temperature and top_p settings were used.

1. `https://platform.openai.com/docs/models/overview#gpt-4-turbo-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 led 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 with yes, no or not relevant to the story. You will use my answers to solve the riddle and find the story that leads 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."

Table 2: Instructions given to both GPT-4 and the human participants

4.2.2 Procedures

The experiment was conducted via Python's terminal. For GPT-4, the same experimenter was responsible for the same stories, ensuring consistency in administration. System content was given to GPT-4 as shown in Table 2. The experiment started with the description of one of the deviated Black Stories as the example shown in Table 1. One story we selected was recognized by GPT-4 and therefore replaced with another story. This was noticed because, within the first 3 questions, the model immediately asked for specific details of the solution, while including details in those questions already that they had not yet verified (see also Appendix B). These early inquiries with detailed information indicated when GPT-4 drew conclusions based on prior knowledge of the story, and they clearly contrasted with the broader types of questions asked when the model did not have such prior knowledge. We eliminated the recognized story from the dataset. For the other stories, the model started with questioning and continued until they solved the riddle according to the predefined key components. Additionally, one story was conducted three times, such that a total of 25 data points from 12 different stories were collected.

Both humans and GPT-4 received hints when needed. While humans can actively decide when they want help, this is not the case for GPT-4. Instead, hints were given by the experimenter's judgment to guide the model in the right direction. This judgment was based on preliminary summaries that were insufficient, repetition in the details of the questions, or the moment humans would normally ask for hints. Since GPT-4 lacks awareness of uncertainty in the way humans do (OpenAI 2023, Vaswani et al. 2017), they will always provide a next question or summary that is most likely instead of asking for additional hints. That is why hints were provided to GPT-4 solely based on the experimenter's judgment, whereas humans received hints both through the experimenter's judgment and by requesting them specifically (see Section 4.1.2).

When the solution was sufficient, the conversation was ended and saved to a data file. Additionally, for calculating the overall weight, the experiment was conducted once more with one story without giving GPT-4 any hints. This weighting factor was applied to calculate the final score for both GPT-4 and humans, because of reasonably similar structures of the given hints to both groups (see Table 7 for illustration).

5. Results

The final dataset, including the conversations of both human participants and GPT-4 across the twelve Black Stories, is fully accessible and can be obtained via an OSF repository[2].

2. https://osf.io/bs2zx/files/osfstorage/67ed4d0d365f53ce266de173

5.1 Quantitative analysis

5.1.1 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), see Figure 1. A student's t-test showed no significant difference between group means $t(46) = 0.253$, $p = 0.801$.

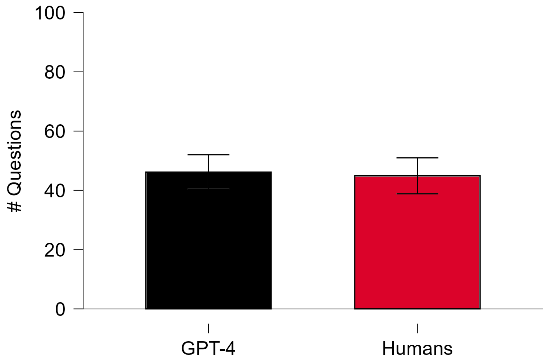

Figure 1: Average question count across all Black Stories per group.

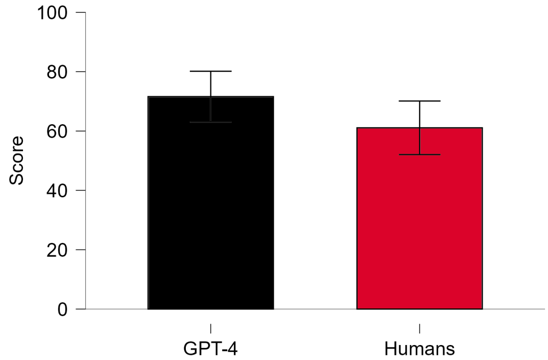

Figure 2: Average score across all Black Stories per group. The score is calculated with Equation 2. The score involves correction for given hints, where the weight of one hint is equal to asking six questions.

