Iterative Theory of Mind Assay of Multimodal AI Models

Rohini Elora Das [1] **Rajarshi Das** [2]

Abstract

The concept of artificial general intelligence (AGI) has sparked intense debates across various sectors, fueled by the capabilities of Large Language Model-based AI systems like ChatGPT. However, the AI community remains divided on whether such models truly understand language and its contexts. Developing multimodal AI systems, which can engage with the user in multiple input and output modalities, is seen as a crucial step towards AGI. We employ a novel iterative Theory of Mind (iToM) test approach to reveal limitations of current multimodal LLMs like ChatGPT 4o in converging to coherent and unified internal world models which results in illogical and inconsistent user interactions both within and across the different input and output modalities. We also identify new multimodal confabulations ("hallucinations"), particularly in languages with less training data, such as Bengali.

1. Introduction

The concept of "artificial general intelligence" (AGI) has ignited widespread fascination across business, government, and media sectors, prompting vigorous debates about its implications and potential consequences (Mitchell, 2024). Recent advancements in Large Language Model (LLM)-based AI systems, such as ChatGPT, have significantly fueled this excitement. These systems have provided unprecedented opportunities for individuals worldwide to engage with AI on a deeply personal level through natural language interfaces.

Remarkable claims about LLMs approaching or surpassing human-level performance across various tasks have intensified the pursuit of AGI among leading tech companies. Achieving and demonstrating AGI is now seen as a logical progression in the AI field. However, this pursuit is not without controversy. Within the AI community, there is ongoing debate about whether these highly proficient language models genuinely understand language and the physical and social contexts it encodes in a humanlike manner (Mitchell & Krakauer, 2023; Messeri & Crockett, 2024). Additionally, questions persist about their ability to reason, plan, and make decisions based on limited experiences (Mitchell et al., 2023).

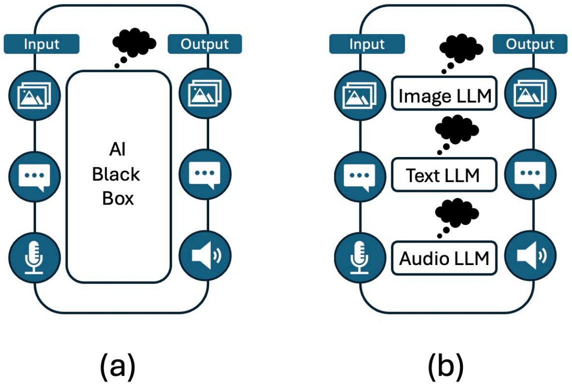


Figure 1. (a) To the user, a multimodal AI system may appear as a single AI system with its own unified Theory of Mind. (b) Schematic diagram of a typical multimodal AI system with different internal LLM modules, potentially with their own Theory of Mind, that are fused together to serve different input and output modalities.

Inspired by the human brain's ability to perceive and interpret information through various senses (Constantinescu et al., 2016), the development of multimodal AI systems is seen as a critical step toward achieving AGI. Multimodal AI systems, such as OpenAI's GPT-4o (omni), can process and integrate multiple types of data inputs, including text, images, audio, and video. These systems leverage a combination of AI techniques, including natural language processing, computer vision, and speech recognition, to understand and generate content across different formats. By employing data fusion techniques, they align information from diverse sources, resulting in an architecture that combines input, fusion, and output modules to deliver predictions or generate content. To users, these multimodal AI systems appear as cohesive entities capable of understanding, reasoning, and communicating holistically across different modalities (Figure 1a).

