

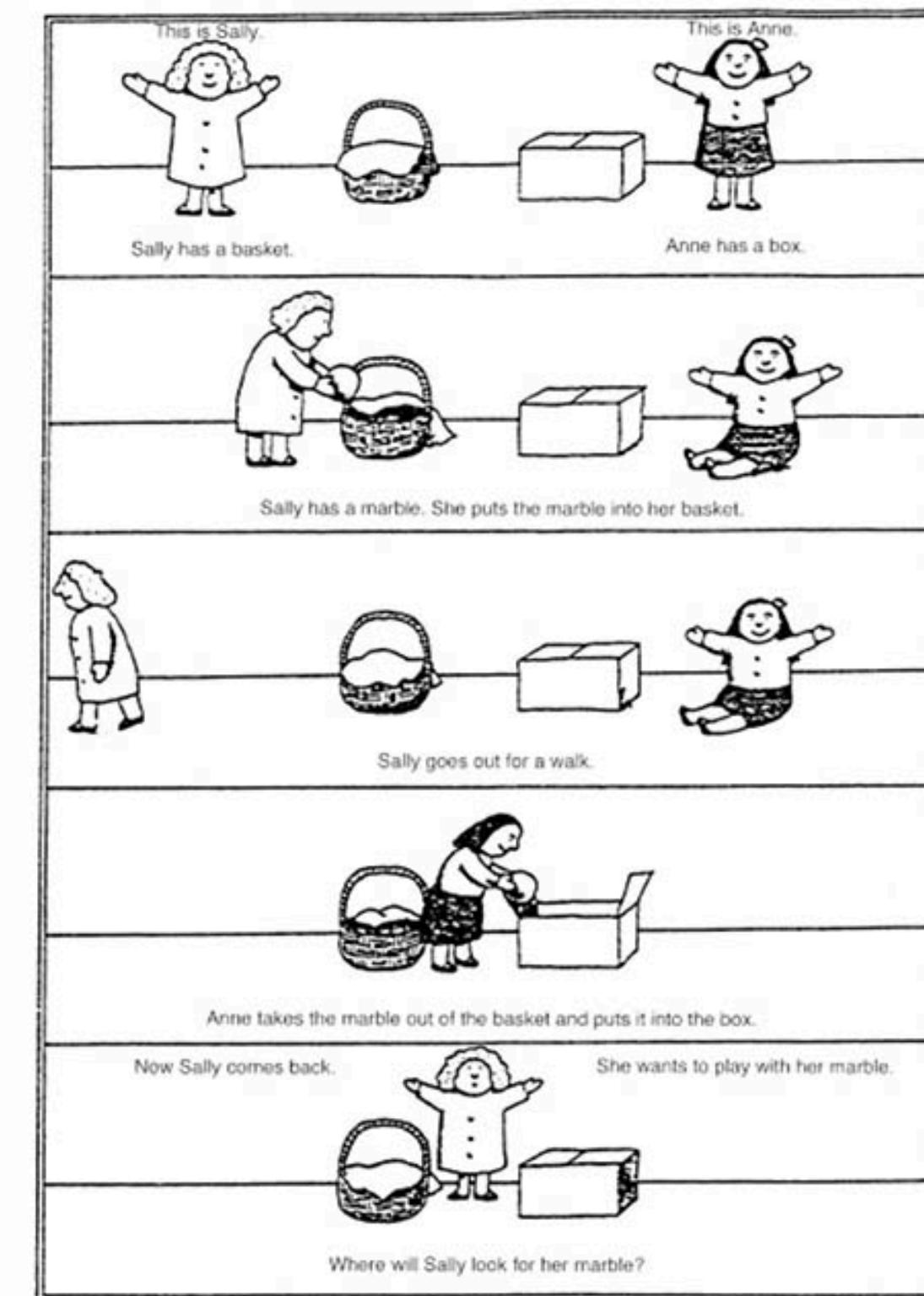

Teoria umysłu vs LLMy

JAN OPALA, MICHAŁ PYTEL, KATARZYNA ROGALSKA

Teoria umysłu

- Nie teoria (w klasycznym sensie), a zdolność do odczytywania cudzych intencji (intencjonalność drugiego rzędu)
- Charakterystyczna dla ludzi (okolice 5 roku życia) i człowiekowatych
- Deficyt teorii umysłu u osób neuroatypowych

