nature machine intelligence

Article

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

Visual cognition in multimodal large language models

Received: 29 November 2023

Accepted: 25 November 2024

Published online: 15 January 2025

Check for updates

Luca M. Schulze Buschoff[1,2,4] ✉, Elif Akata[1,2,3,4], Matthias Bethge[3] & Eric Schulz[1,2]

A chief goal of artificial intelligence is to build machines that think like people. Yet it has been argued that deep neural network architectures fail to accomplish this. Researchers have asserted these models' limitations in the domains of causal reasoning, intuitive physics and intuitive psychology. Yet recent advancements, namely the rise of large language models, particularly those designed for visual processing, have rekindled interest in the potential to emulate human-like cognitive abilities. This paper evaluates the current state of vision-based large language models in the domains of intuitive physics, causal reasoning and intuitive psychology. Through a series of controlled experiments, we investigate the extent to which these modern models grasp complex physical interactions, causal relationships and intuitive understanding of others' preferences. Our findings reveal that, while some of these models demonstrate a notable proficiency in processing and interpreting visual data, they still fall short of human capabilities in these areas. Our results emphasize the need for integrating more robust mechanisms for understanding causality, physical dynamics and social cognition into modern-day, vision-based language models, and point out the importance of cognitively inspired benchmarks.

People are quick to anthropomorphize, attributing human characteristics to non-human agents[1]. The tendency to anthropomorphize has only intensified with the advent of large language models (LLMs)[2]. LLMs apply deep learning techniques to generate text[3], learning from vast datasets to produce responses that can be startlingly human-like[4]. Astonishingly, these models cannot only generate text. When scaled up to bigger training data and architectures, other, so-called 'emergent abilities' appear[5,6]. The current models can, for example, pass the bar exam[7], write poems[8], compose music[9] and assist in programming and data analysis tasks[10]. As a result, the line between human and machine capabilities is increasingly blurred[11,12]. People not only interact with these systems as if they were humans[13], but they also start to rely on them for complex decision-making[14], artistic creation[15] and personal interactions[16]. It is, therefore, natural to ask: Have we built machines that think like people?

Judging whether or not artificial agents can mimic human thought is at the core of cognitive science[17,18]. Therein, researchers have long debated the capabilities of artificially intelligent agents[19–21]. In a seminal paper, Lake and colleagues[22] proposed core domains to consider when making such judgements. Published during the height of the deep learning revolution[23], the authors focused on domains that were easy for people but difficult for deep learning models: intuitive physics, causal reasoning and intuitive psychology.

Research on intuitive physics has studied how people perceive and interpret physical phenomena[24–26]. Past work on this topic has emphasized that humans possess an innate ability to predict and understand the physical properties of objects and their interactions[27], even from a young age[28], a notion sometimes summarized as a 'physics engine' in people's heads[29]. This understanding includes concepts such as gravity[30], inertia[31] and momentum[32]. Some of the most canonical tasks

[1]Max Planck Institute for Biological Cybernetics, Tübingen, Germany. [2]Institute for Human-Centered AI, Helmholtz Munich, Oberschleißheim, Germany. [3]University of Tübingen, Tübingen, Germany. [4]These authors contributed equally: Luca M. Schulze Buschoff, Elif Akata. ✉e-mail: lucaschulzebuschoff@gmail.com

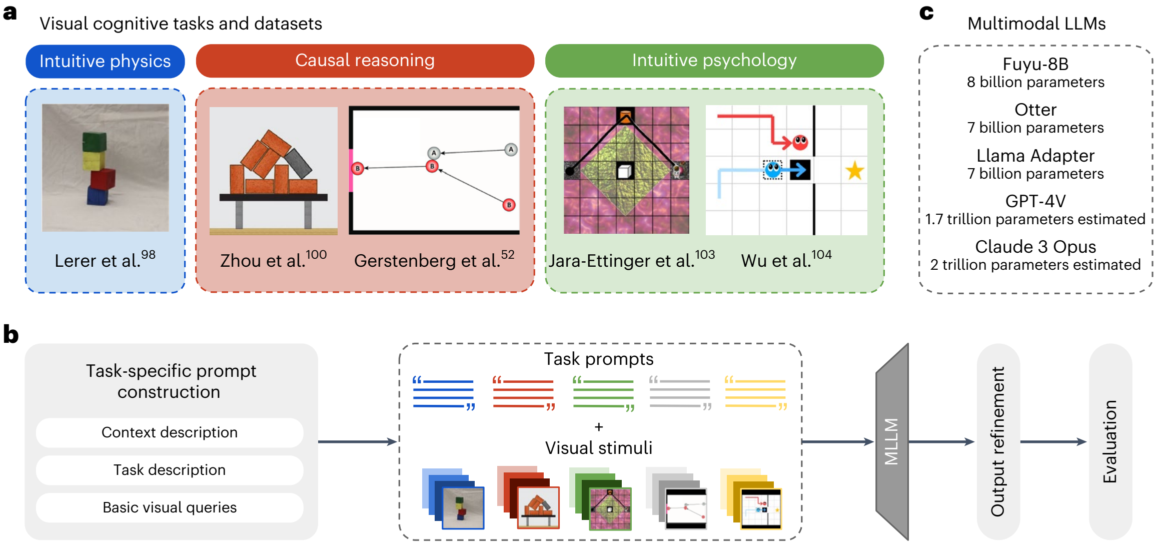

Fig. 1 | Overview of domains, tasks, approach and models. a, Example images for the different experiments. Each experiment was taken from one of three cognitive domains: intuitive physics, causal reasoning and intuitive psychology. **b**, General approach. For every query, an image was submitted to the model, and different questions were asked about the image, that is, we performed visual question answering. **c**, Used multimodal LLMs and their size. MLLM, multimodal LLM.

in this domain involve testing people's judgements about the stability of block towers[33,34]. These tasks have made their way into machine learning benchmarks[35,36], where they are used to test the intuitive physical understanding of neural networks (see ref. 37 for an overview of previous work on building models with human-like physical knowledge).

Research on causal reasoning has studied how individuals infer and think about cause–effect relationships[38–40]. Past work on this topic has proposed that humans possess an intuitive capacity to infer, understand and predict causal relationships in their environment[41–44], oftentimes described using Bayesian models of causal learning[45,46]. This cognitive ability encompasses recognizing patterns[47,48], inferring causes from interventions[49,50], and predicting future events based on hypothetical events[51]. Canonical tasks in this domain often involve assessing individuals' ability to infer causal relationships, for example, when judging the responsibility of one object causing other objects' movement[52,53]. Causal reasoning remains a challenge, even for current machine learning approaches[54,55].

