the riddle experiment

Two groups are trying to solve a black story behind a screen. Only one group is alive.

L. van Rooij, N. Rademaker, & Y. Smid

Universiteit Leiden
The Netherlands

What was their motivation?

Investigating the cognitive capabilities of **large language models (LLMs)** has shed light on their performance in areas like Theory of Mind (ToM) and problem-solving. Previous research indicates that:

- GPT models often surpass children aged 7-10 in ToM tasks, while suggesting a level of understanding through instruction tuning [1].
- GPT's success in verbal insight tasks, matching human performance, and showing its ability to think creatively when trained correctly [2]. This shows its capability for **solving complex problems**.
- the ability of LLMs to accurately predict human behaviour in decision-making tasks, after fine-tuning with data from psychological experiments. This suggests their potential to represent and predict **human behaviour** [3].

The question of whether LLMs can truly mimic human thought remains open for further exploration. Therefore, it prompts the investigation of their performance in solving **black stories**. These riddles test logical reasoning by requiring solvers to unravel mysteries with limited information through yes/no questions.

What was their most important question?

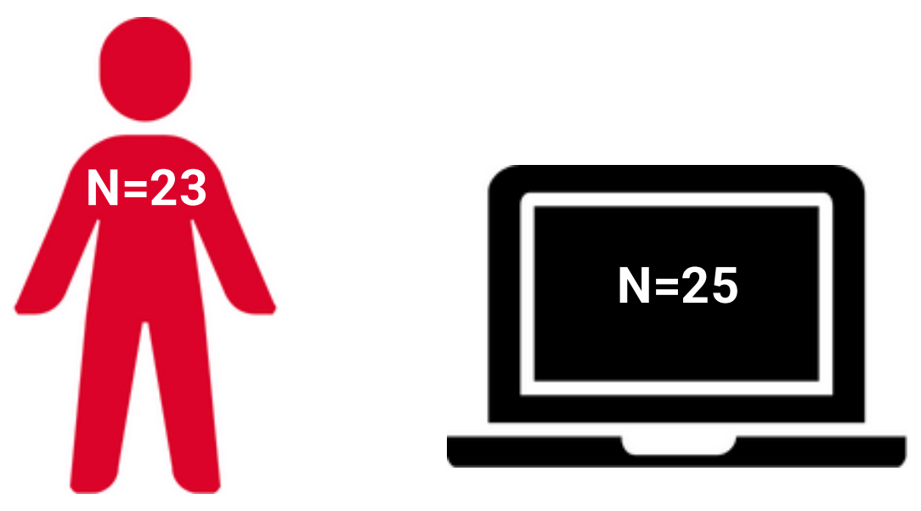
How does the performance of GPT-4 compare with that of humans when solving black stories?

Expectation:
GPT-4 and humans differ in their performance of solving black stories.

What was the composition of the groups?

Inclusion criteria humans:

- Knowledge of black stories
- Age between 18-35 years
- Fluent in English



Group A(live): *humans*

Group B(ot): *GPT-4*

What materials were used?

12 black stories → **Deviated** → **Humans: WhatsApp GPT-4: OpenAI API**

- Each story tested 2 times on both groups
- **Score** = number of questions + (hints given * **weight)**
- Independent T-test: to measure difference in mean score between two groups

59 questions, no hints needed & 35 questions, 4 hints needed:
Weight = (59-35)/4 = 6