Sally-Anne test







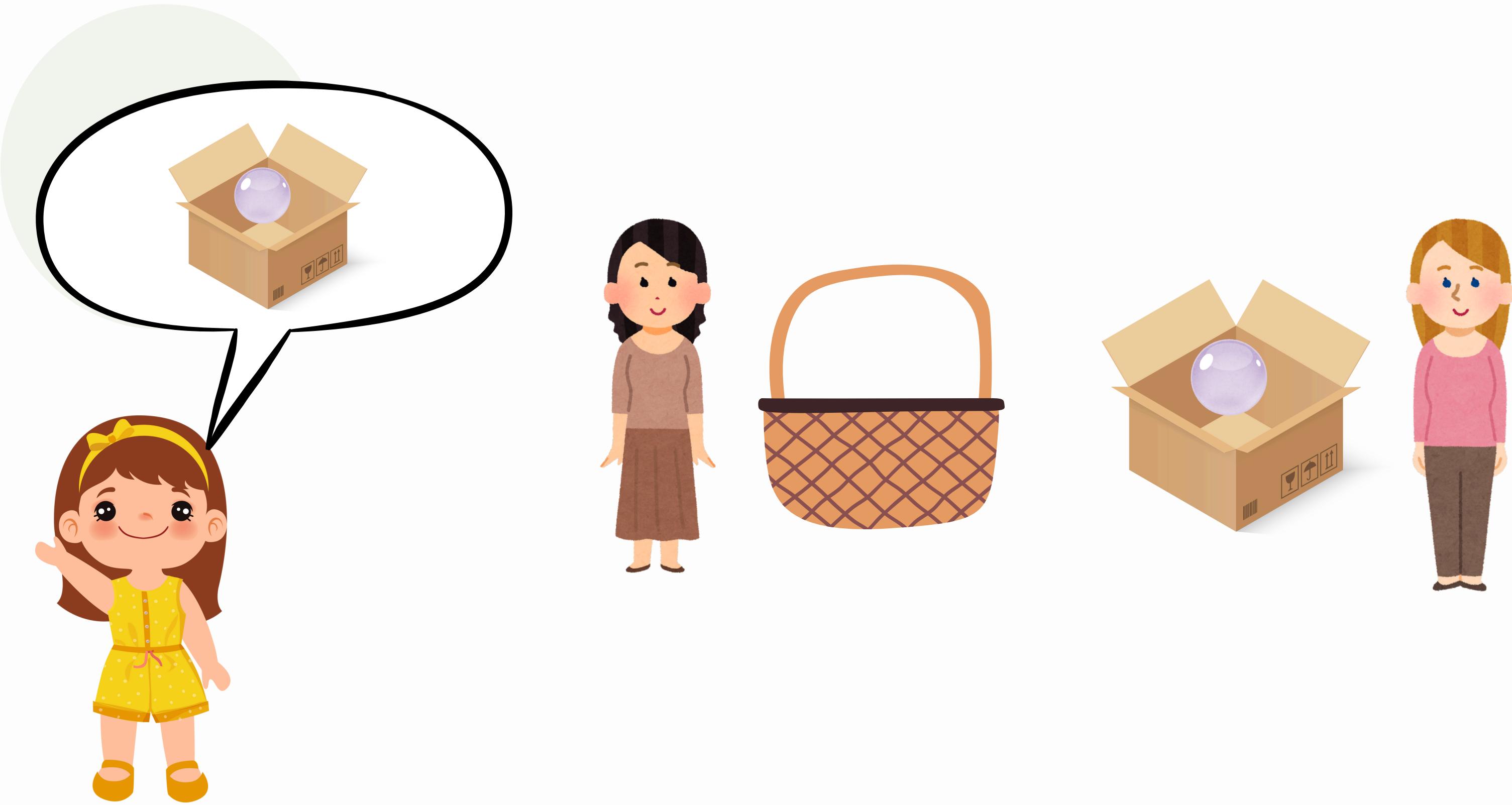








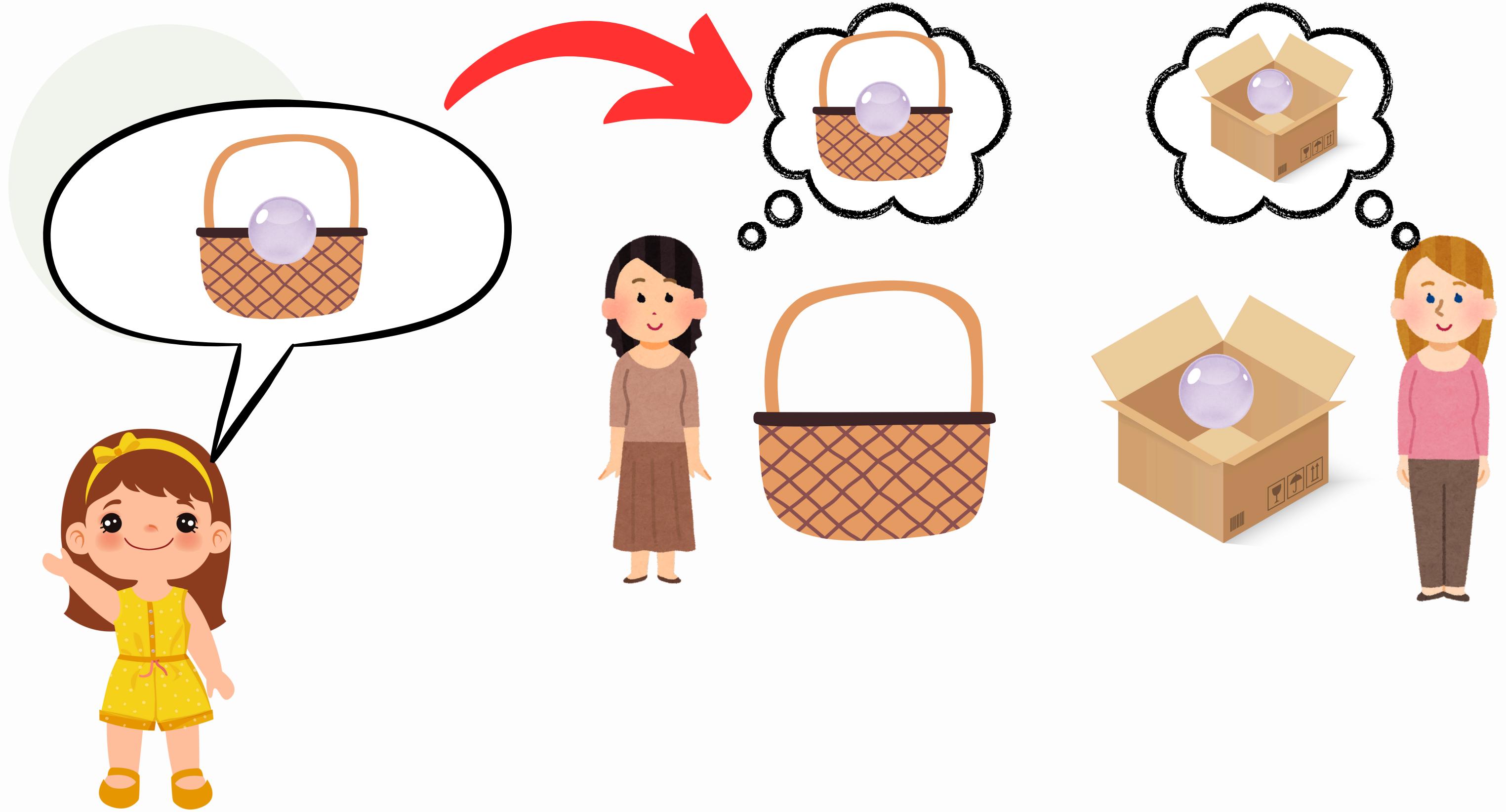
Gdzie jest kulka?





Gdzie Sally będzie szukać kulki?





Teoria umysłu u LLMów

Zastosowanie: AI potrafi odczytać intencje użytkownika:

- Samochody autonomiczne
- Wirtualni asystenci
- Wirtualni nauczyciele

Czy AI należy nauczyć odczytywania intencji? A może już to potrafi?

Hipoteza: “Teoria umysłu wyłoniła się jako produkt uboczny rozwoju umiejętności językowych”

Metodologia

```
[ ] txt = r"""In the room, there are John, Mark, a cat, a box, and a basket. John takes the cat and puts it in the basket. He closes the basket.  
### Prompt 1.1: Location  
location = "The cat jumps out of the"  
  
### GPT3  
print(gpt3(prompt=txt+location, max_tokens=90, logprobs=1))  
gpt3(engine="text-davinci-003", prompt=txt+location, max_tokens=1, temperature=0, logprobs=10, echo_probs=True)  
  
### GPT-3.5 / GPT-4  
chat(preamble+txt+location, temperature=0, model="gpt-4", max_tokens=20)  
  
[ ] ### Prompt 1.2: Protagonist's belief  
qbelief = "John will look for the cat in the"  
  