Research on intuitive psychology has explored how individuals infer, understand and interpret social phenomena and mental states of other agents[56,57]. Past work on this topic has emphasized the concept that humans possess an inherent ability to infer and reason about the mental states[58,59], intentions and emotions of others, often referred to as a 'theory of mind'[60,61]. This ability has been modelled as a Bayesian inference problem[62–64]. Canonical tasks in this domain often involve assessing individuals' capacity to predict actions based on understanding others' perspectives or intentions, such as determining agents' utility functions based on their actions in a given environment[65,66]. It is the subject of ongoing debates whether modern algorithms show any form of intuitive psychology[67–69].

Lake and colleagues argued that some of these abilities act as 'start-up software', because they constitute cognitive capabilities present early in development. Moreover, they proposed that these so-called 'intuitive theories'[70,71] need to be expressed explicitly using the calculus of Bayesian inference[72], as opposed to being learned from scratch, for example, via gradient descent. However, with the increase in abilities of current neural networks, in particular LLMs, we pondered: Can LLMs, in particular vision LLMs, sufficiently solve problems from these core domains?

To address this question, we took canonical tasks from the domains of intuitive physics, causal reasoning and intuitive psychology that could be studied by providing images and language-based questions. We submitted them to some of the currently most advanced LLMs. To evaluate whether the LLMs show human-like performance in these domains, we follow the approach outlined in ref. 73: we treat the models as participants in psychological experiments. This allows us to draw direct comparisons with human behaviour on these tasks. Since the tasks are designed to test abilities in specific cognitive domains, this comparison allows us to investigate in which domains multimodal LLMs perform similar to humans, and in which they don't. Our results showed that these models can, at least partially, solve these tasks. In particular, two of the largest currently available models, OpenAI's Generative Pre-trained Transformer (GPT-4) and Anthropic's Claude-3 Opus, managed to perform robustly above chance in two of the three domains. Yet crucial differences emerged. First, none of the models matched human-level performance in any of the domains. Second, none of the models fully captured human behaviour, leaving room for domain-specific models of cognition such as the Bayesian models originally proposed for the tasks.

Related work

There have been a large number of studies on reasoning abilities in LLMs[74–76]. Previous studies have focused, among others, on testing LLMs' cognitive abilities in model-based planning[73], analogical reasoning tests[77], exploration tasks[78], systematic reasoning tests[79,80], psycholinguistic completion studies[81] and affordance understanding problems[82]. In this sense, our contribution can be seen as a part of a larger movement in which researchers use methods from the behavioural sciences to understand black box machine learning models[83–85]. However, most of the previous studies did not investigate multimodal LLMs but rather remained in the pure language domain. Although there have been recent attempts to investigate vision LLMs' cognitive features, including their reaction to visual illusions[86] as well as how they solve simple intelligence tasks[87], we investigate the proposed core components of cognition in these models.

Previous work has also looked at how LLMs solve cognitive tasks taken from the same domains that we have looked at. In intuitive physics, Zečević and colleagues[88] found that LLMs performed poorly in a task using language descriptions of physical scenarios. Zhang and colleagues[89] extracted programs from text produced by LLMs to improve their physical reasoning abilities. Finally, Jassim and colleagues[90] proposed a new benchmark for evaluating multimodal LLMs' understanding of situated physics. In causal reasoning, Binz

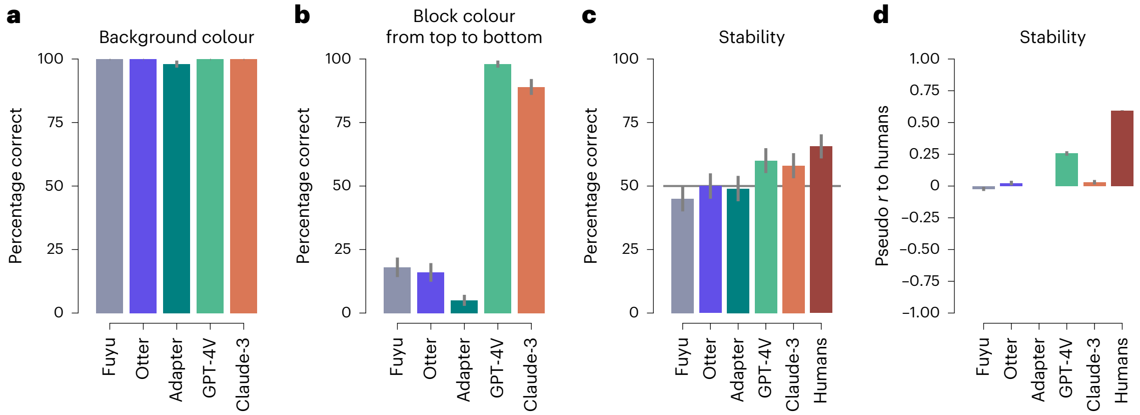

Fig. 2 | Results for five vision LLMs for tasks of increasing complexity given images of real block towers. a–c, We first ask for the background colour in the image (**a**) (images were taken from ref. 98), then the colour of blocks from top to bottom (**b**) and finally a binary stability rating for the block towers (**c**). **d**, The last plot shows the square root of the R^2 value for the Bayesian logistic mixed effects regression between models and human participants. Bars in plots **a–c** show percentage of correct answers with error bars given by the standard deviation of a binomial distribution ($n = 100$). Bars in plot **d** 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 ($n = 10{,}700$, number of images times number of human participants).

and Schulz[73] showed that GPT-3 failed at simple causal reasoning experiments, while Kosoy and colleagues[91] showed that LLMs cannot learn human-like causal over-hypotheses. In research on intuitive psychology, Kosinski argued that theory of mind might have emerged in LLMs[68] which has been criticized other researchers[69]. Akata and colleagues showed that GTP-4 plays repeated games very selfishly and could not pick up on simple conventions such as alternating between options[16]. Finally, Gandhi and colleagues[92] proposed a framework for procedurally generating theory of mind evaluations and found that GPT-4's abilities mirror human inference patterns, although less reliable, while all other LLMs struggled.

Many of the past studies on LLMs have fallen risk to appearing in new models' training sets. Recent work has recognized this issue and, in turn, evaluated language models on many problem variations to minimize training set effects[93]. Our work differs from these approaches as current models could not have just memorized solutions to the given problems because these problems require higher level reasoning. Furthermore, the human data and ground truth are most commonly stored in additional data files, which first have to be extracted and matched to the respective images to be used for model training. Since this requires data wrangling that cannot easily be automated and the number of stimuli to gain is so small, it is extremely unlikely that these stimuli together with the ground truth were entered into the training set of any of the investigated models.

