nature machine intelligence

Supplementary information

https://doi.org/10.1038/s42256-024-00963-y

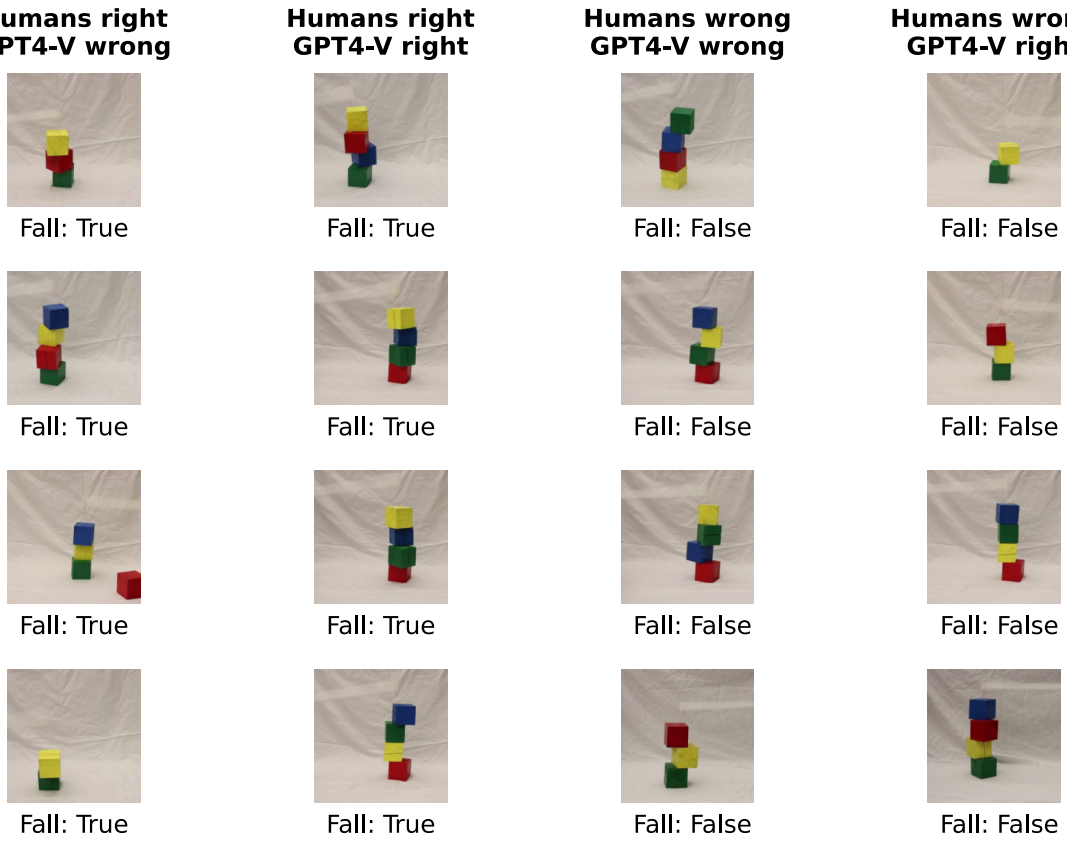
Visual cognition in multimodal large language models