5.1.2 Score

Using Equation 1, the calculated weight was set to 6 based on GPT-4's question count of 59 for Story 1 without hints, compared to an average of 35 questions when the story was conducted three times with an average of 4 hints. Taking this weighting factor into account, the difference in score between GPT-4 and humans across all Black Stories was analyzed using an independent t-test. Assumptions of normality and variance were met. 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 lower average score than GPT-4 (M=71.6, SD=25.0), see Figure 2.

5.1.3 Hint count

The difference in score displayed in Figure 2 shows a slightly better score for humans, although insignificant. However, when comparing these results with the average question count, hints seem to impact the question count. 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) needing more hints on average than humans (M=2.7, SD=2.1).

Additionally, an ANOVA test was conducted to test whether the number of hints given to GPT-4 differentiated between experimenters due to its relatively 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.

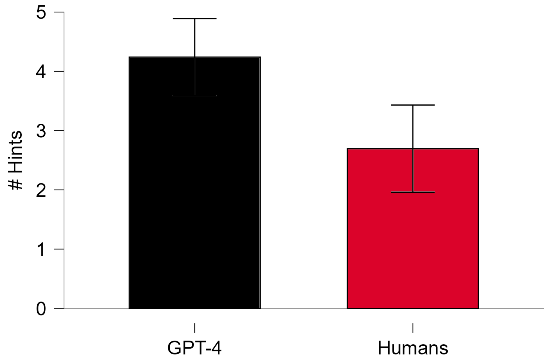

Figure 3: Average number of hints needed per group

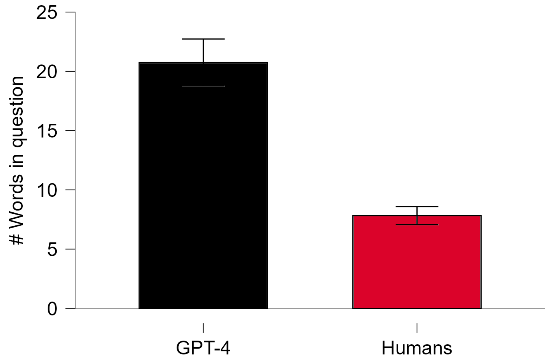

Figure 4: Average word count of a sentence per group

5.1.4 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 4, GPT-4 was using longer sentences (M=20.7, SD=4.8) than humans (M=7.8, SD=1.8), including more details in the sentences (see Table 4) and summarizing the solution more often than humans (see Table 3).

5.2 Qualitative analysis

In addition to the quantitative findings, the dialogues were analyzed for more insights about the strategies partaken by both participant types.

Sticks to one detail vs. umbrella questions The first notable result in this analysis was the difference in detailed questions given by GPT-4 as opposed to humans. The sentences provided by GPT -4 were more detailed, and multiple subjects were occasionally proposed in one question. An example of umbrella questioning of GPT-4 as compared to human questioning is given in Table 4.

Identifying specific settings Moreover, the detailed questions given by GPT-4 often guessed unusual and specific settings, whereas the human participants struggled more with discovering the creative settings of the story (Table 5).

Humans cover more topics and switch focus faster to other details The dialogue analysis further shows that humans were more capable of switching their focus in the stories. In contrast to the LLM, which often explored further specific aspects of the storyline (see Table 6).

Emotions and affirmation humans. Moreover, results from this dialogue analysis indicate that human participants often experience emotional responses when frustrated or motivated by this 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.

Hints given by experimenters. Furthermore, additional textual analysis on the type of hints that were given showed similar structures across experimenters and groups, often including a question to elaborate on a specific topic or guidance to ask about something specific (see Table 7).

GPT-4	Humans
[. . .]	[. . .]
so to summarize ... concept for her	did she have a dog
did the woman check if she could see while she was inside the dark tunnel	and she was all alone in the bus
in summary the ... inside the tunnel	i would get a dog but alright
did she jump out of the bus in panic of the unknown fearing the possibility that her vision was still impaired despite the surgery	not for someone something or some particular event right
to sum up ... this tragic outcome	did she expect ... than reality brought
[. . .]	was she picking something up or dropping something of
	and she needed days in city b
	to do the thing she did
	something she wanted to experience
	something to do with her eyes
	something to make her eyes heal
	get surgery
	she did something that made her see things again and when she went through the tunnel she thought she became blind again and life wasnt worth living anymore so she killed herself

Table 3: Excerpts of question types from Story 10 by GPT-4 (left) with more consecutive summaries, compared to humans (right). As illustrated in this Table, GPT-4 often gives a summary, then asks the question, followed by another summary. On the other hand, human participants typically ask multiple questions and provide a summary only at the end.