Results

We tested five different models on three core components for human-like intelligence as outlined in ref. 22 (Fig. 1a). The models we used are vision LLMs, which are multimodal models that integrate image processing capabilities into LLMs[94,95] (Fig. 1c). These models allow users to perform visual question answering[96,97]: users can upload an image and ask questions about it, which the model interprets and responds to accordingly.

To test the three core components, we used tasks from the cognitive science literature that could be studied in vision LLMs via visual question answering. For every task, we queried the visual reasoning abilities of the LLMs with tasks of increasing complexity. First, we asked about simple features of the shown images such as the background colour or the number of objects shown. Afterwards, we submitted questions taken from the cognitive science experiments. We report results based on comparisons with the ground truth as well as the different models' matches to human data.

Intuitive physics with block towers

To test the intuitive physics capabilities of the different LLMs, we used photographs depicting wooden block towers from ref. 98 (see Supplementary Fig. 1 for an example). We first asked models to determine the background colour of the image. All four models achieved almost perfect accuracy (Fig. 2a). We then asked models to state the colour of blocks from top to bottom. Here, the performance of most models except for GPT-4V and Claude-3 deteriorated (Fig. 2b). Please note that the first two tasks are fairly trivial for humans and we would expect human performance to be at 100% (the background colour is always white and images featured two, three or four blocks in primary colours).

To test the models' physical reasoning abilities, we asked them to give a binary stability judgement of the depicted block towers. Here, only GPT-4V and Claude-3 performed slightly above chance (Fig. 2c; for GPT-4V, Fisher's exact test yielded an odds ratio of 2.597 with a one-sided P value of 0.028). None of the other models performed significantly above chance (the second best performing model, Claude-3, had an odds ratio of 2.016, with a one-sided P value of 0.078). Human participants were also not perfect but showed an average accuracy of 65.608%.

Finally, we determined the relationship between models' and humans' stability judgements using a Bayesian logistic mixed effects regression. We compute a Bayesian R^2 for each regression model based on draws from the modelled residual variances[99]. We then take the square root of this Bayesian R^2 and multiply it with the sign of the main regression coefficient to arrive at a pseudo r value. Around this pseudo r value we plot the square root of the 95% percentiles for the R^2 value (Fig. 2d). We found that GPT-4V was the only model that showed a relation to human judgements, with a regression coefficient of 1.15 (95% credible interval (95% CI) 1.04, 1.27) and an R^2 value of 0.066. However, the regression coefficient between individual humans and the mean over humans was still larger, with a coefficient of 1.46 (95% CI 1.41, 1.52) and an R^2 value of 0.354.

Causal reasoning with Jenga

To test the models' causal reasoning capabilities, we used synthetic images from refs. 100,101, which depicted block towers that were stable but might collapse if one of the blocks was removed (see Supplementary Fig. 2 for an example). We started by asking the models to count the blocks in the image. The images in this task displayed a larger number of blocks (ranging from 6 to 19), which made the basic counting task significantly more challenging than in the previous section. Models' responses

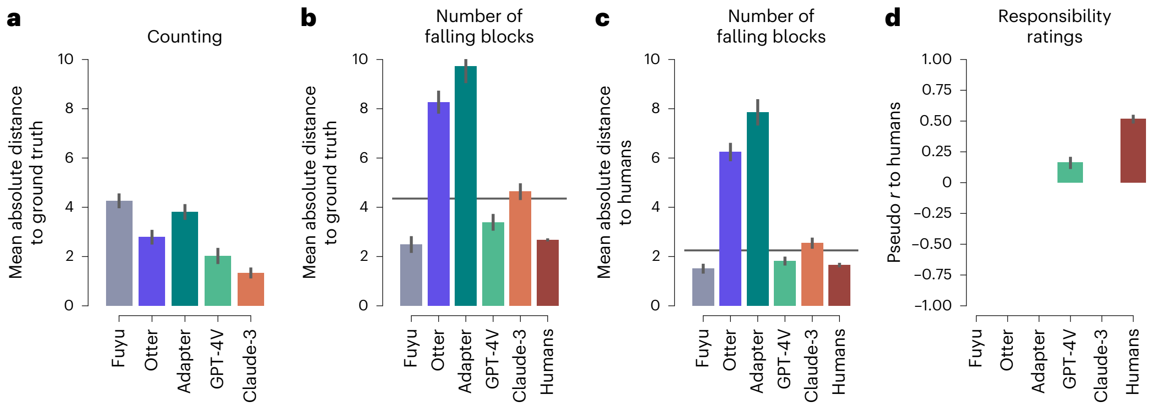

Fig. 3 | Results for Jenga causal reasoning experiment. a–d, We first ask for the number of blocks in the image (**a**), then we ask for the number of blocks that would fall if a specific block is removed and compute the absolute distance to the ground truth (**b**) as well as the absolute distance to human judgements (**c**) and finally a rating between 0 and 100 for how responsible a specific block is for the stability of the tower (**d**). The causal reasoning experiment was taken from ref. 100. For the responsibility ratings, all LLMs except for GPT-4V give constant ratings: Fuyu and Claude-3 always respond with 100, while Otter and LLaMA-Adapter V2 always respond with 50. Bars in plots **a** and **b** show absolute distance to ground truth with error bars given by the standard error of the mean ($n = 42$). Bars in plot **c** show the distance to human answers with errors bars again given by the standard error of the mean ($n = 41$). Bars in plot **d** 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 ($n = 1,470$).

approximated the ground truth, albeit rarely matching it exactly. Therefore, we report the mean absolute distance to the ground truth instead of the percentage of correct answers (Fig. 3a). The models' performance highlighted the challenging nature of this task, with the best performing model (Claude-3) still being on average more than one block off.

We continued by querying the models for the number of blocks that would fall if a specific block was removed from the scene (Fig. 3b,c). We established a baseline performance represented by a horizontal line in Fig. 3b,c, which corresponds to a random agent that gave the mean between 0 and the number of blocks in each image as its' prediction, essentially behaving like a uniform distribution over the possible number of blocks that could fall. Notably, both GPT-4V and Fuyu-8B surpassed the random baseline, their performance levels being close to the human results reported in ref. 100, which is depicted by the rightmost bar in the plot. However, GPT-4V still diverges significantly from the average over human participants ($t(42) = 2.59, P < 0.05$).