In the format provided by the authors and unedited

Supplementary information

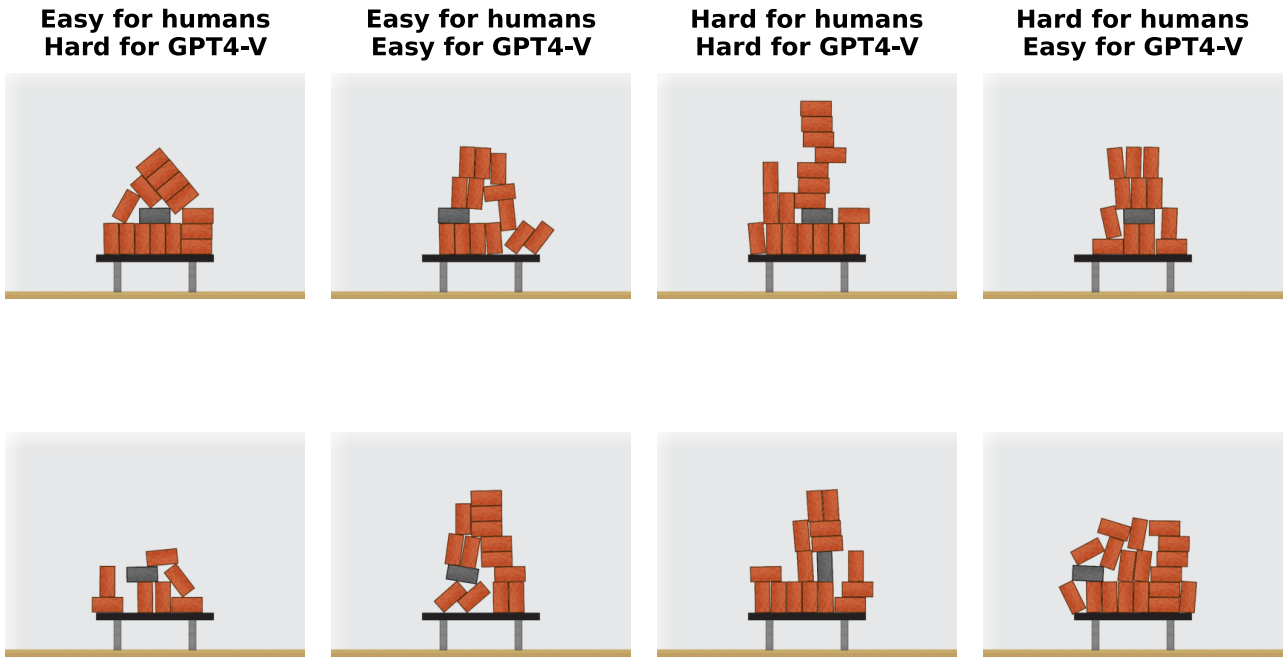
1 Example trials

1.1 Intuitive physics: Block towers



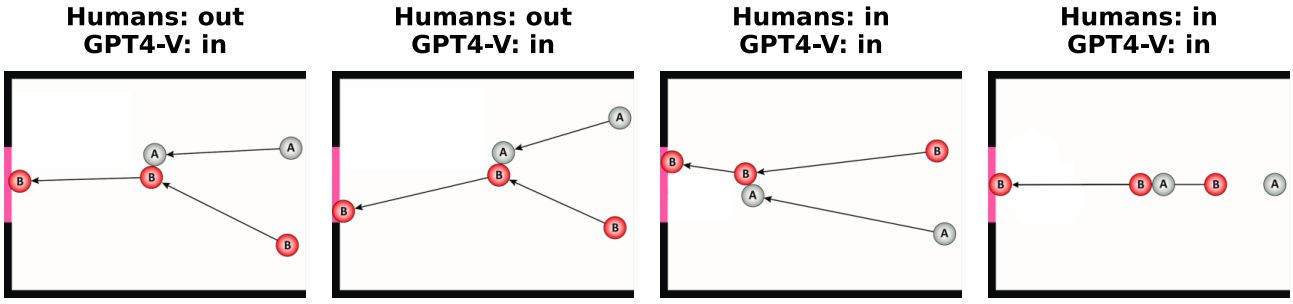
Supplementary Figure 1. Example images from Lerer et al.[99] which are either challenging or easy for Humans, GPT4-V, or both in the binary tower stability judgement task for physical intuition.

1.2 Causal reasoning: Jenga



Supplementary Figure 2. Example images from Zhou et al.[101] which are either challenging or easy for Humans, GPT4-V, or both in the number of blocks that will fall task for causal reasoning. Easy here refers to images where judgements have a low average distance to the that the ground truth. Hard refers to images where judgements have a high average distance to the ground truth.