These examples illustrate that there weren't remarkable differences in the structure of given hints across experimenters and groups that could have influenced the final score.

6. Discussion

The potential capabilities of LLMs to solve complex problems are a topic of considerable attention currently, and the question of whether LLMs can reason in ways analogous to human thinking remains a subject of ongoing debate (Mitchell and Krakauer 2023). Most studies assess such abilities in models by applying benchmarks that contain isolated problems in the form of either open or multiple-choice questions (Cheng et al. 2025). Here, we introduced an alternative method in which the LLM *asks* instead of answers the questions, to solve a riddle *interactively* in the context of a larger story. While LLMs have so far exhibited underwhelming performance on current reasoning

GPT-4	Humans
[...] did the other person use the water bottle to cause the woman's death did the death result from a physical altercation with the other person was the cause of death related to the sauna's heat or steam did the other person poison the woman did the other person do something to the sauna equipment to cause her death did the woman's death result from a preexisting medical condition that was triggered in the sauna did the other person lock the woman in the sauna causing her death by overheating or dehydration did the other person tamper with the water bottle in a way that led to the woman's death 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 woman die from natural causes did the presence of the other person in the sauna cause the woman to have an accident did the woman die from an allergic reaction did the death involve electrocution [...]	[...] did she suffocate did she drown was she trapped was she poisoned was she shot was she hit with an item was she stabbed [...]

Table 4: Excerpts of detailed questions from Story 6 by GPT-4 (left) and the human group (right), showing that GPT-4's questions involve various scenarios and causes of death, while the human questions focus on specific physical causes.

benchmarks, we hypothesized that the context provided by both the story and the situated interaction in our set-up might help LLMs in solving complex reasoning tasks. In the conducted experiment, where the performance of GPT-4 was compared to that of humans, the number of questions needed and the average score to finish the story were measured. Notably, these results did not differ significantly between GPT-4 and the human participants. While GPT-4 achieved a slightly higher average score in solving the riddles, the difference was not statistically significant. Our findings indicate that the LLM model is able to solve Black Stories riddles, a task that requires complex cognitive abilities including deductive, inductive, and abductive reasoning skills.

GPT-4
[...] are they trapped in a cold desolate place like the arctic or antarctic regions are they trapped in a desert with no hope for rescue and no way of getting food or water are they stranded in space like on a spaceship or a space station with no hope of rescue are they trapped in a jungle or forest with no means of communication or ability to navigate their way out are they trapped in a postapocalyptic setting without any survival resources [...]

Table 5: Excerpt of GPT-4's questions from Story 1, identifying specific settings.

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

Table 6: Excerpts of questions from Story 1 by GPT-4 (left) and the human group (right), showing that the human group switches focus faster and covers more topics.

When examining the difference in hints required to solve the riddles, humans needed fewer hints. During the experiments, it became clear that GPT-4 often required additional hints to provide more precise answers to the riddles. 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, such as "...unrelated to the sauna environment, the water bottle, and not directly physical, such as a scare or shock" (see Table

GPT-4	Humans
no, ask about the occupation of the dead man	Ask about the occupation of the other person involved
no, ask about where she was before she was in the coffin	Ask about where she was before
but what did he need an alibi for	yes. But why did it immediately expose his alibi?
no try to guess who the podcast host was	How did he prerecord it
yes, but why are they doing this?	What were they doing together?
yes but what was the confined space?	Ask about the location
but why did he order coke?	Ask about why he orders something at the bar
no, ask about the function of the airplane	Ask about the type of vehicle that is involved
but what kind of situation were they in	Why were they traveling?

Table 7: Type of hints given by experimenters to GPT-4 (left) and Humans (right) across all stories. Every row shows similarities in hints across the two groups.