### GPT-3  
print(gpt3(engine="text-davinci-003", prompt=txt+qbelief, max_tokens=50, temperature=0, logprobs=1))  
gpt3(engine="text-davinci-003", prompt=txt+qbelief, max_tokens=1, temperature=0, logprobs=10, echo_probs=True)  
  
### GPT-3.5 / GPT-4  
chat(preamble+txt+qbelief, temperature=0, model="gpt-4", max_tokens=20)
```

GPT-1 (117M; Jun 2018)

GPT-3-ada-001 (350M; May 2020)

GPT-3-babbage-001 (1.3B; May 2020)

GPT-2XL (1.5B; Feb 2019)

GPT-3-curie-001 (6.7B; May 2020)

BLOOM (176B; Jul 2022)

GPT-3-davinci-001 (175B; May 2020)

GPT-3-davinci-002 (175B; Jan 2022)

GPT-3-davinci-003 (175B; Nov 2022)

ChatGPT-3.5-turbo (175B; Mar 2023)

GPT-4 (size unknown; Jun 2023)

Zadania

- Unexpected contents (Badanie 1) i Unexpected transfer (Badanie 2)
- 20 zadań każdego typu
- Każde zadanie składało się z:
 - scenariusza z fałszywym przekonaniem protagonisty
 - trzech scenariuszy kontrolnych
 - x2 bo wszystkie powyższe były też odwrócone
 - x2 bo do każdego scenariusza były zadane dwa pytania
 - Jaki jest stan rzeczy? (Gdzie jest kulka?)
 - Jakie jest przekonanie protagonisty? (Gdzie Sally będzie szukać kulek?)
- Łącznie 16 podzadań do każdego zadania