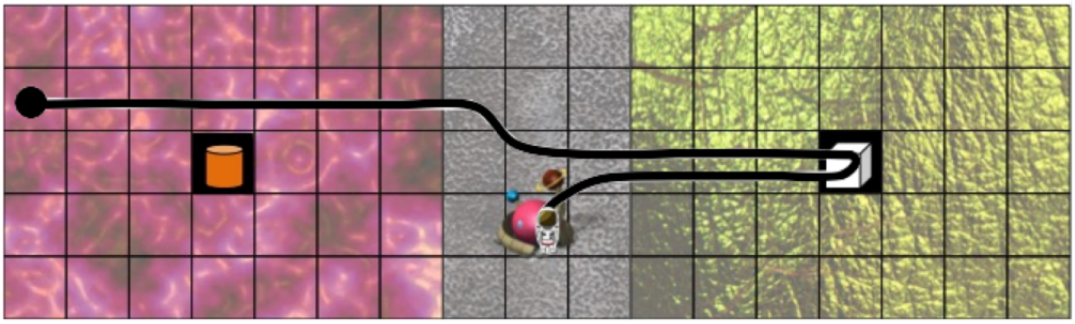
1.3 Causal reasoning: Michotte



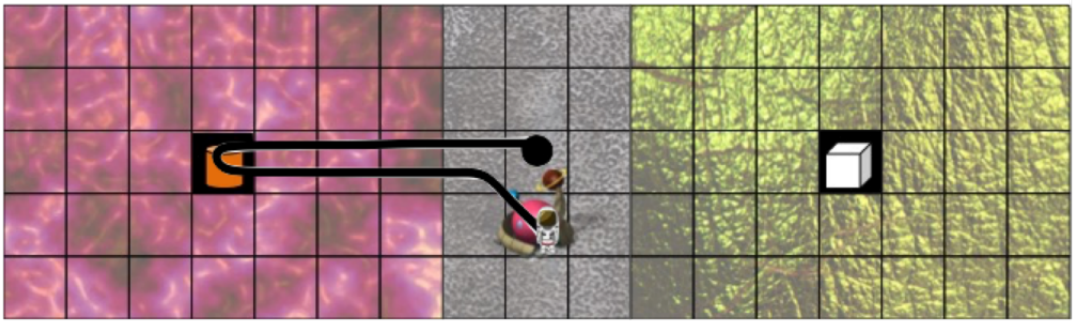
Supplementary Figure 3. Example images from Gerstenberg et al.[52]. Humans and models were asked if ball "B" would have gone through the gate if ball "A" had not been present in the scene.

1.4 Intuitive psychology: Astronaut

A

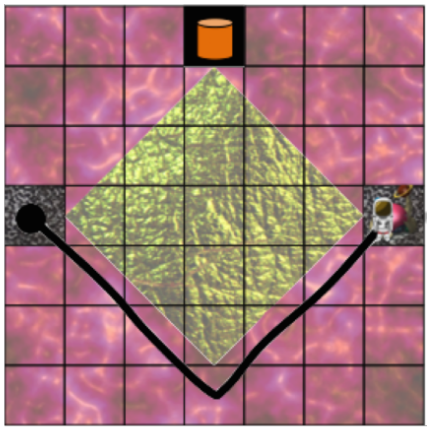


B

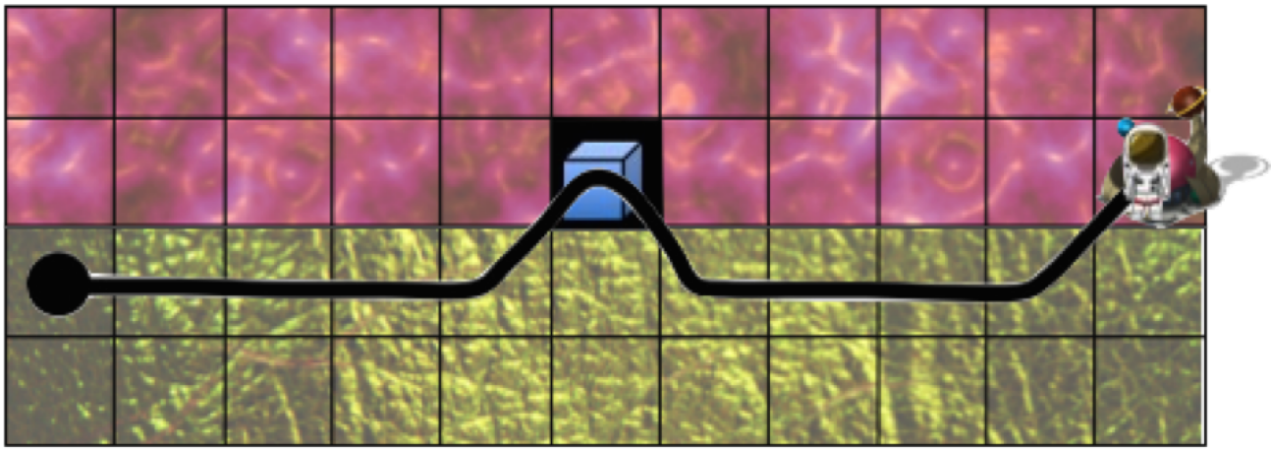


Supplementary Figure 4. Example images from Jara-Ettinger et al.[104], where the ratings given by GPT4-V match or diverge from human psychological intuition. **A**: GPT4-V answers 7, 5, 2, 7 for left cost, right cost, left reward, right reward. These values make sense given the path of the agent. It crosses straight through the left terrain, indicating that it might have a high cost associated with it and that the agent does not find the left care package rewarding. It then crosses through the yellow terrain to pick up the white care package, indicating that this package has a high reward associated with it. **B**: GPT4-V answers 5, 4, 2, 7 for left cost, right cost, left reward, right reward. These values are counterintuitive. The agent is seen crossing into the left terrain and picking up the left reward, which indicates that the left care package has a significant reward associated with it and that the left terrain should not incurr a large cost upon the agent. GPT4-V however assigns the left terrain a higher cost than the right terrain and also the left care package a lower reward than the right care package.