Finally, we asked the models to rate the responsibility of a specific block for the stability of the other blocks (Fig. 3d). Notably, all models except for GPT-4V gave constant ratings for this task (Fuyu and Claude-3 always responded with 100, while Otter and LLaMA-Adapter V2 always responded with 50). The regression coefficient for GPT-4V with human values is 0.16 (95% CI 0.10, 0.21) with an R^2 value of 0.027. The human-to-human regression has a coefficient of 0.54 (95% CI 0.45, 0.63) and an R^2 value of 0.268.

Causal reasoning with Michotte

For the second test for causal reasoning abilities, we ran an experiment from ref. 52 that is based on the classic Michotte launching paradigm[102]. It uses simple synthetic two-dimensional (2D) depictions of two balls labelled 'A' and 'B' with arrows showing their trajectories in front of a white background (see Supplementary Fig. 3 for an example). We started by asking the models to determine the background colour of the image (Fig. 4a). Most models perform fine with the exception of Fuyu, which always answers 'pink' (probably since pink is mentioned as the colour of the gate in the prompt). Then, we asked models to infer the trajectory of ball movement. This proved challenging for most models (Fig. 4b), which is surprising given that the prompt explicitly mentions that the arrows in the stimuli depict the trajectory of the balls and the balls always move from right to left.

We then queried the models for their agreement on a scale from 0 to 100 with the following questions: either 'Ball B went through the middle of the gate' (if ball B entered the gate) or 'Ball B completely missed the gate' (if ball B missed the gate) (Fig. 4c). No model performs close to the human results reported in ref. 52. The best performing model is Fuyu with a regression coefficient of 0.26 (95% CI −0.08, 0.61) and an R^2 value of 0.067. Interestingly, Claude-3 shows a negative relationship with human judgements, with a regression coefficient of −0.22 (95% CI −0.39, −0.06) and an R^2 value of 0.076. The human-to-human regression coefficient is 0.85 (95% CI 0.69, 1.03) with an R^2 value of 0.556.

Finally, we asked the models for their agreement on a scale from 0 to 100 with the counterfactual question of whether 'Ball B would have gone through the gate had Ball A not been present in the scene' (Fig. 4d). Notably, the closed-source models perform worse than some open-source models for both tasks. Here, Fuyu is again the best performing model with a regression coefficient of 0.42 (95% CI 0.28, 0.57) and an R^2 value of 0.185. Pseudo r values for LLaMA-Adapter V2 and GPT-4V are missing, since the former gave only non-valid answers and latter always responded with 100. The human-to-human regression coefficient is 0.85 (95% CI 0.76, 0.93) with an R^2 value of 0.698.

Intuitive psychology with the astronaut task

As a first test for the intuitive psychology understanding of the different LLMs, we used synthetic images depicting an astronaut on a coloured background from ref. 103 (see Supplementary Figs. 4 and 5 for an example). The images featured different terrains and care packages. Depending on which terrain the astronaut crossed or which care package they chose to pick up, it was possible to infer the costs associated with the terrains and rewards associated with the care packages.

Again, we first tasked models with determining the background colour of the images. Here, the performance of the models was worse compared with the intuitive physics dataset (Fig. 5a), which might be due to the fact that the background colour here was not uniform (Supplementary Fig. 5). We then asked models to count the number of care packages in the scene. Most models except for GPT-4V struggled here (Fig. 5b).

Afterwards, we asked them to infer the costs associated with the different terrains (Fig. 5c) and the rewards associated with different care packages (Fig. 5d). All models only showed weak relations with the average over human participants in their judgements about the costs and rewards associated with the environment. The regression

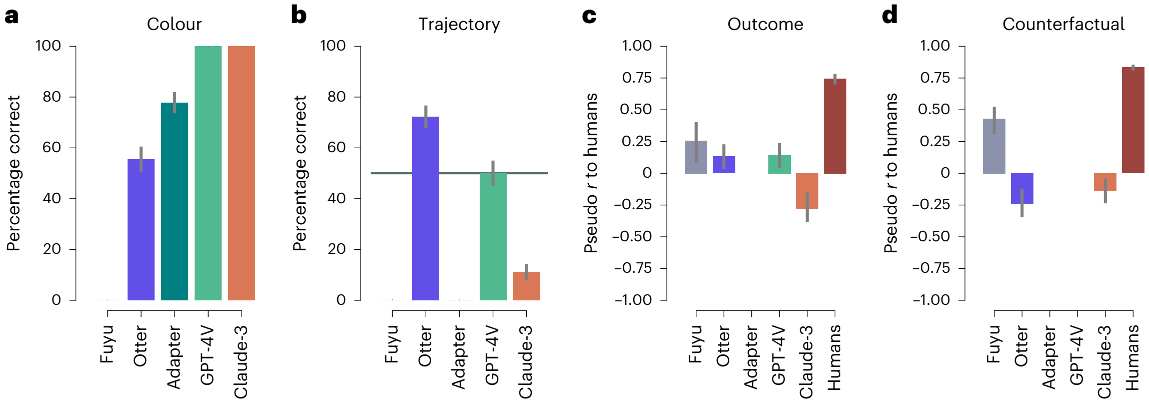

Fig. 4 | Results for Michotte causal reasoning experiment. a–**d**, We first ask for the background colour in the image (**a**), then the direction of ball movement (**b**), a judgement between 0 and 100 on whether ball 'B' goes through the gate (**c**) and finally a counterfactual judgement between 0 and 100 on whether ball 'B' would have gone through the gate, had ball 'A' not been present in the scene (**d**). The causal reasoning experiment was taken from ref. 52. Bars in plots **a** and **b** show percentage of correct answers with error bars given by the standard deviation of a binomial distribution ($n = 18$). Bars in plots **c** and **d** 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 ($n = 252$ and 234, respectively).