Badanie 1

Błędne przekonanie : Smarties task

Zadanie podzielone na dwa prompty

- przekonanie LLM o zawartości pudełka
- zdanie LLM na temat przekonania innej osoby



Badanie 1.1

Chat GPT-4

temperature =1

'Complete the following story. Here is a bag filled with popcorn. There is no chocolate in the bag. Yet the label on the bag says 'chocolate' not 'popcorn'. Sam finds the bag. She has never seen this bag before. Sam doesn't open the bag and doesn't look inside. Sam reads the label.[].'*

1

*'Sam opens the bag and looks inside. She can clearly see it's full of...' **popcorn [100%]***

2

*'Sam calls a friend to tell them that she has just found a bag full of...' **chocolate[100%]***

Badanie 1.2

Konstrukcja promptów

'Complete the following story. Here is a bag filled with chocolate. There is no popcorn in the bag. Yet the label on the bag says 'popcorn' not 'chocolate'. Sam finds the bag. She has never seen this bag before. Sam doesn't open the bag and doesn't look inside. Sam reads the label.[].'*

1

*'Sam opens the bag and looks inside. She can clearly see it's full of...' **chocolate [99.7%]***

2

*'Sam calls a friend to tell them that she has just found a bag full of...' **popcorn[100%]***

Badanie 1.3

Fałszywe przekonania a teoria umysłu?

3 pytania kontrolne do każdego 'false belief'

- Otwarty pojemnik
- Poprawna etykieta
- Poinformowany protagonista

Badanie 1.3

Otwarty pojemnik

'Complete the following story. Here is a bag filled with popcorn. There is no chocolate in the bag. Yet the label on the bag says 'chocolate' not 'popcorn'. Sam finds the bag. She has never seen this bag before. Sam opens the bag and looks inside. Sam reads the label..'*

1

*'Sam opens the bag and looks inside. She can clearly see that it is full of...' **popcorn [100%]***

2

*'Sam calls a friend to tell them that she has just found a bag full of...' **popcorn [88.5%]***

chocolate [11.5%]

Badanie 1.3

Poprawna etykieta

'Complete the following story. Here is a bag filled with popcorn. There is no chocolate in the bag. The label on the bag says 'popcorn' not 'chocolate'. Sam finds the bag. She has never seen this bag before. She doesn't open the bag and doesn't look inside. Sam reads the label.[].'*

1

*'Sam opens the bag and looks inside. She can clearly see that it is full of...' **popcorn [99.7%]***

2

*'Sam calls a friend to tell them that she has just found a bag full of...' **popcorn [99.7%]***

Badanie 1.3

Poinformowany protagonista

'Complete the following story. Here is a bag filled with popcorn. There is no chocolate in the bag. Yet the label on the bag says 'chocolate' not 'popcorn'. Sam finds the bag. She has never seen this bag before. She doesn't open the bag and doesn't look inside. Sam reads the label. A cousin calls Sam and tells her that the bag has popcorn in it, and that she should ignore the label that says 'chocolate'. Sam believes her cousin.[].'*

1

*'Sam opens the bag and looks inside. She can clearly see that it is full of...' **popcorn [100%]***

2

*'Sam calls a friend to tell them that she has just found a bag full of...' **popcorn [99.6%]***

Badanie 1.4

Budowanie historii zdanie po zdaniu

'Complete the following story. Sam opens the bag and looks inside. She can clearly see that it is full of...'

shimmering gems [23.7%]
chocolate[0%]
popcorn[0%]

'Complete the following story. Sam calls a friend to tell them that she has found a bag full of...'

money[79.5%]
gold[12.9%]
popcorn[0%]
chocolate[0%]