A

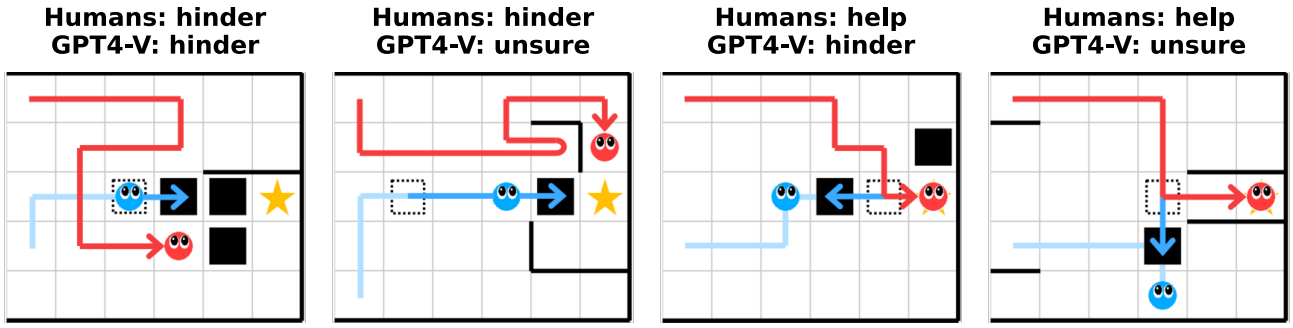


B



Supplementary Figure 5. Example images from from Jara-Ettinger et al.[104]. **A**: GPT4-V answers 7, 7, 0 for inside cost, outside cost, top reward. That the care package has no value for the agent is sensible given the image, however we would expect the costs for inside and outside terrains to diverge, as the agent seemingly takes a detour to avoid crossing the inside terrain. **B**: GPT4-V answers 5, 5, 7 for bottom cost, top cost, top reward. The reward here should have a high value seeing as the agent takes a detour to collect the care package. However, the bottom terrain should be associated with a lower cost compared to the top terrain, as the agent is seen taking a detour to cross the bottom terrain instead of the top terrain after picking up the care package.

1.5 Intuitive psychology: Help or hinder

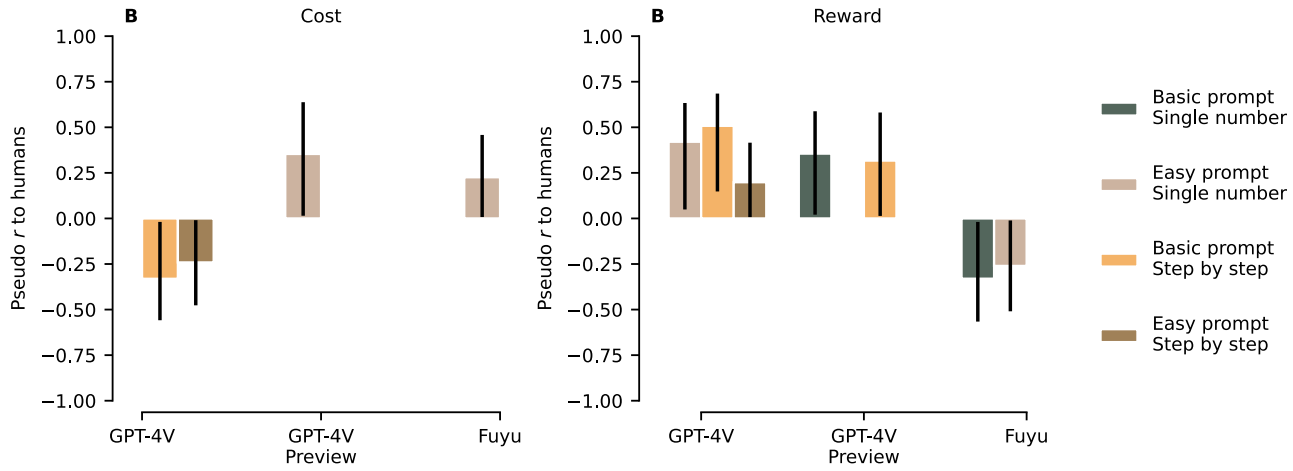


Supplementary Figure 6. Example images from Wu et al.[105]. Humans and models were asked if the red agent would have succeeded in reaching the star if the blue agent had not been present in the scene.