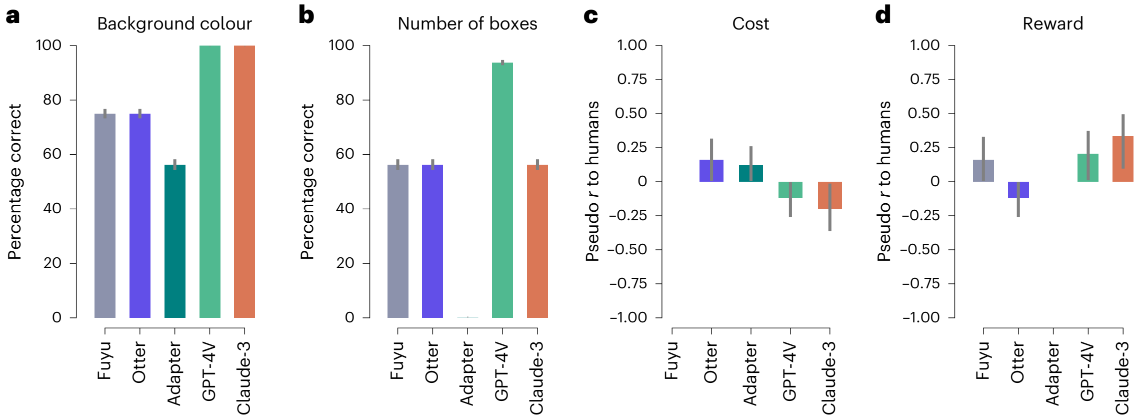

Fig. 5 | Results on astronaut task for intuitive psychology. a,**b**, Again, we first ask for the background colour (**a**) and the number of boxes in the scene (**b**). **c**,**d**, Models are then asked to make inferences about the costs (**c**) and rewards (**d**) in an environment depending on the path an agent has taken. The tasks for intuitive psychology were taken from ref. 103. Regression coefficients for Fuyu and LLaMA-Adapter V2 are missing as they always responded with constant ratings for either cost or reward questions. Bars in plots **a** and **b** show percentage of correct answers with error bars given by the standard deviation of a binomial distribution ($n = 16$). Bars in plots **c** and **d** 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 ($n = 81$ and 70, respectively).

coefficients of the models with the *z*-scaled mean over human participants ranged from −0.24 to 0.16 with R^2 values between 0.025 to 0.04 for cost questions, and from −0.02 to 0.39 (Claude-3, 95% CI 0.11, 0.66) with R^2 values between 0.015 and 0.110 for reward questions.

Intuitive psychology with the help or hinder task

The new intuitive psychology dataset we added is taken from ref. 104. This task shows a simple 2D depiction of two agents in a grid environment (see Supplementary Fig. 6 for an example). On each time step, the agents can move up, down, left or right, or stay in place, but cannot move through walls or boxes. The red agent has the objective of reaching a star in ten time steps. If the agent runs out of time they fail. The blue agent has the objective of either helping or hindering the red agent by pushing or pulling boxes around.

We first asked models to determine the background colour in the scene and to determine the number of boxes in the scene (Fig. 6a,b). The closed-source models are able to perfectly determine the background colour (always white) but they nonetheless struggle with determining the number of boxes in the scene (always 1, 2 or 3). Model answers for the counting task ranged from 1 to 4, with only LLaMA-Adapter V2 giving constant responses of 2.

We then asked the models whether the blue agent tried to help or hinder the red agent (Fig. 6c). Here, Otter shows the highest regression coefficient with human answers with 0.19 (95% CI 0.13, 0.25) and an R^2 value of 0.038. Claude-3 shows a negative relationship with human answers with a coefficient of −0.25 (95% CI −0.31, −0.20) and an R^2 value of 0.066. No model showed coefficients even close to the human-to-human coefficient of 0.93 (95% CI 0.90, 0.96) with an R^2 value of 0.858.

Finally, we asked the model whether the red agent would have succeeded in reaching the star, had the blue agent not been there. We show the square root of the R^2 for the Bayesian linear mixed effects regression with 95% percentiles in Fig. 6d. Interestingly, the results here flip, with Otter now showing a stronger negative relationship with a coefficient of −0.40 (95% CI −0.47, −0.33) and an R^2 value of 0.161 (this makes sense, since this task is essentially a counterfactual simulation question similar to Fig. 4d, where Otter already showed a negative relation to human judgements). GPT-4V and Claude-3 both show small positive regression coefficients with humans answers: 0.15 (95% CI 0.09, 0.21) with an R^2 value of 0.025, and 0.17 (95% CI 0.11, 0.23) with an R^2 value of 0.032, respectively. Again, no model coefficient is close to the human-to-human coefficient of 0.83 (95% CI 0.80, 0.87) with an R^2 value of 0.688.

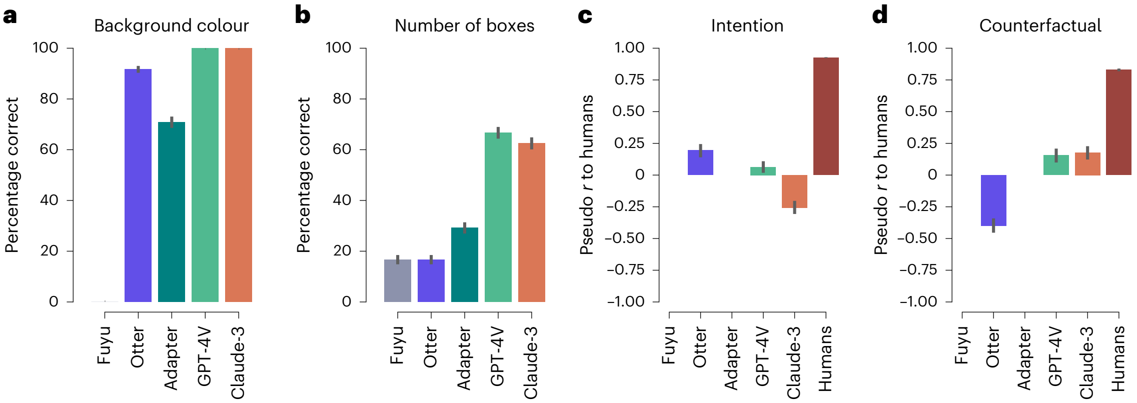

Fig. 6 | Results on help or hinder task for intuitive psychology. a–d, We first ask for the background colour in the image (**a**), then the number of boxes in the scene (**b**), a judgement between 0 and 100 on whether an agent in the scene tried to hinder the other agent (**c**) and finally a counterfactual judgement between 0 and 100 on whether an agent in the scene would have successfully reached the goal, had the other agent not been present (**d**). The intuitive psychology dataset was from taken from ref. 104. Bars in plots **a** and **b** show percentage of correct answers with error bars given by the standard deviation of a binomial distribution ($n = 24$). Bars in plots **c** and **d** 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 ($n = 1{,}200$).

Discussion

We started by asking whether, with the rise of modern LLMs, researchers have created machines that—at least to some degree—think like people. To address this question, we took four recent multimodal LLMs and probed their abilities in three core cognitive domains: intuitive physics, causal reasoning and intuitive psychology.