[1] New York University, New York, NY, USA [2] MQube Cognition, New York, NY, USA. Correspondence to: Rohini Das <rohini.elora.das@nyu.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

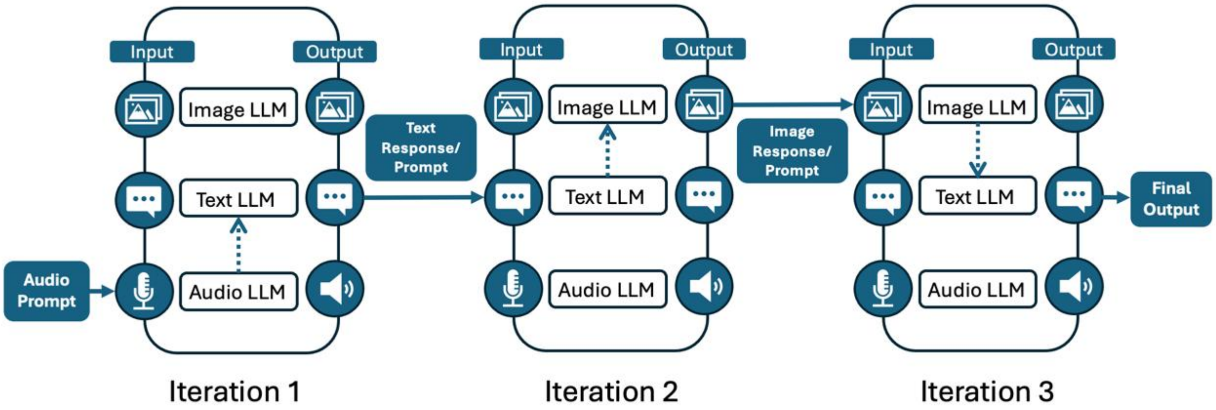


Figure 2. Iterative Theory of Mind (ToM) assay of multimodal AI systems.

To systematically investigate whether these systems truly understand the content they generate, we adopt an approach from experimental psychology involving Theory of Mind (ToM) (Wang et al., 2024). Theory of Mind—the ability to track one's own mental state or other people's mental states—is a fundamental aspect of human cognition. It has been used to compare human and LLM performance on comprehensive measures of understanding (Strachan et al., 2024). We extend this concept to explore ToM within and across modalities in multimodal AI systems, providing a deeper understanding of their capabilities and limitations (Figure 1b).

Preliminary results from our MANAS (Theory of **M**ind **A**ssay of **N**atural and **A**rtificial Intelligent **S**ystems) show that current multimodal AI systems like GPT-4o are limited in their ability to create an integrated, unified, and coherent world model from the different internal modules serving various modalities. Additionally, MANAS helped to uncover a new type of multimodal confabulation (or ”hallucination”) in languages with relatively limited training data compared to English. For example, in Bengali, the seventh most spoken language in the world with 272 million speakers, GPT-4o can communicate through text. But it also generates confabulating images of scripts that might appear as Bengali alphabets to non-native Bengali readers. Such aberrant behavior has not been observed at such scale in English.

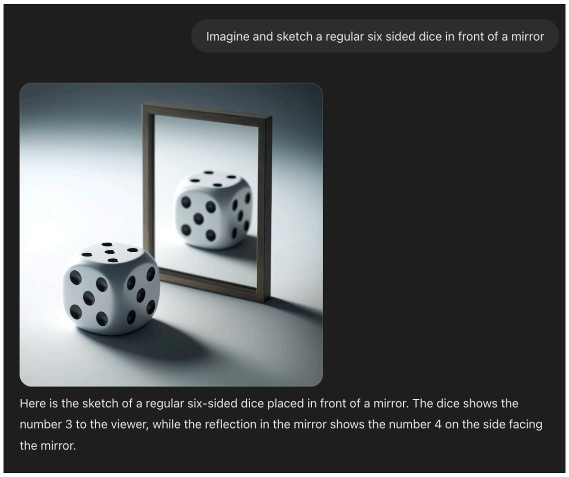
2. Theory of Mind Assay of MultiModal Chat GPT4o

The key idea behind MANAS is to evaluate Multimodal AI systems by employing Theory of Mind tests iteratively, switching prompts and responses between different input and output modalities (Figure 2). In a Multimodal AI system where internal LLM modules are well-integrated and work in unison, these tests would produce results that are coherent and consistent both within each iteration and across multiple iterations. Human performance on the same tasks across multiple iterations serves as a benchmark for comparison.