2 Prompting analysis

In this additional analysis, we explore two different types of context and response constraints: the standard context and a simplified context. The standard context is close to the original task description used in the human experiment. The simplified context is a modified version of the standard context that is closer to the grid level and therefore should require less abstraction from the input image: "The image shows a grid of squares. The white figure on the center right square is the astronaut. Each square represents a single block of terrain that the astronaut can cross in one step [...]".

We also explore two different types of response constraints: a short single number response and a step by step reasoning response. For the single number response, the number of output tokens is set to 3 and we begin the prompt with "Please answer the following question with a number only:" and end it with "You are only allowed to answer with a number!". For the step by step response, we allow for 500 output tokens and begin the prompt with "Answer the following question and give your reasoning why:" and end it with "Let's think step by step and then give your final answer as a single number in the format [X]." We test all four resulting combinations for GPT-4V, GPT-4V-Preview and Fuyu.



Supplementary Figure 7. Results for the prompting analysis. We find that no combination of context and response constraints works reliably for all models for (A) cost and (B) reward questions in the "Intuitive psychology: Astronaut" experiment. Note that the missing bars are due to uninterpretable or constant model responses. Bars in plots A & B show the square root of the R^2 values for Bayesian logistic mixed effects regressions with error bars given by the square root of the 95% percentiles for this R^2 value (n = 26 and 25, respectively).

We find that, for all models, the correlation coefficients to human answers change substantially depending on the prompting variation. For GPT-4V-Preview, the computed coefficients are most stable, however at the same time there are a lot of missing conditions where the model either returns uninterpretable or constant model responses. Noteably, it seems that GPT-4 gives responses that correlate most strongly with human answers in the step by step conditions. The highest correlation to humans is achieved for reward judgements in the basic context and the step by step response. However, this comes at the *cost* of a negative cost correlation with humans. Fuyu, on the other hand, does not produce sensible outputs when given the instruction to perform step by step reasoning. For both cost and reward questions, it keeps repeating uninterpretable sentences that rephrase the input (such as "The white object on a scale of 7 (not a lot) is rewarding for the astronaut because it represents multiple blocks of terrain that can be crossed simultaneously [...]"). Please note that while the basic prompt and single number condition is analogous to the prompting strategy used for the main experiments, the results reported here are different. This is because we only investigate experiment 1A from Jara-Ettinger et al.[104] for the prompting analysis, while we report results summarized over experiments 1A-C in the main text.

2.1 Conclusion

We show that prompting strategies affect behavior for all models for the intuitive psychology astronaut task. While strategies such as step by step reasoning can improve the performance of powerful models such as GPT-4V, the performance of less powerful models such as Fuyu can deteriorate (likely due to the increased prompt complexity). Additionally, we found that models can be impacted by small changes to prompts on a character level, likely due to the effects of tokenization. For example, in the counterfactual task for the Michotte experiment, we query Fuyu with:

> "Q: The scene shows two balls labeled "A", and "B". On the left side there is a pink gate. The solid arrows show the trajectories of the balls. Indicate your agreement with the assessment that ball "B" would have gone through the gate if ball "A" had not been present in the scene on a scale from 0 (not at all) to 100 (very much). You are only allowed to answer with a single number."

If we instead introduce the slight variation of removing the quotation marks around A and B in the second to last sentence: "Indicate your agreement with the assessment that ball B would have gone through the gate if ball A had not been present [...]", the ratings given by Fuyu switch completely for some sequences (from 0 to 100 or 100 to 0). While this small change heavily impacts Fuyu's behavior, GPT-4V-Preview only switches from 100 to 0 for one sequence, and Claude-3 only varies ratings between 90 and 100. Taken together these results not only indicate that model behavior can be sensitive to small changes to prompts on a character level but also that the behavior of more capable models might be less impacted.

Prompts

Intuitive physics: Block towers

For the intuitive physics experiment, we used a task from Lerer et al.[99]. For each trial, we asked the models three questions:

Question 1	What is the background color?
Question 2	What are the colors of the blocks from top to bottom? You are only allowed to answer with the color names.
Question 3	Will this block tower fall? Give a boolean answer.

Causal reasoning: Jenga

For the first causal reasoning experiment, we used a task from Zhou et al.[101,102]. For each trial, we again asked the models three questions:

Question 1	How many blocks are there in the image? You are only allowed to answer with a single number. No words allowed!
Question 2	How many of the red bricks would fall off the table if the dark grey brick wasn't there? You are only allowed to answer with a single number between 0 and {num_blocks} corresponding to how many blocks would fall. No words allowed!
Question 3	How responsible is the dark grey brick for the red bricks staying on the table? You are only allowed to answer with a number on a scale from 0% (not at all responsible) to 100% (fully responsible). No words allowed!