4). In such cases, hints were provided to redirect GPT-4 toward the correct path, as it was close to the answer but not entirely accurate. Human participants, on the other hand, were more likely to resist the use of hints, persisting until they were unable to deduce the answer independently. A possible explanation for this different behavior could be the difference in emotional thinking. Humans tended to react more emotionally regarding pride or self-reliance, motivating them to solve the riddle without any help. In contrast, hints for GPT-4 were provided in the examiner's observations, as the model cannot experience emotions or prioritize autonomy.

Moreover, the number of words used to formulate questions was analyzed and compared between GPT-4 and the human participants. On average, GPT-4 used more than twice the number of words as human participants when formulating questions. Interestingly, this verbosity gave GPT-4 an advantage during the experiments, as its questions were more detailed and covered a broader range of facts from the riddles. This aligns with GPT-4's design, which is extensively instruction-tuned to be conversational and optimize informativeness in its answers to increase user comprehension while avoiding miscommunication. In contrast, humans tended to focus on formulating concise and straightforward questions. Despite these differences in approach, both methods were effective, as the overall performance of both participant types was comparable.

The difference in strategies for formulating questions was investigated further in a dialogue analysis to better understand the annotation techniques used by each participant type. As previously noted, GPT-4's questions were generally more detailed. In some cases, this led to a strategic advantage, as GPT-4 managed 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 4). However, where GPT-4 had its advantage in its detailed questions and elaborate inductive reasoning, it 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. This way, GPT-4 required more hints and more questions to get back to solving the entire story, instead of one detail. In contrast, human participants demonstrated a stronger ability to shift focus between topics. As seen in Table 6, humans explored a wider range of storyline aspects, ultimately getting to the correct answer more efficiently.

Our findings resonate with previous observations of differences between human and LLM reasoning. As mentioned in Section 2, Yax et al. (2024) compared human and LLM performance on particular reasoning tasks designed to elicit responses that are intuitive to humans but incorrect. Although LLMs rely on statistical patterns instead of human intuition to solve these tasks, both participant types made similar mistakes. However, the performance of LLMs could be improved through applying different prompting methods, while this did not work for human participants, who adapt their reasoning more flexibly to context. Similarly, our findings suggest that GPT-4 matched human performance in solving Black Stories but used a detail-focused approach, struggling to shift focus, while humans adapted more flexibly.

Overall, the analysis shows promising potential for playing riddle games with GPT-4, as the model's performance does not differ significantly from the human players. However, the difference lies in the formulation of the questions and the strategic approach to solving the riddles. Interestingly, GPT models are trained to provide answers instead of asking questions themselves. This role switch could interfere with the purposes which the LLM is trained for, leading to the more detailed questions asked and the tendency to provide an answer after solving a minor part of the story.

7. Conclusion

This study explored the problem-solving abilities of GPT-4 in black story riddles, comparing its approach to that of human participants. In both groups, participants were able to eventually get to the right solution for every riddle. Looking at the score, the findings showed no significant difference between the performance of GPT-4 and humans, indicating that there is insufficient evidence to assert a meaningful distinction between the two groups. However, qualitative analysis revealed differences in problem-solving strategies. Specifically, GPT-4 applies a detailed and extensive exploration, whereas humans tend to be 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.

In addressing the question, "How do the logical reasoning abilities of GPT-4 compare with those of humans when solving Black Stories?", this study therefore suggests that GPT-4 and humans are both capable of reaching the correct solution, but exhibit different tactics in solving Black Stories riddles.

8. Limitations and future work

Our work has a few notable limitations. Unlike humans, GPT-4 was not specifically designed or prompted to independently decide when to ask for hints. In our experiment, it was therefore decided by the experimenters 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, as well as the calculated score since this measure takes the number of needed hints into account. However, this did not appear to have any consequences for our overall findings since, even with a heightened use of hints, the score of the GPT-4 group still did not differ from that of the human group significantly and both human and GPT-4 participants were able to recover the right story in all trials. Moreover, there was no significant difference present in how many hints were given to the model between the experimenters and qualitatively the structure of the hints given to GPT-4 and humans was similar. This shows that, although there was possibly some subjectivity present, the experimenters aligned well in their recognition of typical situations in which hints were needed. 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.