In intuitive physics and causal reasoning, the models managed to solve some of the given tasks and GPT-4V showed a slight match with human data. However, while they performed well in some tasks, the models did not show a conclusive match with human data for the causal reasoning experiments. Finally, in the intuitive psychology tasks, none of the models showed a strong match with human data. Thus, an appropriate answer to the question motivating our work would be 'No', or—perhaps more optimistically—'Not quite'.

Although we have tried our best to give all models a fair chance and set up the experiments in a clean and replicable fashion, some shortcomings remain that should be addressed in future work. First of all, we have tested only a handful of multimodal models on just three cognitive domains. While we believe that the used models and tasks provide good insights into the state-of-the-science of LLMs' cognitive abilities, future studies should look at more domains and different models to further tease apart when and why LLMs can mimic human reasoning. For example, it would be interesting to see whether scale is the only important feature influencing model performance[105,106]. Currently, our evidence suggests that even smaller models, for example, Fuyu, with its 8 billion parameters, can sometimes perform as well as GPT-4V in some tasks. Additionally, we applied all models out of the box and without further fine-tuning. Future studies could attempt to fine-tune multimodal LLMs to better align with cognitive data[107] and assess whether this improves their reasoning abilities more generally. Similar to other recent work[108], we found that many models were already constrained in their basic visual processing. While the more powerful closed-source models performed more robustly on simple scene understanding tasks, we found that they still failed simple questions that would be trivial for human observers. Thus, we think that the models' weak performance in some domains can partially be explained by their poor basic visual processing capabilities.

Another shortcoming of the current work is the simplicity of the used stimuli. While the block towers used in our first study were deliberately designed to be more realistic[98] than commonly used psychological stimuli[33], this was not true for the experiments in the other two domains. For the intuitive psychology experiments, in particular, we would expect the models to perform better if the stimuli contained more realistic images of people, which has been shown to work better in previous studies[109]. Interestingly, using more realistic stimuli can also change people's causal judgements[110]; how realistic the stimuli used in cognitive experiments should be remains an open question[111].

On a related point, we used only static images in our current experiments, which severely limits the breadth and level of detail of the questions we could ask. For example, some of the most canonical tasks investigating people's causal reasoning abilities involve videos of colliding billiard balls[52]. As future LLMs will probably be able to answer questions about videos[112], these tasks represent the next frontier of cognitively inspired benchmarks.

For the comparisons with human data, we used the participant data collected in the original studies for all experiments, except for the intuitive physics task, and assessed the correspondence between models and these data via a Bayesian mixed effects regression and R^2 values. Future work could expand on this approach by collecting new data from human participants choosing which of the model's judgements they prefer. This could lead to a more detailed comparison, similar to what has been proposed to discriminate among deep learning models for human vision[113] and language[114].

A crucial weakness of most studies using LLMs is that they can be sensitive to specific prompts[115–117]. While we have attempted to use prompts that elicited good behaviour, thereby giving LLMs a chance to perform well, future work could try to further optimize these prompts using available methods[118–120], while also assessing how the models respond to paraphrased versions of the same tasks. We present an exploratory analysis of the effects of response constraints and context complexity on human behaviour in the intuitive psychology astronaut task in Supplementary Fig. 7. While response constraints and context complexity both influence model outputs, we also find that small variations to prompts on a character level can impact model behaviour, probably due to tokenization. Taken together, this shows that evaluations of visual LLMs are not only dependent on the specific models and experiments used, but also on the prompts and probably even how these prompts are tokenized. While it could be possible to further engineer the used prompts, we believe that our current approach was sufficient to showcase these models' abilities.

Our work has shown that multimodal LLMs have come a long way, showing some correspondence to human behaviour and often performing above chance. Moreover, machine learning researchers have put forward various ideas about how to close the remaining gap between humans and machines[121], including self-supervised learning[122], translating from natural into probabilistic languages[123] or grounding LLMs in realistic environments[124]. This continuous evolution in models' capabilities necessitates a re-evaluation of the metaphors and tools we use to understand them. We believe that cognitive science can offer tools, theories and benchmarks to evaluate how close we have come to 'building machines that learn and think like people'.

Methods

Code

The open-source models were installed per the instructions on their related GitHub or Huggingface repositories and evaluated on a Slurm-based cluster with a single A100. For the results reported as GPT-4V, we used the public ChatGPT interface and the OpenAI application programming interface (API), specifically the November 2023 release of gpt4-vision-preview model which is available via the completions endpoint. For Claude-3, we used the Anthropic API. Code for replicating our results is available on GitHub (github.com/lsbuschoff/multimodal). All models were evaluated in Python using PyTorch[125]. Additional analyses were carried out using NumPy[126], Pandas[127] and SciPy[128]. Matplotlib[129] and Seaborn[130] were used for plotting. Bayesian mixed effects models were computed using brms[131] in R[132].

Models

Open-source. Fuyu is an 8 billion parameter multimodal text and image decoder-only transformer. We used the Huggingface implementation with standard settings and without further fine-tuning (available at https://huggingface.co/adept/fuyu-8b). The maximum number of generated tokens was set to 8 and responses were parsed by hand. Otter is a multimodal LLM that supports in-context instruction tuning and is based on the OpenFlamingo model. We used the Huggingface implementation of OTTER-Image-MPT7B (available at https://huggingface.co/luodian/OTTER-Image-MPT7B), again with standard settings and without fine-tuning. The maximum number of generated tokens was left at 512 and responses were parsed by hand. For LLaMA-Adapter V2, which adds adapters into LLaMA's transformer to turn it into an instruction-following model, we used the GitHub implementation of llama-adapter-v2-multimodal7b with standard settings and again without further fine-tuning (available at https://github.com/OpenGVLab/LLaMA-Adapter/tree/main/llama_adapter_v2_multimodal7b). The maximum number of generated tokens was left at 512 and responses were parsed by hand.

Closed-source. We initially queried GPT-4V through the ChatGPT interface, since the OpenAI API was not publicly available at the outset of this project. The intuitive psychology task responses were collected using the gpt4-vision-preview model variant after its November 2023 release in the API. We set the maximum number of generated tokens for a given prompt to 1 to get single numerical responses. All other parameters were set to their default values. Note that this model does not currently feature an option for manually setting the temperature, and the provided documentation does not specify what the default temperature is. We query Claude-3 using the Anthropic API. We use the model version claude-3-opus-20240229 with a temperature of zero and the maximum number of new tokens between 3 and 6 depending on the task.