Causal reasoning: Michotte

For the second causal reasoning experiment, we used a task from Gerstenberg et al.[52]. For each trial, we asked the models four questions:

Question 1	The scene shows two balls labeled "A", and "B". On the left side there is a pink gate. The solid arrows show the trajectories of the balls. What is the background color of the image? You are only allowed to answer with a single color name!
Question 2	The scene shows two balls labeled "A", and "B". On the left side there is a pink gate. The solid arrows show the trajectories of the balls. Do the balls move "left to right" or "right to left"? You are only allowed to respond with one of the two options!
Question 3	The scene shows two balls labeled "A", and "B". On the left side there is a pink gate. The solid arrows show the trajectories of the balls. Indicate your agreement with the assessment that ball "B" [completely missed / went through the middle of] the gate on a scale from 0 (not at all) to 100 (very much). You are only allowed to answer with a single number.
Question 4	The scene shows two balls labeled "A", and "B". On the left side there is a pink gate. The solid arrows show the trajectories of the balls. Indicate your agreement with the assessment that ball "B" would have gone through the gate if ball "A" had not been present in the scene on a scale from 0 (not at all) to 100 (very much). You are only allowed to answer with a single number.

Intuitive psychology: Astronaut

For the first intuitive psychology experiment, we used three tasks from Jara-Ettinger et al.[104]. We again asked the model two descriptive questions in order to assess their basic comprehension of the scene:

Question 1	Please answer the following question with a single color name only: What color is the background of the central part of the image? You are only allowed to answer with a single color name!
Question 2	Please answer the following question with a number only: How many orange or white containers are in the image? You are only allowed to answer with a number!

For experiment 1A we first gave the models the following basic prompt, which was combined with different trial specific questions:

> "This task is about astronauts. The astronauts are exploring planets with alien terrains depicted with different colours and textures. Each astronaut has different skills, making each terrain more or less exhausting or easy for them to cross. All astronauts can ultimately cross all terrains, even if it's exhausting. The astronauts land far from the base and have to walk there. In each image, the black circle on the left indicates where the astronaut landed. The base is on the middle right part of the image. Sometimes care packages are dropped from above and the astronauts can pick them up. There are two kinds of care packages depicted with an orange cylinder and a white cube. Each astronaut has different preferences and likes each kind of care package in different amounts. The astronauts don't actually need the care packages. They can go straight to the base, or they can pick one up. You will see images of different astronauts with different skills and preferences travelling from their landing location to the home base. The astronauts always have a map. So they know all about the terrains and the care packages. Please answer the following question with a number only:"

For experiment 1B we first gave the models the following basic prompt, which was again combined with different trial-specific questions:

> "This task is about astronauts. The astronauts are exploring planets with alien terrains depicted with different colours and textures. Each astronaut has different skills, making each terrain more or less exhausting or easy for them to cross. All astronauts can ultimately cross all terrains, even if it's exhausting. Sometimes, the astronauts land far from the base and have to walk there. In each image, the black circle indicates where the astronaut landed. The base is in the center of the image. Sometimes care packages are dropped from above and the astronauts can pick them up. There are two kinds of care packages depicted with an orange cylinder and a white cube. Sometimes both care packages are identical. The astronauts cannot pick both care packages. Each astronaut has different preferences and likes each kind of care package in different amounts. The astronauts don't actually need the care packages. They can go straight to the base, or they can pick one up. You will see images of different astronauts with different skills and preferences travelling from their landing location to the home base. The astronauts always have a map. So they know all about the terrains and the care packages. Please answer the following question with a number only:"

Finally, for experiment 1C we first gave the models the following basic prompt:

> "This task is about astronauts. The astronauts are exploring planets with alien terrains depicted with different colours and textures. Each astronaut has different skills, making each terrain more or less exhausting or easy for them to cross. All astronauts can ultimately cross all terrains, even if it's exhausting. The astronauts land far from the base and have to walk there. In each image, the black circle on the left indicates where the astronaut landed. The base is on the right part of the image. The path astronauts take from where they land to their base is indicated by a thick black line between the black circle on the left and the astronaut on the right. Sometimes care packages depicted by a blue cube on a black background are dropped from above and the astronauts can pick them up. Each astronaut has different preferences and likes each care package in different amounts. The astronauts don't actually

need the care packages. They can go straight to the base, or they can pick one up. You will see images of different astronauts with different skills and preferences travelling from their landing location to the home base. Your task is to judge how easy/exhausting it is for the astronaut in each image to cross each terrain, and how much they like each care package. The astronauts always have a map. So they know all about the terrains and the care packages. Please answer the following question with a number only:"