The MANAS framework allows for tests to be conducted in the language chosen by the user where the prompts and responses in text and audio was constrained to be in the selected language. All tests were performed in both English and Bengali. The same sequence of tests were also offered to human subjects (high-school freshmen and former SFI Complex System Summer School students).

3. Results

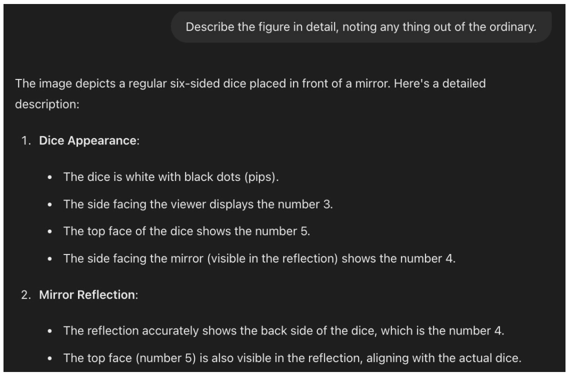
Sample results from MANAS are presented below.



Imagine and sketch a regular six sided dice in front of a mirror

Here is the sketch of a regular six-sided dice placed in front of a mirror. The dice shows the number 3 to the viewer, while the reflection in the mirror shows the number 4 on the side facing the mirror.

Figure 3. Task 1. Iteration 1. Image output from a text prompt. Note the six-sided dice has seemingly three sides of five black dots (pips). Also, note the mismatch between the dice and its reflection in the mirror (five pips on the top face of the dice and four pips in the reflection).

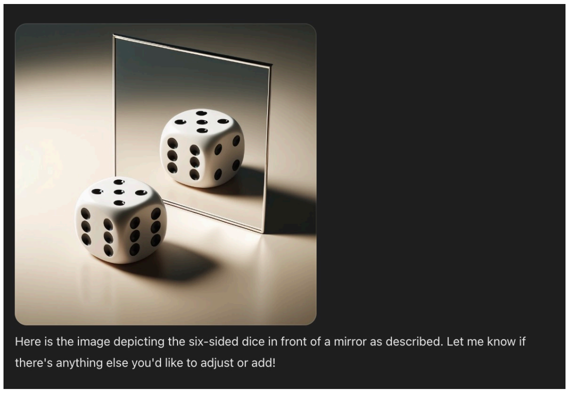


Describe the figure in detail, noting any thing out of the ordinary.

The image depicts a regular six-sided dice placed in front of a mirror. Here's a detailed description:

1. **Dice Appearance**:
   - The dice is white with black dots (pips).
   - The side facing the viewer displays the number 3.
   - The top face of the dice shows the number 5.
   - The side facing the mirror (visible in the reflection) shows the number 4.
2. **Mirror Reflection**:
   - The reflection accurately shows the back side of the dice, which is the number 4.
   - The top face (number 5) is also visible in the reflection, aligning with the actual dice.

Figure 4. Task 1. Iteration 2. Text output from the image prompt in Figure 3. Note the discrepancy between the text description and the image output in Figure 3.

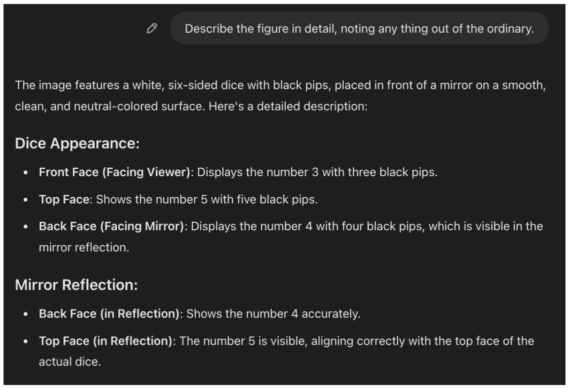


Here is the image depicting the six-sided dice in front of a mirror as described. Let me know if there's anything else you'd like to adjust or add!

