Universiteit Leiden
The Netherlands

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

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?

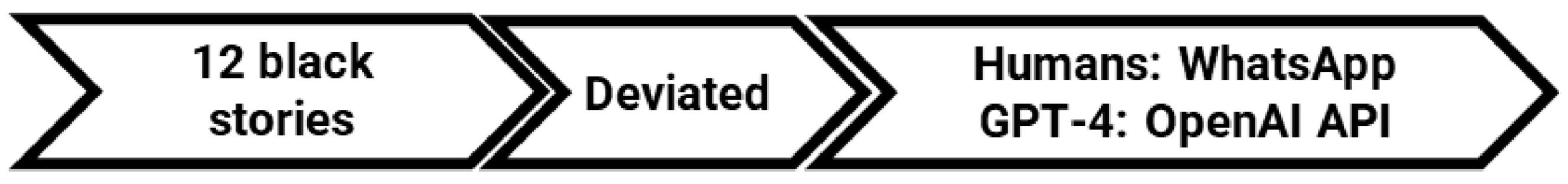
Inclusion criteria humans:

- Knowledge of black stories
- Age between 18-35 yrs.
- 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**

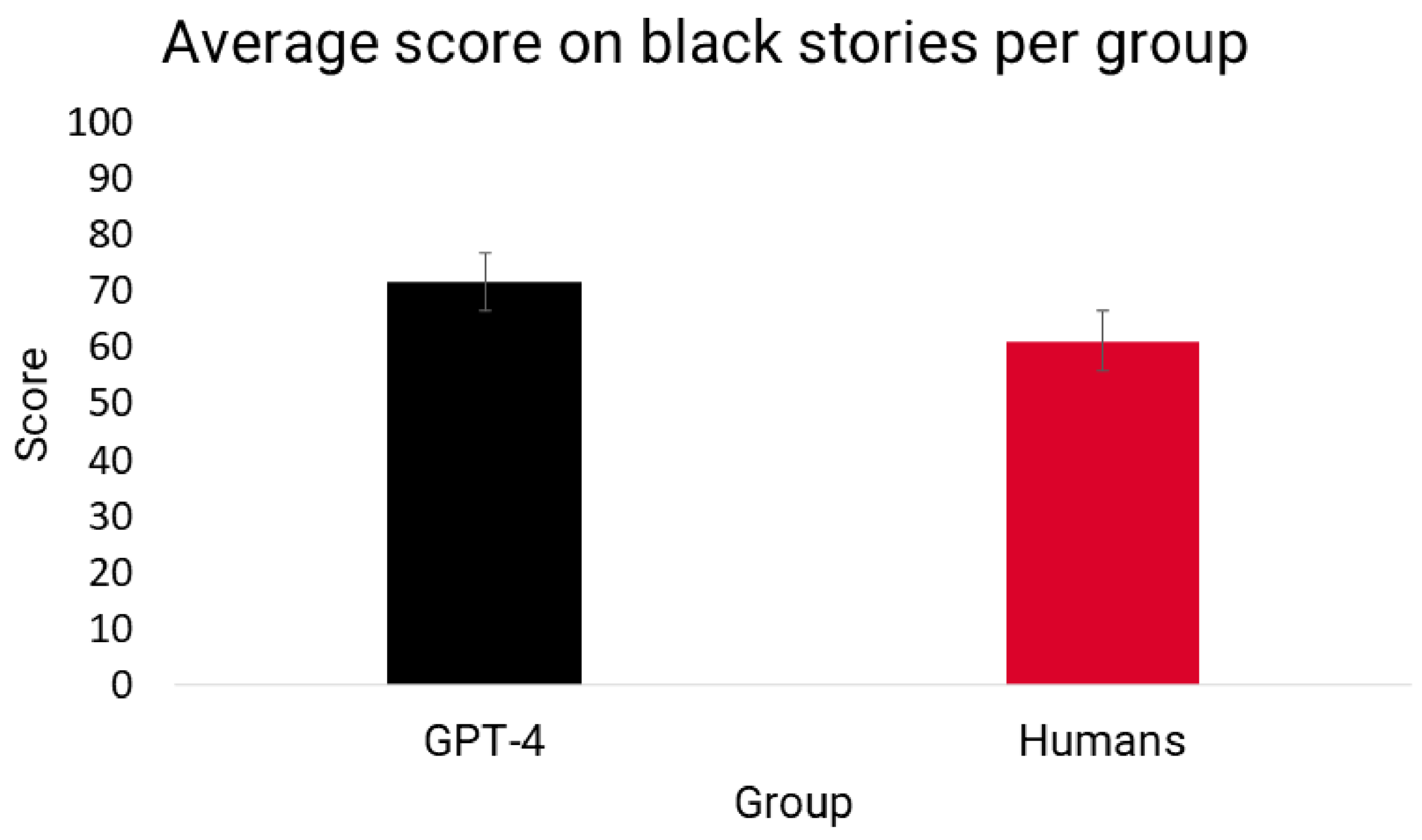
Who solved the riddle the quickest and how?

- 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