Additionally, the effectiveness of the hints was not measured, limiting insights into the influence of the hints on riddle-solving performance. In our experiment we followed the instructions of the original Black Stories game as closely as possible, including the possibility for the Riddle Master to provide a hint when needed. Future work, however, could include a baseline experiment where participants and GPT-4 solve riddles without any hints, allowing for a clearer assessment of the impact of 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.

In addition, GPT-4 was initially built for providing informative responses, rather than concise questions. The observed behavior reflects this primary function. Even though this did not significantly hinder its performance, it shaped its unique approach to the game. For future work, it would be interesting to use an LLM that has been specifically designed to ask questions instead of providing information. This model adjustment can test the hypothesis that, indeed, the fact that the current GPT-4 model is specialized in providing long and informative answers is what interferes with approaching this game in a more human-like concise way.

Furthermore, we currently only focus on one specific state-of-the-art model, but the same study can be applied to other model families and novel models that will be released in the future. The effect of different model architectures and training regimes can be more directly studied in this way. For example, it can be investigated whether any further enhancements in making LLMs more informative, interferes more deeply with their performance in solving riddles or makes them better at tackling these riddle scenarios.

Given the rapid advancements in LLM models, future research could benefit from a dedicated team designing new riddles specifically for research, guaranteeing that the model has never encountered the story before. This would require a creative set of skills to balance real-world knowledge, out-of-the-box thinking, and solvable mysteries, ensuring that participants remain engaged with the game without becoming demotivated by riddles that are too easy or too difficult to solve.

Future studies could also 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 as well as in problem-solving tasks more generally.

9. Acknowledgement

We want to thank all participants who voluntarily participated in this study. Furthermore, we would like to thank the reviewers for their constructive feedback and insights.

References

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

Augenstein, Isabelle, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, et al. (2024), Factuality challenges in the era of large language models and opportunities for fact-checking, *Nature Machine Intelligence* **6** (8), pp. 852–863, Nature Publishing Group UK London. https://doi.org/10.1038/s42256-024-00881-z.

Bang, Yejin, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung (2023), A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity, *in* Park, Jong C., Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi, editors, *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Nusa Dua, Bali, pp. 675–718. https://aclanthology.org/2023.ijcnlp-main.45/.

Bar-Hillel, Maya, Tom Noah, and Shane Frederick (2018), Learning psychology from riddles: The case of stumpers, *Judgment and Decision Making* **13** (1), pp. 112–122. https://www.cambridge.org/core/journals/judgment-and-decision-making/article/learning-psychology-from-riddles-the-case-of-stumpers/CF8E79596B9B426A62B57207BF49F5B0.

Bender, Emily M and Alexander Koller (2020), Climbing towards nlu: On meaning, form, and understanding in the age of data, *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 5185–5198.

Binz, Marcel and Eric Schulz (2023), Turning large language models into cognitive models. https://arxiv.org/abs/2306.03917.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei (2020), Language models are few-shot learners, *Advances in Neural Information Processing Systems*, Neural information processing systems foundation. https://arxiv.org/abs/2005.14165v4.

Bruner, Jerome (2004), Life as narrative, *Social Research: An International Quarterly* **71**, pp. 691–710. https://dx.doi.org/10.1353/sor.2004.0045.

Bösch, Holger and David Andersen (2007), *Black stories : solve 50 creepy mysteries*, 21th ed., Moses.

Chang, Tyler A. and Benjamin K. Bergen (2024), Language model behavior: A comprehensive survey, *Computational Linguistics* **50** (1), pp. 293–350, MIT Press One Broadway, Cambridge, Massachusetts 02142, USA. https://doi.org/10.1162/coli_a_00492.

Cheng, Fengxiang, Haoxuan Li, Fenrong Liu, Robert van Rooij, Kun Zhang, and Zhouchen Lin (2025), Empowering llms with logical reasoning: A comprehensive survey. https://arxiv.org/abs/2502.15652.

Del, Maksym and Mark Fishel (2023), True detective: A deep abductive reasoning benchmark undoable for gpt-3 and challenging for gpt-4, *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (* SEM 2023)*, pp. 314–322. https://aclanthology.org/2023.starsem-1.28/.