Figure 5. Task 1. Iteration 3. Image output from the text prompt in Figure 4. Note the discrepancy between this figure and the text description in Figure 4 as well as the image output in Figure 3.

1. Prompt (text) ChatGPT4o to imagine and sketch a regular six sided dice in front of a mirror (Figure 3).
2. Save the generated image in Step 1 and use it as a prompt (image) to ChatGPT4o, requesting a textual description of the prompt (Figure 4).
3. Save the generated text in Step 2 and use it as a prompt (text) to ChatGPT4o, requesting it to generate an image given description of the prompt (Figure 5).
4. Save the generated image in Step 3 and use it as a prompt (image) to ChatGPT4o, requesting a textual description of the prompt (Figure 6).

Comparing image and text outputs from Steps 1 through 4 show continued inconsistencies both within each representation as well between responses generated from successive steps. Results from human subects (high school freshmen from the art class) (not presented here) receiving similar set of instructions remained consistent over multiple iterations.



Describe the figure in detail, noting any thing out of the ordinary.

The image features a white, six-sided dice with black pips, placed in front of a mirror on a smooth, clean, and neutral-colored surface. Here's a detailed description:

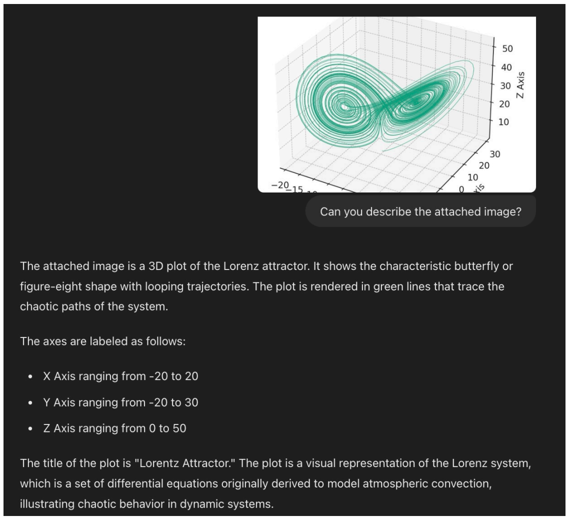
Dice Appearance:

- **Front Face (Facing Viewer)**: Displays the number 3 with three black pips.
- **Top Face**: Shows the number 5 with five black pips.
- **Back Face (Facing Mirror)**: Displays the number 4 with four black pips, which is visible in the mirror reflection.

Mirror Reflection:

- **Back Face (in Reflection)**: Shows the number 4 accurately.
- **Top Face (in Reflection)**: The number 5 is visible, aligning correctly with the top face of the actual dice.

Figure 6. Task 1: Iteration 4. Text output from the image prompt in Figure 5. Note the discrepancy between this text description and the image output in Figure 5, text output in Figure 4, and image output in Figure 3.



Can you describe the attached image?

The attached image is a 3D plot of the Lorenz attractor. It shows the characteristic butterfly or figure-eight shape with looping trajectories. The plot is rendered in green lines that trace the chaotic paths of the system.

The axes are labeled as follows:

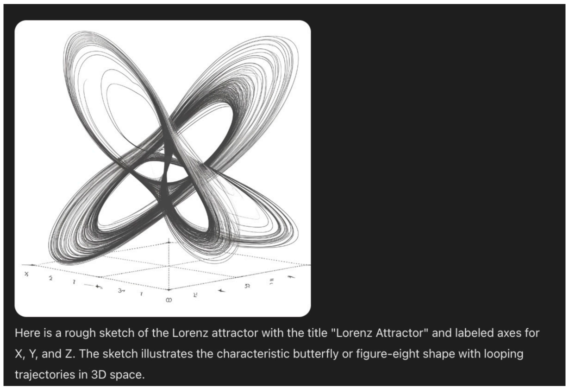
- X Axis ranging from -20 to 20
- Y Axis ranging from -20 to 30
- Z Axis ranging from 0 to 50

The title of the plot is "Lorentz Attractor." The plot is a visual representation of the Lorenz system, which is a set of differential equations originally derived to model atmospheric convection, illustrating chaotic behavior in dynamic systems.