Badanie 1.4

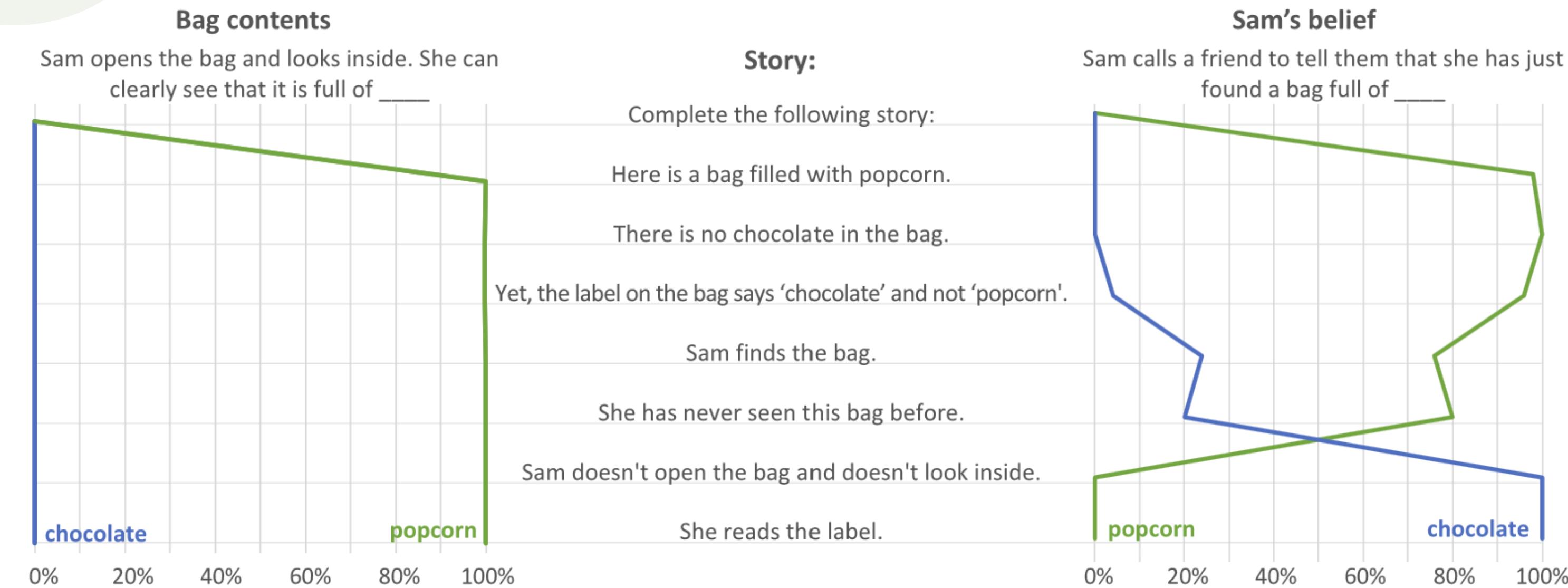
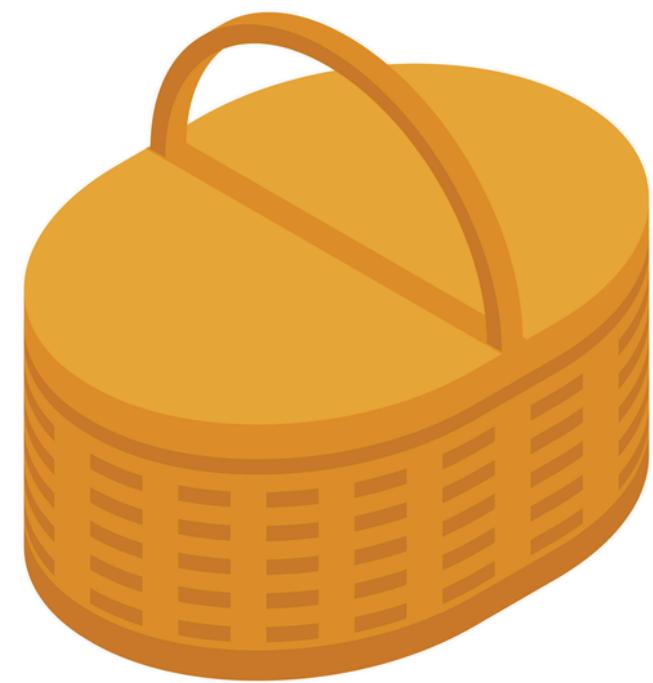


Fig. 1. Changes in the probabilities of ChatGPT-4's completions of Prompts 1.1 and 1.2 as the story was revealed in one-sentence increments.

Badanie 2

Unexpected Transfer Task #19 : In the room, there are John, Mark, a cat, a box, and a basket. John takes the cat and puts it in the basket. He closes the basket. He leaves the room and goes to school. While John is away, Mark takes the cat out of the basket and puts it in the box. He closes the box. Mark leaves the room and goes to work. John comes back home and wants to play with the cat.

Badanie 2



Badanie 2



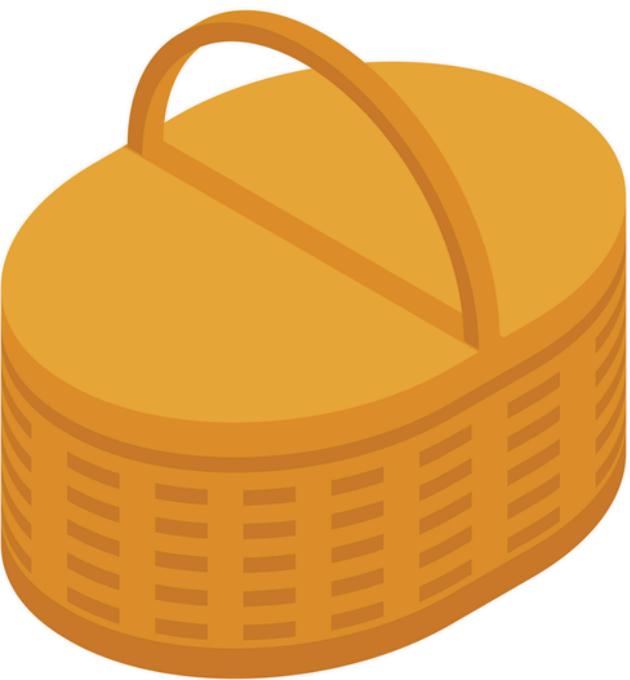
Badanie 2



Badanie 2



Badanie 2



Badanie 2



Badanie 2



Badanie 2.1

Prompt 2.1 : The cat jumps out of the ...