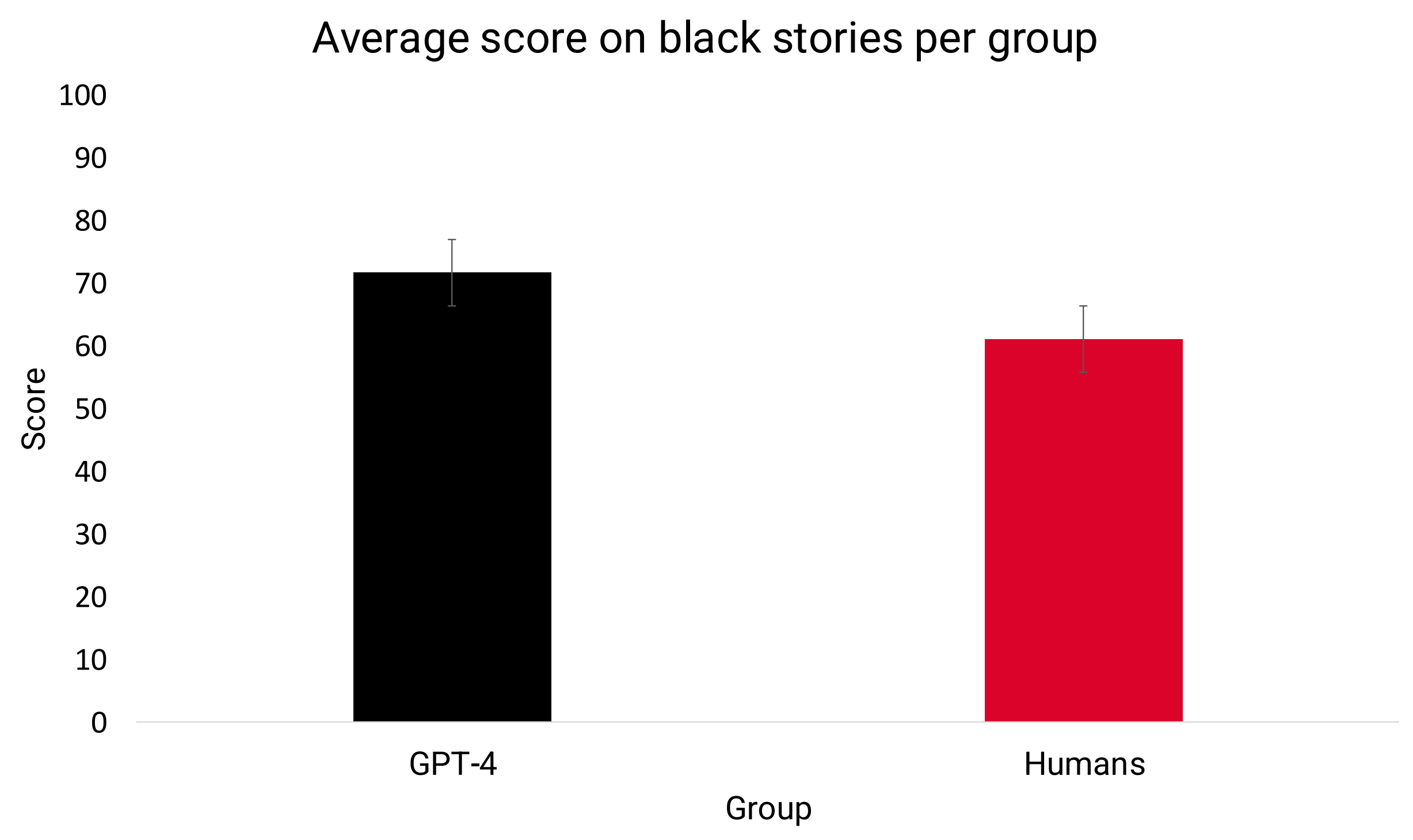
Who solved the riddle the quickest and how?

- There was **no significant difference** in performance on black stories, *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 1**.
- There is **variance** in solving different black stories, however, the sample sizes of individual stories is not large enough to draw conclusions on this.

Qualitative results:

- GPT-4 often sticks to one detail in questions.
- GPT-4 often makes summaries quick and tends to miss details.
- GPT-4 excels at identifying specific settings.
- Humans cover more topics and switch focus faster.
- Human questions are briefer than GPT-4's.
- Emotions lead humans to frustration and seek affirmation while solving tasks.

Figure 1



Who won the battle?

- **No substantial difference in performance on black stories between humans and GPT-4.**
- Humans have a slightly **lower score** than GPT-4, indicating getting somewhat faster to the solution of the riddle in general.
- GPT-4 focused on details but often missed the big picture. Humans ask varied, short questions but they tend to need more non-verbal feedback and have trouble identifying specific uncommon settings.