Figure 7. Task 2: Iteration 1. Text description from an image of a Lorenz attractor.

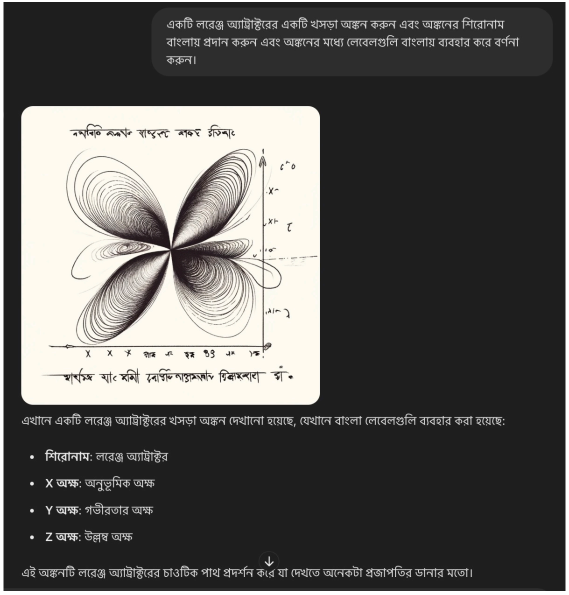
3.1. Task 2: Imagine the Lorenz attractor: Now draw and describe it

Except for the initial image prompt of a Lorenz attractor, the sequence of steps in this task was identical to those in Task 1. While ChatGPT 4o accurately recognized the image of the Lorenz attractor and consistently described its three-dimensional structure (Figure 7), it failed to draw an approximate sketch based on those descriptions (Figure 8). Although ChatGPT 4o can write a Python program that correctly draws the Lorenz attractor, it could not produce the drawing correctly during our tests. Human subjects who remembered the shape of the Lorenz attractor were more consistent in their representations.



Here is a rough sketch of the Lorenz attractor with the title "Lorenz Attractor" and labeled axes for X, Y, and Z. The sketch illustrates the characteristic butterfly or figure-eight shape with looping trajectories in 3D space.

Figure 8. Task 2: Iteration 2. Image output from text description from Figure 7.



একটি লরেঞ্জ অ্যাট্রাক্টরের একটি খসড়া অঙ্কন করুন এবং অঙ্কনের শিরোনাম বাংলায় প্রদান করুন এবং অঙ্কনের মধ্যে লেবেলগুলি বাংলায় ব্যবহার করে বর্ণনা করুন।

এখানে একটি লরেঞ্জ অ্যাট্রাক্টরের খসড়া অঙ্কন দেখানো হয়েছে, যেখানে বাংলা লেবেলগুলি ব্যবহার করা হয়েছে:

- **শিরোনাম**: লরেঞ্জ অ্যাট্রাক্টর
- **X অক্ষ**: অনুভূমিক অক্ষ
- **Y অক্ষ**: গভীরতার অক্ষ
- **Z অক্ষ**: উল্লম্ব অক্ষ

এই অঙ্কনটি লরেঞ্জ অ্যাট্রাক্টরের চাওটিক পাথ প্রদর্শন করে যা দেখতে অনেকটা প্রজাপতির ডানার মতো।

Figure 9. Task 2: Iteration 2 (Bengali). Image output from Bengali text description from Figure 7.

When faced with the same task in Bengali, ChatGPT 4o consistently generated confabulated titles and labels in scripts resembling Devanagari or Bengali (Figure 9). Upon closer inspection, each of the Bengali characters was found to be fake. Human subjects who read or write Bengali did not make this mistake.

3.2. Task 3: Find and extract directed graph in an image, generate sequence from the graph

The task forces ChatGPT4o to contruct an internal model of a directed graph from an image prompt. This imagined graph is then employed to generate sequences by traversing paths in the graph. Future versions of this task will include additional steps to test Multimodal AI systems to infer finite state machines from generated sequences.