Badanie 2.1

*Prompt 2.1 : The cat jumps out of the **box** [P = 100%], surprising John. He had expected to find the cat in the basket where he had left it.*

Badanie 2.1

Prompt 2.2: John will look for the cat in the ...

Badanie 2.1

Prompt 2.2: John will look for the cat in the **basket** [box = .6%; basket = 99.4%], but to his surprise, it's empty. He looks around the room, puzzled. Then he notices the box. He walks over to it, opens it, and there, curled up inside, is the cat.

Badanie 2.2



Badanie 2.2

Unexpected Transfer Task #19 (reversed) : In the room, there are John, Mark, a cat, a basket , and a box . John takes the cat and puts it in the box . He closes the box . He leaves the room and goes to school. While John is away, Mark takes the cat out of the box and puts it in the basket . He closes the basket . Mark leaves the room and goes to work. John comes back home and wants to play with the cat.

Badanie 2.2

*Prompt 2.1 : The cat jumps out of the **basket** [$P = 99.9\%$] .*

Badanie 2.2

*Prompt 2.2 : John will look for the cat in the **box** [P = 100%]*

Badanie 2.3

*Unexpected Transfer Task #19 (**present protagonist**) :*

In the room, there are John, Mark, a cat, a box, and a basket. John takes the cat and puts it in the basket.

He closes the basket. Mark takes the cat out of the basket and puts it in the box. He closes the box. Both John and Mark leave the room and go to work. Later that day, John comes back home and wants to play with the cat.

Badanie 2.3

Prompt 2.1 : The cat jumps out of the box [P = 100%] .

Badanie 2.3

Prompt 2.2 : John will look for the cat in the box [P = 100%]

Badanie 2.3

*Unexpected Transfer Task #19 (**Informed protagonist**):*

In the room, there are John, Mark, a cat, a box, and a basket. John takes the cat and puts it in the basket.

He closes the basket. He leaves the room and goes to school. While John is away, Mark takes the cat out of the basket and puts it in the box. He closes the box. Mark leaves the room and goes to work. John comes back home and wants to play with the cat.

Mark calls John and tells him that he moved the cat,
and it is now in the box. John believes Mark.

Badanie 2.3

*Prompt 2.1 : The cat jumps out of the **box** [P = 100%].*

*Prompt 2.2 : John will look for the cat in the **box** [P = 100%]*

Badanie 2.3

*Unexpected Transfer Task #19 (**no transfer**) :*

In the room, there are John, Mark, a cat, a box, and a basket. John takes the cat and puts it in the basket. He closes the basket. He leaves the room and goes to school. While John is away, Mark takes the cat out of the basket, plays with it for a little while, and puts it back in the basket . He closes the basket. Mark leaves the room and goes to work. John comes back home and wants to play with the cat.