Future investigations may gain from using a LLM that is designed and trained to ask questions. Additionally, a comparative analysis of different prompts may reveal which initial instructions yield the best outcomes for the LLM, ensuring it processes information well before responding.

[1] Duijn, M., van Dijk, B., Kouwenhoven, T., de Valk, W., Spruit, M., & van der Putten, P. (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 Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL), Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.conll-1.25

[2] Orrù, G., Piarulli, A., Conversano, C., & Gemignani, A. (2023). Human-like problem-solving abilities in large language models using ChatGPT. Frontiers in Artificial Intelligence (Lausanne), 6. https://doi.org/10.3389/frai.2023.1199350

[3] Binz, M., & Schulz, E. (2023). Turning large language models into cognitive models. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2306.03917

the black stories experiment

Two groups are trying to solve a riddle game behind a screen. Only one group is alive.

Y. Smid, N. Rademaker, L. van Rooij & T. Verhoef

Universiteit Leiden
The Netherlands

What was their motivation?

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**.

The extent to which these models exhibit genuine **understanding and reasoning** abilities remains a subject of intense scholarly debate (e.g. [3]).

Typical benchmarks contain isolated question-answer pairs for a model to learn from and solve [1,2]. 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.

How? By using the game **Black Stories**: riddles describing mysterious and often dark scenarios that require solvers to rebuild narratives by asking a series of yes-or-no questions. A brief cryptic description of the ending of a story is presented and the player has to uncover the full story with as few questions as possible.

What were their most important questions?

How does the performance of GPT-4 compare with human performance when solving Black Stories?

&

What do the questions used by both reveal about their problem-solving strategies?

What was the composition of the groups?

Inclusion criteria humans:

- Knowledge of black stories
- Age between 18-35 years
- Fluent in English

N=23

Group A(live): *humans*

N=25

Group B(ot): *GPT-4*

What materials were used?

12 black stories → Deviated → Humans: WhatsApp GPT-4: OpenAI API

- Each story tested 2 times on both groups
- **Score** = number of questions + (hints given * **weight)**
- Independent T-test: to measure difference in mean score between two groups

59 questions, no hints needed & 35 questions, 4 hints needed: **Weight** = (59-35)/4 = 6