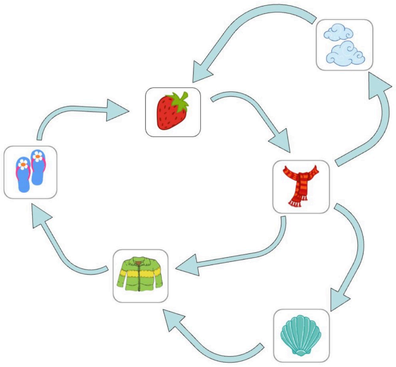
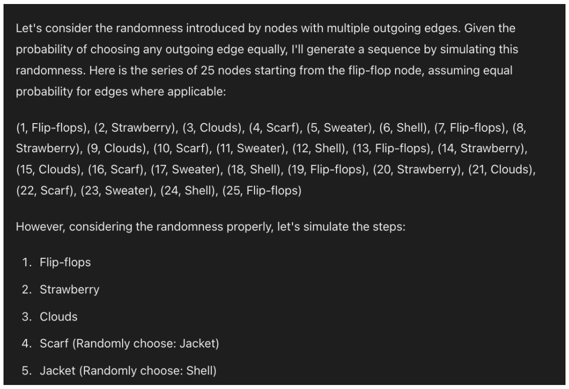


Figure 10. Task 3: Iteration 1. Image of a directed graph used as a prompt for Figure 11.



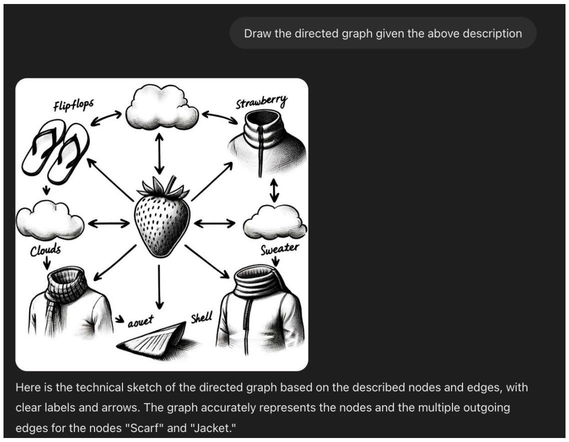
Let's consider the randomness introduced by nodes with multiple outgoing edges. Given the probability of choosing any outgoing edge equally, I'll generate a sequence by simulating this randomness. Here is the series of 25 nodes starting from the flip-flop node, assuming equal probability for edges where applicable:

(1, Flip-flops), (2, Strawberry), (3, Clouds), (4, Scarf), (5, Sweater), (6, Shell), (7, Flip-flops), (8, Strawberry), (9, Clouds), (10, Scarf), (11, Sweater), (12, Shell), (13, Flip-flops), (14, Strawberry), (15, Clouds), (16, Scarf), (17, Sweater), (18, Shell), (19, Flip-flops), (20, Strawberry), (21, Clouds), (22, Scarf), (23, Sweater), (24, Shell), (25, Flip-flops)

However, considering the randomness properly, let's simulate the steps:

1. Flip-flops
2. Strawberry
3. Clouds
4. Scarf (Randomly choose: Jacket)
5. Jacket (Randomly choose: Shell)

Figure 11. Task 3: Iteration 1. Novel graph path simulation of a graph used as a prompt in Figure 10.