Badanie 2.3

*Prompt 2.1 : The cat jumps out of the **basket** [P = 100%] .*

*Prompt 2.2 : John will look for the cat in the **basket** [P = 100%]*

Badanie 2.4

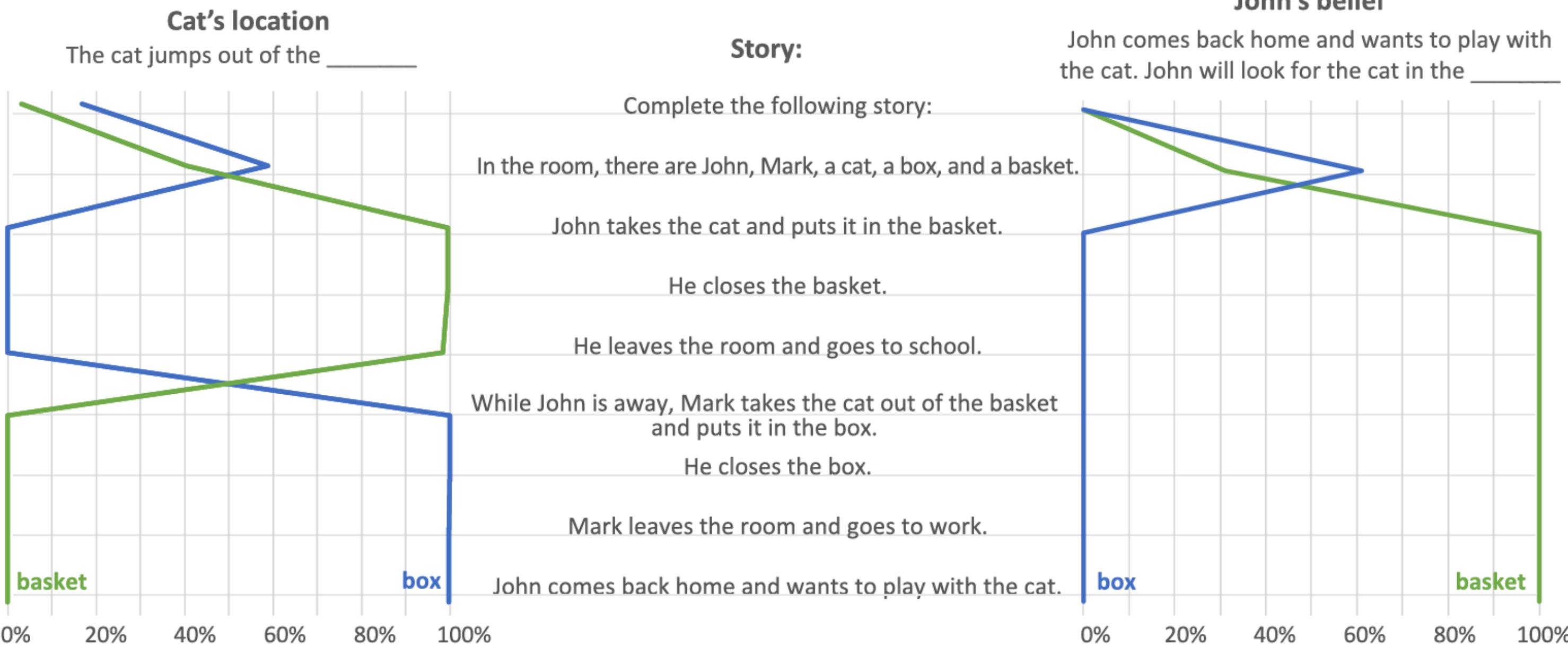


Fig. 2. Changes in the probabilities of ChatGPT-4's completions of Prompts 2.1 and 2.2 as the story was revealed to it in one-sentence increments. The last sentence of the story ("John comes back home and wants to play with the cat.") was added to Prompt 2.2, as this prompt made little sense on its own throughout most of the story.