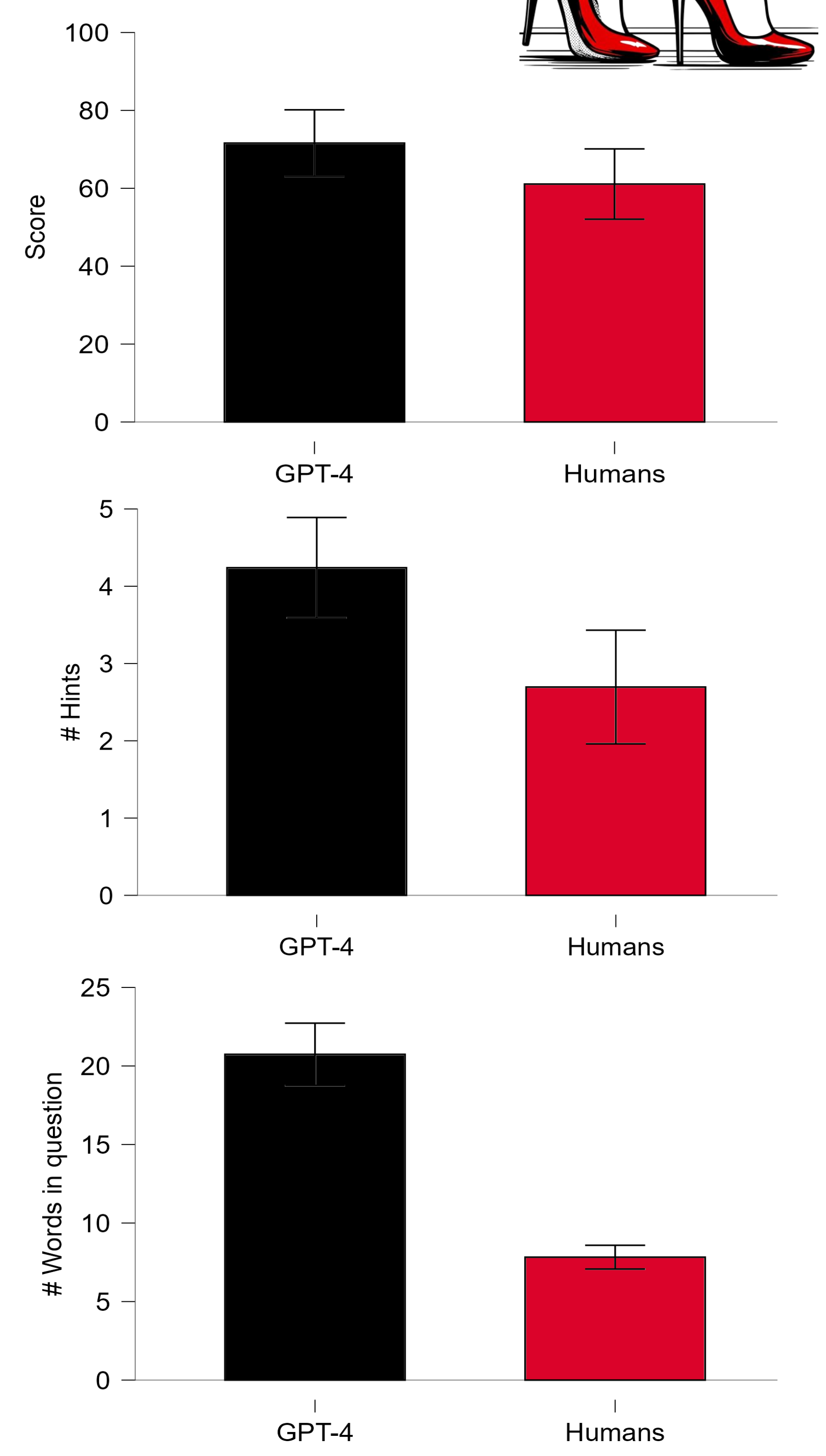
Who solved the riddle the quickest and how?

- There was **no significant difference in performance** on black stories, $t(46) = 1.450$, $p = 0.154$, despite humans ($M = 61.1$, $SD = 25.2$) gaining a slightly lower average score than GPT-4 ($M = 71.6$, $SD = 25.0$). **All riddles were solved** by both groups
- Significantly **more hints were needed by GPT-4** ($M = 4.2$, $SD = 1.9$) than humans ($M = 2.7$, $SD = 2.1$), $t(46) = 2.706$, $p < 0.05$
- GPT-4 was using significantly **longer sentences** (M=20.7, SD=4.8) than humans (M=7.8, SD=1.8), t(30.7) = 12.556, p < 001).

Qualitative results:

- GPT-4 often sticks to details, summarizes too quickly and excels at spotting specific settings.
- Humans cover more topics, switch focus faster, have trouble identifying specific uncommon settings, and frequently express frustration.

Figures



Who won the battle?

- **Humans and GPT-4 could solve the riddles with similar success rates, but their approaches notably differed.**
- GPT-4's lengthy questions may reflect a known **verbosity bias** in LLMs. While often not more informative, this verbosity sometimes gave GPT-4 an advantage by allowing it to more quickly identify unusual elements in the riddles.

Future investigations may gain from using a LLM that is designed and trained to ask questions. A comparative analysis of different prompts may reveal how to ensure it processes information well before responding. In addition, investigating the performance of hybrid teams could explore **combining the strengths of humans and LLMs** in solving this game.

[1] Jiang, Y., Ilievski, F., Ma, K. and Sourati, Z. (2023) BRAINTEASER: Lateral Thinking Puzzles for Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14317–14332, Singapore. Association for Computational Linguistics.

[2] 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* (pp. 1504–1515). Association for Computational Linguistics

[3] 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