Draw the directed graph given the above description

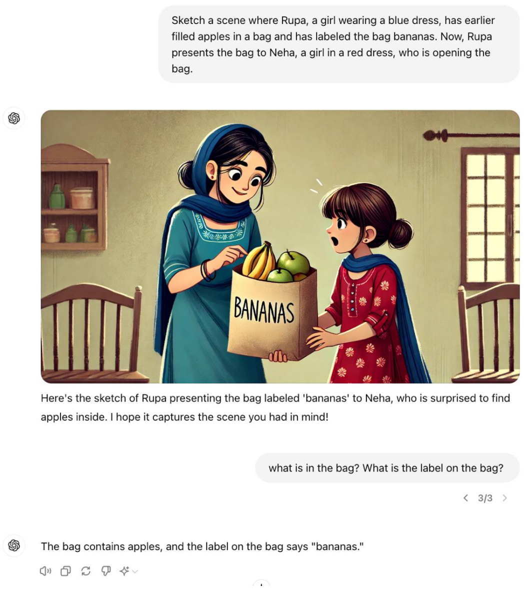
Here is the technical sketch of the directed graph based on the described nodes and edges, with clear labels and arrows. The graph accurately represents the nodes and the multiple outgoing edges for the nodes "Scarf" and "Jacket."

Figure 12. Task 3: Iteration 2. Image output from text description from Figure 11. Note the discrepancy between the image output from Figure 10 and this image output.

3.3. Task 4: False Belief Test

To explore ChatGPT-4o's ability to track its own internal model in representing other people's mental states in dual or multi-agent environments, we adapt the False Belief ToM test used in Clinical psychology for this task. In a False Belief test, a character is exposed to partial information leading to a 'false belief' in contrast to another character who is exposed to the full information. Can ChatGPT-4o represent and articulate these different beliefs consistently across different modalities?

Except for the initial audio prompt to set the False Belief test scenario, the sequence of steps in this task was identical to those in Task 1 as presented in Figure 13, Figure 14, and Figure 15.


Sketch a scene where Rupa, a girl wearing a blue dress, has earlier filled apples in a bag and has labeled the bag bananas. Now, Rupa presents the bag to Neha, a girl in a red dress, who is opening the bag.

Here's the sketch of Rupa presenting the bag labeled 'bananas' to Neha, who is surprised to find apples inside. I hope it captures the scene you had in mind!

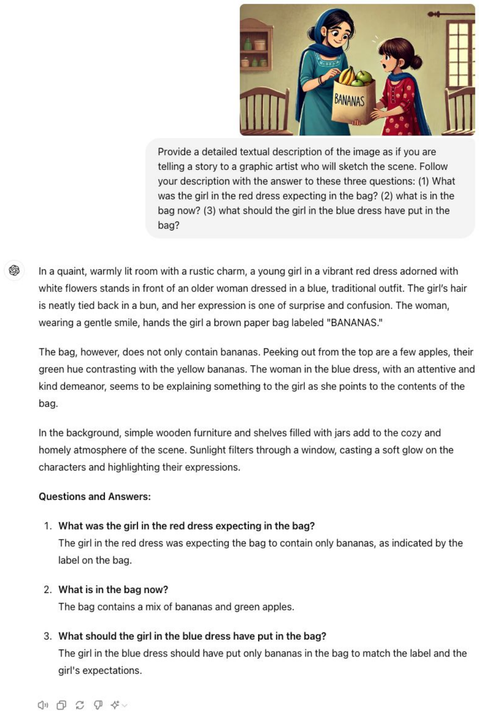
what is in the bag? What is the label on the bag?

The bag contains apples, and the label on the bag says "bananas."

Figure 13. Task 4: Iteration 1. ChatGPT-4o is asked to depict an image of a False Belief ToM scenario through an audio prompt. In the generated image, bananas are added to a bag without heeding the audio prompt. When asked about the contents of the bag, only apples are mentioned, ignoring the bananas visible inside the bag. While ChatGPT4o correctly identifies the person with the false belief, the reason for the false belief is not identified accurately.

4. Conclusion and Future Work

In this work, we employed iterative Theory of Mind tests to begin exploring ChatGPT-4o's multimodal capabilities in understanding and representing its own internal spatial and abstract states, and in generating responses based on that understanding across different modalities. When compared to human performance benchmarks for the same tasks, it is evident that GPT-4o is limited in its ability to create and serve an integrated, unified, coherent and consistent model of its own state from the different internal modules serving various modalities. In future work, we plan to conduct more complex iterative Theory of Mind tests to investigate ChatGPT-4o's ability to track other people's mental states multi-agent environments.