Experiment 1A

Images	Questions
0, 1	1. How easy is it for the astronaut to cross the pink terrain on a scale from 0 (extremely easy) to 10 (extremely exhausting)? You are only allowed to answer with a number! 2. How much does the astronaut like the orange care package on a scale from 0 (not at all) to 10 (a lot)? You are only allowed to answer with a number!
2, 3, 4, 5	1. How easy is it for the astronaut to cross the pink terrain on a scale from 0 (extremely easy) to 10 (extremely exhausting)? You are only allowed to answer with a number! 2. How much does the astronaut like the white care package on a scale from 0 (not at all) to 10 (a lot)? You are only allowed to answer with a number! 3. How much does the astronaut like the orange care package on a scale from 0 (not at all) to 10 (a lot)? You are only allowed to answer with a number!"
6, 7, 8, 9, 10	1. How easy is it for the astronaut to cross the purple terrain on a scale from 0 (extremely easy) to 10 (extremely exhausting)? You are only allowed to answer with a number! 2. How easy is it for the astronaut to cross the pink terrain on a scale from 0 (extremely easy) to 10 (extremely exhausting)? You are only allowed to answer with a number! 3. How much does the astronaut like the orange care package on a scale from 0 (not at all) to 10 (a lot)? You are only allowed to answer with a number!
11, 12, 13, 14, 15	1. How easy is it for the astronaut to cross the purple terrain on a scale from 0 (extremely easy) to 10 (extremely exhausting)? You are only allowed to answer with a number! 2. How easy is it for the astronaut to cross the pink terrain on a scale from 0 (extremely easy) to 10 (extremely exhausting)? You are only allowed to answer with a number! 3. How much does the astronaut like the white care package on a scale from 0 (not at all) to 10 (a lot)? You are only allowed to answer with a number! 4. How much does the astronaut like the orange care package on a scale from 0 (not at all) to 10 (a lot)? You are only allowed to answer with a number!"

Experiment 1B

Images	Questions
1, 2, 3, 4, 5, 6, 7	1. How easy is it for the astronaut to cross the pink terrain on a scale from 0 (extremely easy) to 10 (extremely exhausting)? You are only allowed to answer with a number! 2. How easy is it for the astronaut to cross the yellow terrain on a scale from 0 (extremely easy) to 10 (extremely exhausting)? You are only allowed to answer with a number! 3. How much does the astronaut like the orange care package on a scale from 0 (not at all) to 10 (a lot)? You are only allowed to answer with a number! 4. How much does the astronaut like the white care package on a scale from 0 (not at all) to 10 (a lot)? You are only allowed to answer with a number!
8, 9, 10	1. How easy is it for the astronaut to cross the pink terrain on a scale from 0 (extremely easy) to 10 (extremely exhausting)? You are only allowed to answer with a number! 2. How easy is it for the astronaut to cross the yellow terrain on a scale from 0 (extremely easy) to 10 (extremely exhausting)? You are only allowed to answer with a number! 3. How much does the astronaut like the orange care package on a scale from 0 (not at all) to 10 (a lot)? You are only allowed to answer with a number!
11, 12, 13, 14, 15, 16, 17	1. How easy is it for the astronaut to cross the pink terrain on a scale from 0 (extremely easy) to 10 (extremely exhausting)? You are only allowed to answer with a number! 2. How much does the astronaut like the orange care package on a scale from 0 (not at all) to 10 (a lot)? You are only allowed to answer with a number! 3. How much does the astronaut like the white care package on a scale from 0 (not at all) to 10 (a lot)? You are only allowed to answer with a number!

Experiment 1C

Images	Questions
All images	1. How easy is it for the astronaut to cross the yellow terrain on a scale from 0 (extremely easy) to 10 (extremely exhausting)? You are only allowed to answer with a number! 2. How easy is it for the astronaut to cross the pink terrain on a scale from 0 (extremely easy) to 10 (extremely exhausting)? You are only allowed to answer with a number! 3. How much does the astronaut like the blue care package on a scale from 0 (not at all) to 10 (a lot)? You are only allowed to answer with a number!