Datasets

Intuitive physics with block towers. We tested the intuitive physical understanding of the models using images from ref. 98. The photos depict a block tower consisting of coloured wooden blocks in front of a white fabric (see Supplementary Fig. 1 for an example). The images are of size 224 × 244. In the dataset, there are a total of 516 images of block towers. We tested the models on 100 randomly drawn images. We first tested the models on their high-level visual understanding of the scenes: we tasked them with determining the background colour and the number of blocks in the image. To test their physical understanding, we tested them on the same task as the original study: we asked them to give a binary rating on the stability of the depicted block towers. For the first two tasks, we calculated the percentage of correct answers for each of the models. For the third task, we calculated a Bayesian linear mixed effects regression between human and model answers.

Due to the limited sample size of the original human experiment, we reran the human experiment from ref. 98 on Prolific with 107 participants (55 female and 52 male native English speakers with a mean age of 27.73 (s.d. = 4.21)). All participants agreed to take part in the study and were informed about the general purpose of the experiment. Experiments were performed in accordance with the relevant guidelines and regulations approved by the ethics committee of the University of Tübingen. Participants first saw an example trial, followed by 100 test images. In a two-alternative forced choice paradigm, participants were asked whether the block tower in a given image was stable or not stable. They were paid £1.50 and the median time they took to complete the experiment was 08:08 min, making the average base reward £11.07 per hour. Additionally, they received a bonus payment of up to £1 depending on their performance (1 penny for each correct answer).

Causal reasoning with Jenga. For the first causal reasoning experiment, we used images from ref. 100. The images show artificial block stacks of red and grey blocks on a black table (see Supplementary Fig. 2 for an example). The dataset consists of 42 images on which we tested all models. We again first tested the models on their high-level visual understanding of the scene and therefore tasked them with determining the number of blocks in the scene. The ground truth number of blocks in the scenes ranged from 6 to 19. Since this task is rather challenging due to the increased number of blocks, we do not report the percentage correct as for the intuitive physics dataset, but the mean over the absolute distance between model predictions and the ground truth for each image (Fig. 3a).

To test the causal reasoning of the models, we adopted the tasks performed in the original study[100,101]. We asked models to infer how many red blocks would fall if the grey block was removed. For this condition, Zhou and colleagues[100] collected data from 42 participants. We again report the absolute distance between model predictions and the ground truth for each image (Fig. 3b). We calculate a random baseline which uses the mean between 0 and the number of blocks for each specific image as the prediction. We also ask the models for a rating between 0 and 100 for how responsible the grey block is for the stability of the tower. Here, data for 41 human participants were publicly available. For both the number of blocks that would fall if the grey block was removed, and its responsibility for the stability of the tower, we calculate the mean Pearson correlation to human participants from the original study (Fig. 3c).

Causal reasoning with Michotte. For the second test for causal reasoning abilities, we used a task from ref. 52. It features 18 images which show a 2D view of two balls and their trajectories on a flat surface (see Supplementary Fig. 3 for an example). This experiment is a variation of the classic Michotte launching paradigm[102], used to test visual causal perception. We again first tested the models on their high-level visual understanding of the scene: we first asked them to determine the background colour (Fig. 4a) and then the direction of ball movement (Fig. 4b) from the two options 'left to right' or 'right to left' (the balls always moved from right to left).

To test the causal reasoning of the models, we adopted the tasks performed in the original study. We asked models about the actual outcome of the scene: 'Did ball A enter the gate?' As in the original experiments, models had to indicate their agreement with this statement on a scale from 0 (not at all) to 100 (completely). We then also asked the counterfactual question: 'Would ball A have entered the gate had it not collided with ball B?' The original authors[52] collected the responses of 14 participants in the 'outcome' condition and 13 participants in the 'counterfactual' condition. We here report the regression between model and human responses (Fig. 4c,d).

Intuitive psychology with astronaut images task. To test the intuitive psychology of the different LLMs, we used stimuli from ref. 103. This part consisted of three different experiments, each consisting of 16, 17 and 14 images showing a 2D depiction of an astronaut and care packages in different terrains (see Supplementary Figs. 4 and 5 for an example). To check their high-level understanding of the images, we again asked the models to determine the background colour of the images. Since this background colour is not uniform, we counted both 'Pink' and 'Purple' as correct answers. We report the percentage of correct answers for the background colour in Fig. 5a.

In accordance with the original study, analyses for the intuitive psychological capabilities of the models are split into cost questions (passing through a terrain is associated with a cost for the agent) and reward questions (collecting a care package yields some sort of reward for the agent). We pooled cost and reward questions over all three experiments and reported the mean Pearson correlation with the data of 90 human participants collected in ref. 103 (Fig. 5b,c). This heuristic calculates the costs and rewards associated with the environment from the amount of time an agent spends in each terrain and which care package the agent collects.

Intuitive psychology with the help or hinder task. The second intuitive psychology experiment is taken from ref. 104. It consists of 24 images showing a 2D depiction of two agents in a grid world (see Supplementary Fig. 6 for an example). To check the models' basic understanding of the images, we again asked the models to determine the background colour of the images and the number of boxes in the scene. We report the percentage of correct answers for both tasks in Fig. 6a,b.

We then asked the models whether the blue agent tried to help or hinder the red agent on a scale from 'definitely hinder RED' (0) to 'definitely help RED' (100), with the midpoint 'unsure' (50). We show the regression to human judgements in Fig. 6c. Finally, we asked the model the counterfactual question if the red agent would have succeeded in reaching the star had the blue agent not been there on a scale from 'not at all' (0) to 'very much' (100)? The original authors collected the responses of 50 participants for each of the two conditions ('intention' and 'counterfactual'). We show the mixed linear regression coefficients between model and human answers for all models with 95% credible intervals in Fig. 6d.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data used in our experiments are available on GitHub (github.com/lsbuschoff/multimodal). We have used subsets of openly available datasets from Lerer et al. (https://github.com/facebookarchive/UETorch/issues/25#issuecomment-235688223)[98], Gerstenberg et al. (https://github.com/tobiasgerstenberg/eye_tracking_causality)[52], Zhou et al. (https://github.com/cicl-stanford/mental_jenga)[100], Wu et al. (https://github.com/cicl-stanford/counterfactual_agents)[104] and Jara-Ettinger et al. (https://osf.io/uzs8r/)[103].