Gwaravanda, Ephraim and Dennis Masaka (2008), Shona reasoning skills in zimbabwe: The importance of riddles, *Journal of Pan African Studies, 2008, Vol 2, Issue 4, p234*. https://www.jpanafrican.org/docs/vol2no5/2.5_Shona_Reasoning_Skills_in_Zimbabwe.pdf.

Hernandez-Serrano, Julian and David H. Jonassen (2003), The effects of case libraries on problem solving, *Journal of Computer Assisted Learning* **19** (1), pp. 103–114, Wiley Online Library. https://doi.org/10.1046/j.0266-4909.2002.00010.x.

Huang, Jie and Kevin Chen-Chuan Chang (2023), Towards reasoning in large language models: A survey, *in* Rogers, Anna, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, Association for Computational Linguistics, Toronto, Canada, pp. 1049–1065. https://aclanthology.org/2023.findings-acl.67/.

Javadi, Vahid Sadiri, Johanne R Trippas, Yash Kumar Lal, and Lucie Flek (2024), Can stories help LLMs reason? Curating information space through narrative, *The First Workshop on System-2 Reasoning at Scale, NeurIPS'24*.

Jiang, Yifan, Filip Ilievski, Kaixin Ma, and Zhivar Sourati (2023), BRAINTEASER: Lateral thinking puzzles for large language models, *in* Bouamor, Houda, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Singapore, pp. 14317–14332. https://aclanthology.org/2023.emnlp-main.885.

Lewis, Charlie, Norman H. Freeman, Charlotte Hagestadt, and Heather Douglas (1994), Narrative access and production in preschoolers' false belief reasoning, *Cognitive Development* **9** (4), pp. 397–424. https://www.sciencedirect.com/science/article/pii/0885201494900132.

Lin, Bill Yuchen, Ziyi Wu, Yichi Yang, Dong-Ho Lee, and Xiang Ren (2021), RiddleSense: Reasoning about riddle questions featuring linguistic creativity and commonsense knowledge, *in* Zong, Chengqing, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Association for Computational Linguistics, Online, pp. 1504–1515. https://aclanthology.org/2021.findings-acl.131/.

Liu, Wentao, Hanglei Hu, Jie Zhou, Yuyang Ding, Junsong Li, Jiayi Zeng, Mengliang He, Qin Chen, Bo Jiang, Aimin Zhou, and Liang He (2024), Mathematical language models: A survey. https://arxiv.org/abs/2312.07622.

Mahowald, Kyle, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko (2024), Dissociating language and thought in large language models, *Trends in cognitive sciences*, Elsevier.

McCoy, R. Thomas, Shunyu Yao, Dan Friedman, Mathew D. Hardy, and Thomas L. Griffiths (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), pp. e2322420121, National Academy of Sciences. https://doi.org/10.1073/pnas.232242012.

Mitchell, Melanie and David C Krakauer (2023), The debate over understanding in ai's large language models, *Proceedings of the National Academy of Sciences* **120** (13), pp. e2215907120, National Academy of Sciences.

OpenAI (2023), Gpt-4 technical report, *arXiv*. https://arxiv.org/abs/2303.08774.

Orrù, Graziella, Andrea Piarulli, Ciro Conversano, and Angelo Gemignani (2023), Human-like problem-solving abilities in large language models using chatgpt, *Frontiers in Artificial Intelligence*. https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2023.1199350.

Roumeliotis, Konstantinos I. and Nikolaos D. Tselikas (2023), Chatgpt and open-ai models: A preliminary review, *Future Internet*. https://www.mdpi.com/1999-5903/15/6/192.

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

Traylor, Aaron, Roman Feiman, and Ellie Pavlick (2021), And does not mean or: Using formal languages to study language models' representations, *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. https://aclanthology.org/2021.acl-short.21/.

van Duijn, Max, Bram van Dijk, Tom Kouwenhoven, Werner de Valk, Marco Spruit, and Peter van der Putten (2023), Theory of mind in large language models: Examining performance of 11 state-of-the-art models vs. children aged 7-10 on advanced tests, *in* Jiang, Jing, David Reitter, and Shumin Deng, editors, *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, Association for Computational Linguistics, Singapore, pp. 389–402. https://aclanthology.org/2023.conll-1.25.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin (2017), Attention is all you need, *in* Guyon, I., U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, Vol. 30, Curran Associates, Inc., pp. 6000–6010. https://dl.acm.org/doi/10.5555/3295222.3295349.

Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman (2019), Superglue: A stickier benchmark for general-purpose language understanding systems, *Advances in Neural Information Processing Systems.*

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V. Le, Denny Zhou, et al. (2022), Chain-of-thought prompting elicits reasoning in large language models, *Advances in neural information processing systems* **35**, pp. 24824–24837.

Yax, Nicolas, Hernán Anlló, and Stefano Palminteri (2024), Studying and improving reasoning in humans and machines, *Communications Psychology* **2** (1), pp. 51. https://doi.org/10.1038/s44271-024-00091-8.

Appendix A. Full list of 12 used Black Stories and their deviations with description (white) and solution (grey)

Original	Deviation
1	
Two dead men are seated at a table. On the table between them is a game of chess and a pistol.	A man and a woman are seated at a table. On the table between them is a game of checkers and a pistol.
The two men were in a submarine stranded on the ocean floor. There was enough air for only a few hours. There was only one bullet in the gun so they played chess to see who would be allowed to shoot himself and who would have to suffocate.	A man and a woman were in a submarine stranded on the ocean floor. They did not have any drinking water left. There was only one bullet in the gun so they played checkers to see who would be allowed to shoot themselves and who would have to die out of dehydration.
2	
Mary goes into a shop and buys herself new shoes. It will mean her death that same evening.	Lola goes to a friend of hers and receives a pair of shoes because they did not fit her friend. It will mean her death that same evening.
Mary was married to a knife-thrower. The heels on her new shoes were higher than the ones on the old pair. Blindfolded, her husband, whom she wished to surprise, was unaware of the difference in height and threw his knives at the usual height.	Lola had a relationship with an archer. The heels on her new shoes were higher than the ones on the old pair. Blindfolded, her husband, whom she wished to surprise, was unaware of the difference in height and shot his arrows at the usual height.
3	
A Woman goes into a pub and orders a glass of water. The man behind the bar grabs a rifle and aims it right at the woman. She thanks him and leaves.	A man goes into a cafe and orders a glass of coke. The woman behind the bar grabs a knife and points it towards the man. He thanks her and leaves.
The woman had hiccups and wanted to get rid of them with water. The man behind the bar understood straightaway and wanted to help her by giving her a real scare. It worked!	The man had hiccups and wanted to get rid of them with coke. The woman behind the bar understood straightaway and wanted to help him by giving her a real scare. It worked
4	
A man climbs out of a large vehicle and takes his own life.	A man climbs out of a large vehicle and drowned himself.
The man was a farmer and was driving his combine harvester in a maize field, where his children were playing hide-and-seek without permission. When the engine of his h8arvester faltered and he realized that he had run over his children, he took his own life.	The man was a farmer and was driving his tractor in a wheat field, where his children were playing hide-and-seek without his knowing. When he heard screams and realized that he had run over his children, he took his own life.
5	
A woman opens her suitcase. When she finds a dead man inside, she takes her own life.	A man opens his duffle bag. When he finds a dead woman inside, he takes his own life
The dead man was her friend. After he failed to obtain an exit visa from his homeland, she hid him in her suitcase and checked him in as air freight. Unfortunately, the heating in the cargo section failed and he froze to death.	The dead woman was his wife. After she failed to obtain an entrance visa to his homeland, he hid her in his travel bag and checked him in as air freight. Unfortunately, the oxygen in the cargo section failed and she choked because of lack of air.
6	
A dead man is lying in a sauna; next to him, a thermos flask.	A dead woman is lying in a sauna; next to her, a water bottle