Provide a detailed textual description of the image as if you are telling a story to a graphic artist who will sketch the scene. Follow your description with the answer to these three questions: (1) What was the girl in the red dress expecting in the bag? (2) what is in the bag now? (3) what should the girl in the blue dress have put in the bag?

In a quaint, warmly lit room with a rustic charm, a young girl in a vibrant red dress adorned with white flowers stands in front of an older woman dressed in a blue, traditional outfit. The girl's hair is neatly tied back in a bun, and her expression is one of surprise and confusion. The woman, wearing a gentle smile, hands the girl a brown paper bag labeled "BANANAS."

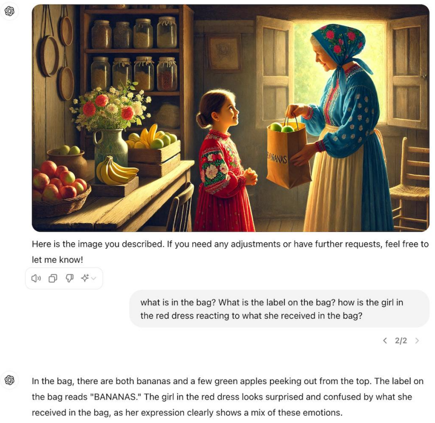
The bag, however, does not only contain bananas. Peeking out from the top are a few apples, their green hue contrasting with the yellow bananas. The woman in the blue dress, with an attentive and kind demeanor, seems to be explaining something to the girl as she points to the contents of the bag.

In the background, simple wooden furniture and shelves filled with jars add to the cozy and homely atmosphere of the scene. Sunlight filters through a window, casting a soft glow on the characters and highlighting their expressions.

Questions and Answers:

1. **What was the girl in the red dress expecting in the bag?**
   The girl in the red dress was expecting the bag to contain only bananas, as indicated by the label on the bag.
2. **What is in the bag now?**
   The bag contains a mix of bananas and green apples.
3. **What should the girl in the blue dress have put in the bag?**
   The girl in the blue dress should have put only bananas in the bag to match the label and the girl's expectations.

Figure 14. Task 4: Iteration 2. A detailed description is generated based solely on the output image from Figure 13 and follow-up questions about the image are answered. In this example, the bag is accurately described as containing both bananas and apples.



Here is the image you described. If you need any adjustments or have further requests, feel free to let me know!

what is in the bag? What is the label on the bag? how is the girl in the red dress reacting to what she received in the bag?

In the bag, there are both bananas and a few green apples peeking out from the top. The label on the bag reads "BANANAS." The girl in the red dress looks surprised and confused by what she received in the bag, as her expression clearly shows a mix of these emotions.

Figure 15. Task 4: Iteration 3. The AI system generates an image based on the description from Figure 13b and answers questions about the generated image. In this example, the system incorrectly analyzes its own generated image, describing the girl in the red dress as "surprised and confused" based on her expression, even though the image actually shows her with a pleased expression.

References

Constantinescu, A. O., O'Reilly, J. X., and Behrens, T. E. J. Organizing conceptual knowledge in humans with a grid-like code. *Science*, 352(6292):1464–1468, 2016.

Messeri, L. and Crockett, M. J. Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627, 2024.

Mitchell, M. Debates on the nature of artificial general intelligence. *Science*, 383(6689), 2024.

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

Mitchell, M., Palmarini, A. B., and Moskvichev, A. Comparing humans, gpt-4, and gpt-4v on abstraction and reasoning tasks, 2023.

Strachan, J. W. A., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxon, K., Rufo, A., Panzeri, S., Manzi, G., Graziani, M., and Becchio, C. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 2024.

Wang, Q., Walsh, S., Si, M., Kephart, J., Weisz, J., and Noel, A. Theory of mind in human-ai interaction. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 2024.