Intuitive psychology: Help or hinder

For the second intuitive psychology experiment, we used a task from Wu et al.[105]. For each trial, we gave the models the following basic prompt, followed by one of the four questions:

> "The scene shows two agents in a grid world in which agents and objects can interact. On each timestep, agents can move up, down, left, right, or stay in place, but cannot move through walls or boxes. One agent, RED, has a physical goal of reaching a star in 10 timesteps. If they run out of time, then they fail. Another agent, BLUE, has a social goal of helping or hindering RED. BLUE has the ability to push or pull boxes around."

Question 1	What is the background color of the image? You are only allowed to answer with a single color name!
Question 2	How many boxes are in the scene? You are only allowed to respond with a single number.
Question 3	What was BLUE intending to do? Give your answer on a scale from "definitely hinder RED" (0) to "definitely help RED" (100) with the midpoint "unsure" (50). You are only allowed to respond with a single number.
Question 4	How much do you agree that RED would have (still) succeeded if BLUE hadn't been there on a scale from "not at all" (0) to "very much" (100)? You are only allowed to answer with a single number.

Prompting analysis

For the prompting analysis, we used two different types of context (standard and simplified) and two response constraints (single number response and step by step reasoning). The standard context is close to the context of the original task and reads as follows:

> "This task is about astronauts. The astronauts are exploring planets with alien terrains depicted with different colours and textures. Each astronaut has different skills, making each terrain more or less exhausting or easy for them to cross. All astronauts can ultimately cross all terrains, even if it's exhausting. The astronauts land far from the base and have to walk there. In each image, the black circle on the left indicates where the astronaut landed. The base is on the middle right part of the image. Sometimes care packages are dropped from above and the astronauts can pick them up. There are two kinds of care packages depicted with an orange cylinder and a white cube. Each astronaut has different preferences and likes each kind of care package in different amounts. The astronauts don't actually need the care packages. They can go straight to the base, or they can pick one up. You will see images of different astronauts with different skills and preferences travelling from their landing location to the home base. The astronauts always have a map. So they know all about the terrains and the care packages."

The simplified context is a modified version of the standard context that is closer to the grid level and therefore should require less abstraction from the input image:

> "The image shows a grid of squares. The white figure on the center right square is the astronaut. Each square represents a single block of terrain that the astronaut can cross in one step. The astronaut moves along the black path that starts at the center left square and ends at the center right square. The squares have different colors which relate to how hard or easy they are for the astronaut to cross. In some squares there is also an orange cylinder or a white cube in front of a black back ground. These objects may be rewarding for the astronaut and he might choose to pick them up if it is worth it for him. The astronauts don't actually need the objects. They can go straight to the final square, or they can pick one up. You will see images of different astronauts travelling from the center left square to the center right square. The astronauts always have a map. So they know all about the terrains and the objects."

For the *single number response*, the context was followed by the sentence "Please answer the following question with a number only:", followed by the respective question and finally "You are only allowed to answer with a number!". For the *step by step response*, the context was followed by the sentence "Answer the following question and give your reasoning why:", followed by the respective question and finally "Let's think step by step and then give your final answer as a single number in the format [X]." The respective questions were the same as in experiment 1A of the "Intuitive physics: Astronaut" task:

Images	Questions
0, 1	1. How easy is it for the astronaut to cross the pink terrain on a scale from 0 (extremely easy) to 10 (extremely exhausting)? 2. How much does the astronaut like the orange care package on a scale from 0 (not at all) to 10 (a lot)?
2, 3, 4, 5	1. How easy is it for the astronaut to cross the pink terrain on a scale from 0 (extremely easy) to 10 (extremely exhausting)? 2. How much does the astronaut like the white care package on a scale from 0 (not at all) to 10 (a lot)? 3. How much does the astronaut like the orange care package on a scale from 0 (not at all) to 10 (a lot)?
6, 7, 8, 9, 10	1. How easy is it for the astronaut to cross the purple terrain on a scale from 0 (extremely easy) to 10 (extremely exhausting)? 2. How easy is it for the astronaut to cross the pink terrain on a scale from 0 (extremely easy) to 10 (extremely exhausting)? 3. How much does the astronaut like the orange care package on a scale from 0 (not at all) to 10 (a lot)?
11, 12, 13, 14, 15	1. How easy is it for the astronaut to cross the purple terrain on a scale from 0 (extremely easy) to 10 (extremely exhausting)? 2. How easy is it for the astronaut to cross the pink terrain on a scale from 0 (extremely easy) to 10 (extremely exhausting)? 3. How much does the astronaut like the white care package on a scale from 0 (not at all) to 10 (a lot)? 4. How much does the astronaut like the orange care package on a scale from 0 (not at all) to 10 (a lot)?