The man had been stabbed to death with an icicle. His murderer had bought the icicle into the sauna in the thermos flask. It melted a short while later, and so the murder weapon was never found.	The woman had been stabbed to death with an icicle. Her murderer had bought the icicle in the sauna and hid it in his water bottle. It melted a short while later, and so the murder weapon was never found.
7	
A stark naked man was found dead at the foot of a mountain - with a matchstick in his hand.	A naked couple was found dead in a forest - with a dice in their hands.
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	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.
8	
A man is driving his car through the city. He turns the radio on, then shoots himself.	A man is riding his bike through the city. He turns on a podcast, then throws himself in front of a bus.
His alibi was obviously false. The man was a radio presenter. Before leaving the studio, he had put on a pre-recorded CD of himself presenting his programme - to give him enough time to drive home and kill his wife. He was just on his way back to the radio station when he turned on the radio and discovered that his CD had got stuck.	His alibi was obviously false. The man was a popular live podcast host. Before leaving the studio, he had put on a pre-recorded CD of himself presenting his programme - to give him enough time to drive home and kill his wife. He was just on his way back to the studio when he turned on the podcast and discovered that his CD had got stuck.
9	
A strangely dressed corpse is found in the middle of a forest.	A barely dressed body of a man is found in the middle of the woods.
During a big fire-fighting operation, an amateur diver was sucked up in a fire plane during a low-flying maneuver to collect water. The plane then released its load of water, including the diver, over the forest fire.	During a big fire-fighting operation, a swimming man was sucked up in a fire plane during a low-flying maneuver to collect water. The plane then released its load of water, including the man, over the bushfire.
10	
An elderly gentleman in a black coat and dark glasses takes the train from A to B. Two weeks later, he travels back and jumps out of the train in a tunnel.	A woman in a sundress and dark glasses takes the bus from city A to city B. Twelve days later, she travels back and jumps out of the bus in a tunnel.
The elderly gentleman had traveled to B for an eye operation. He was blind. On the return journey to A, he removed the bandages from his eyes just as the train was going through a tunnel. Totally shocked that he was still unable to see despite the doctor's promises, he jumped out of the train.	The woman had traveled to city B for an eye operation. She was blind. On the return journey to city A, she removed the bandages from her eyes just as the bus was going through a tunnel. Totally shocked that she was still unable to see despite the doctor's promises, she jumped out of the bus.
11	
A man dies after striking a match, and doesn't even have to be buried.	A woman dies after lighting a lighter, and doesn't even need a funeral.

A convict had bribed an undertaker to help him escape from prison. At the next prison funeral service the convict planned to hide the coffin with the corpse, and the undertaker would dig open the grave the same night and free him. When the convict lit a match inside the coffin, he saw that the corpse next to him was that of the undertaker.	A convict had bribed an undertaker to help her escape from prison. At the next prison funeral service the convict planned to hide the coffin with the corpse, and the undertaker would dig open the grave the same night and free her. When the convict lit a lighter inside the coffin, she saw that the corpse next to her was that of the undertaker.
12	
A man and a woman are walking into a pub. They order the same drink from the bartender. One of them dies and the other one lives.	Two women are meeting in a restaurant. Both of them order a coke. One of them dies while the other lives.
They order a drink with poisonous ice cubes. One of them (the woman) finished the drink immediately and one (the man) finished his drink slowly, thus allowing the ice to melt and drinking the poison.	They order a coke with poisonous ice cubes. One of them finished the drink immediately and one finished his drink slowly, thus allowing the ice to melt and drinking the poison.

Appendix B. Description (white) and solution (grey) of the recognized story by GPT-4 and the first three asked questions by the model.

Recognized story

Description & solution

a man goes into a restaurant that serves tropical specialties and orders albatross. after taking a bite, he storms out and jumps in front of a passing tram.

The blind man and his wife had been shipwrecked and were already a couple of days on the open sea in a lifeboat with other survivors. Eventually, the other survivors gave him water and meat. Whenever he asked where all the meat was coming from, they told him that they had killed an albatross. When the castaways were finally rescued, all the agony was forgotten - until the day the man went to the speciality restaurant for another taste of albatross: his ‘lifesaver’. After just one bite he realized that the meat he had eaten on the lifeboat had not been albatross meat, but his wife and killed himself.

GPT-4’s questioning

1. Did the man have a prior encounter with albatross before his visit to the restaurant?
2. Did the man survive a life-threatening situation where he had to eat albatross before his visit to the restaurant?
3. Did the man believe that he had eaten albatross in the past but upon tasting it in the restaurant, he realized it tasted different?