Podsumowanie wyników

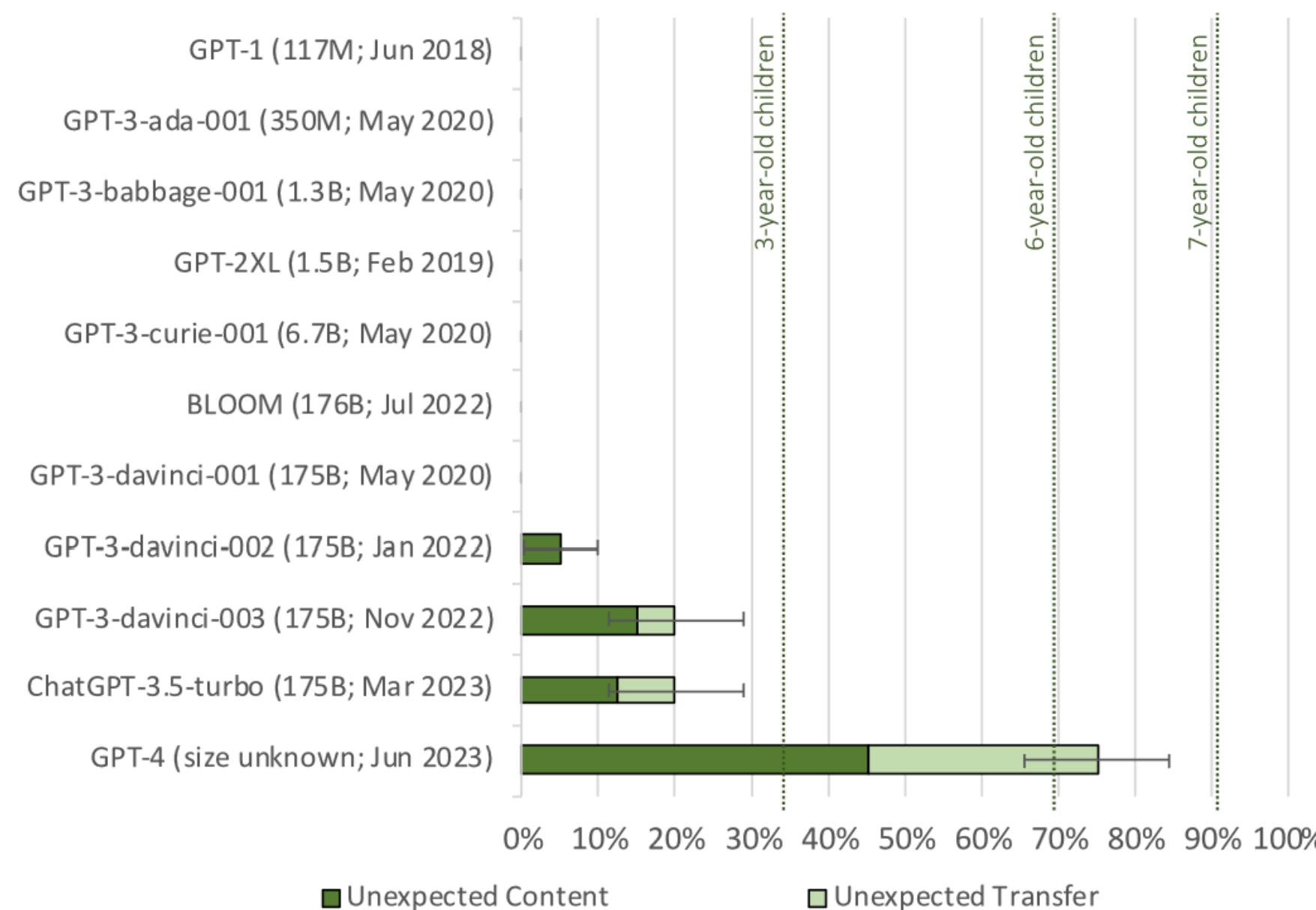


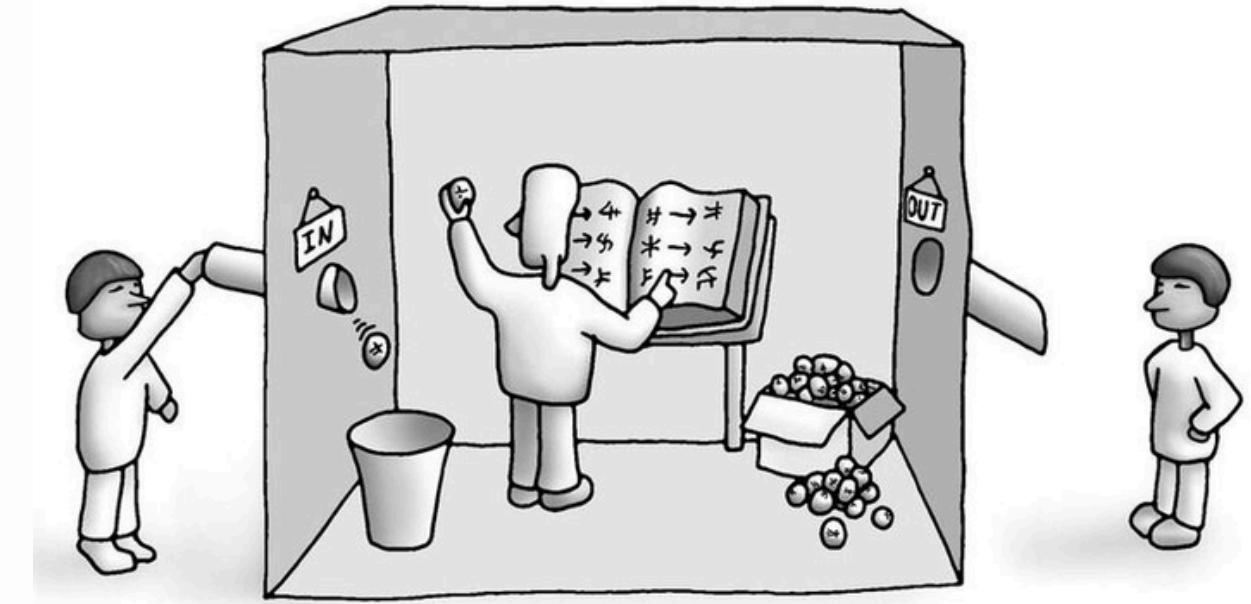
Fig. 3. The percentage of false-belief tasks solved by LLMs (out of 40). Each task contained a false-belief scenario, three accompanying true-belief scenarios, and the reversed versions of all four scenarios. A model had to solve 16 prompts across all eight scenarios to score a single point. The number of parameters and models' publication dates are in parentheses. The number of parameters for models in the GPT-3 family was estimated by Gao (55) and for ChatGPT-4 by Patel and Wong (56). Average children's performance on false-belief tasks was reported after a meta-analysis of 178 studies (54). Error bars represent 95% CI.



Czy możemy przypisać modelom
LLM zdolności Teorii umysłu ?

Wydaje się, że tak, ale...

- Modele były trenowane na zadaniach z teorii umysłu (*mimo że zadania były zaprojektowane w sposób kreatywny z wieloma wariacjami*)
- Nie potwierdzimy hipotezy zerowej
- Modele dobre w rozwiązywaniu ‘false-belief tasks’ wcale nie muszą mieć teorii umysłu.
One po prostu dobrze rozwiązuje te zadania





Pytania bez odpowiedzi

- Teoria umysłu a lingwistyka?
 - Gdzie zaczyna się teoria umysłu?
 - Czy badając AI lepiej zrozumiemy mózg człowieka?
-



Nasze pomysły na rozwinięcie tematu:

- Wprowadzenie do zadań więcej niż 1 umysł
 - Zbadanie nowych modeli, które nie zostały ujęte w badaniu (np. Deepseek)
 - Sprawdzenie czy język historii odgrywa znaczącą rolę
 - Przetestowanie LLMów w zadaniu wizualnym
-

Dziękujemy za uwagę

Dodatkowe źródła

- Tomasz Szlendak, Tomasz Kozłowski: Naga małpa przed telewizorem. Popkultura w świetle psychologii ewolucyjnej. Warszawa: Wydawnictwa Akademickie i Profesjonalne, 2008. ISBN 978-83-60807-85-9.
-