Code availability

All code needed to reproduce our results is available on GitHub (github.com/lsbuschoff/multimodal; and via the Zenodo repository at https://doi.org/10.5281/zenodo.14050104 (ref. 133)). We use openly available implementations of all LLMs except for GPT-4V and Claude-3. The code includes instructions on how to install and evaluate these LLMs. All prompts are listed in the Supplementary Information.

References

1. Mitchell, M. *Artificial Intelligence: A Guide for Thinking Humans* (Penguin, 2019).
2. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (eds Burstein, J. et al.) 4171–4186 (Association for Computational Linguistics, 2019).
3. Vaswani, A. et al. Attention is all you need. In *Proc. Advances in Neural Information Processing Systems 30 (NIPS 2017)* (eds Guyon, I. et al.) 5998–6008 (2017).
4. Brown, T. et al. Language models are few-shot learners. In *Proc. Advances in Neural Information Processing Systems 33 (NeurIPS 2020)* (eds Larochelle, H. et al.) 1877–1901 (Curran Associates, 2020).
5. Bubeck, S. et al. Sparks of artificial general intelligence: early experiments with GPT-4. Preprint at https://arxiv.org/abs/2303.12712 (2023).
6. Wei, J. et al. Emergent abilities of large language models. *Trans. Mach. Learn. Res*. https://openreview.net/forum?id=yzkSU5zdwD (2022).
7. Katz, D. M., Bommarito, M. J., Gao, S. & Arredondo, P. GPT-4 passes the bar exam. *Phil. Trans. R. Soc. A* **382**, 2270 (2024).
8. Sawicki, P. et al. On the power of special-purpose gpt models to create and evaluate new poetry in old styles. In *Proc. 14th International Conference on Computational Creativity (ICCC'23)* (eds Pease, A. et al.) 10–19 (Association for Computational Creativity, 2023).
9. Borsos, Z. et al. Audiolm: a language modeling approach to audio generation. *IEEE/ACM Trans. Audio Speech Lang. Process* **31**, 2523–2533 (2023).
10. Poldrack, R. A., Lu, T. & Beguš, G. Ai-assisted coding: experiments with GPT-4. Preprint at https://arxiv.org/abs/2304.13187 (2023).
11. Kasneci, E. et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* **103**, 102274 (2023).
12. Bommasani, R. et al. On the opportunities and risks of foundation models. Preprint at https://arxiv.org/abs/2108.07258 (2021).
13. Elkins, K. & Chun, J. Can GPT-3 pass a writer's Turing test? *J. Cult. Anal.* **5**, 2 (2020).
14. Dell'Acqua, F. et al. *Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality*. Harvard Business School Technology & Operations Mgt. Unit Working Paper (Harvard Business School, 2023).
15. Bašić, Ž., Banovac, A., Kružić, I. & Jerković, I. Better by you, better than me? ChatGPT-3.5 as writing assistance in students' essays. *Humanit. Soc. Sci. Commun.* **10**, 750 (2023).
16. Akata, E. et al. Playing repeated games with large language models. Preprint at https://arxiv.org/abs/2305.16867 (2023).
17. Simon, H. A. Cognitive science: the newest science of the artificial. *Cogn. Sci.* **4**, 33–46 (1980).
18. Rumelhart, D. E. et al. *Parallel Distributed Processing*, Vol.1 (MIT Press, 1987).
19. Wichmann, F. A. & Geirhos, R. Are deep neural networks adequate behavioral models of human visual perception? *Annu. Rev. Vis. Sci.* **9**, 501–524 (2023).
20. Bowers, J. S. et al. On the importance of severely testing deep learning models of cognition. *Cogn. Syst. Res.* **82**, 101158 (2023).
21. Marcus, G. Deep learning: a critical appraisal. Preprint at https://arxiv.org/abs/1801.00631 (2018).
22. Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. Building machines that learn and think like people. *Behav. Brain Sci.* **40**, e253 (2017).

23. Sejnowski, T. J. *The Deep Learning Revolution* (MIT, 2018).
24. Smith, K. A., Battaglia, P. W. & Vul, E. Different physical intuitions exist between tasks, not domains. *Comput. Brain Behav.* **1**, 101–118 (2018).
25. Bates, C. J., Yildirim, I., Tenenbaum, J. B. & Battaglia, P. Modeling human intuitions about liquid flow with particle-based simulation. *PLoS Comput. Biol.* **15**, e1007210 (2019).
26. Battaglia, P. et al. Computational models of intuitive physics. *Proc. Annu. Meet. Cogn. Sci. Soc.* **34**, 32–33 (2012).
27. Bramley, N. R., Gerstenberg, T., Tenenbaum, J. B. & Gureckis, T. M. Intuitive experimentation in the physical world. *Cogn. Psychol.* **105**, 9–38 (2018).
28. Ullman, T. D. & Tenenbaum, J. B. Bayesian models of conceptual development: learning as building models of the world. *Annu. Rev. Dev. Psychol.* **2**, 533–558 (2020).
29. Ullman, T. D., Spelke, E., Battaglia, P. & Tenenbaum, J. B. Mind games: game engines as an architecture for intuitive physics. *Trends Cogn. Sci.* **21**, 649–665 (2017).
30. Hamrick, J., Battaglia, P. & Tenenbaum, J. B. Probabilistic internal physics models guide judgments about object dynamics. *Proc. Annu. Meet. Cogn. Sci. Soc.* **33**, 1545–1550 (2011).
31. Mildenhall, P. & Williams, J. Instability in students' use of intuitive and Newtonian models to predict motion: the critical effect of the parameters involved. *Int. J. Sci. Educ.* **23**, 643–660 (2001).
32. Todd, J. T. & Warren, W. H. Jr Visual perception of relative mass in dynamic events. *Perception* **11**, 325–335 (1982).
33. Battaglia, P. W., Hamrick, J. B. & Tenenbaum, J. B. Simulation as an engine of physical scene understanding. *Proc. Natl Acad. Sci. USA* **110**, 18327–18332 (2013).
34. Hamrick, J. B., Battaglia, P. W., Griffiths, T. L. & Tenenbaum, J. B. Inferring mass in complex scenes by mental simulation. *Cognition* **157**, 61–76 (2016).
35. Bakhtin, A., van der Maaten, L., Johnson, J., Gustafson, L. & Girshick, R. PHYRE: a new benchmark for physical reasoning. In *Proc. Advances in Neural Information Processing Systems 32 (NeurIPS 2019)* (eds Wallach, H. et al.) 5082–5093 (Curran Associates, 2019).
36. Riochet, R. et al. Intphys: a framework and benchmark for visual intuitive physics reasoning. Preprint at https://arxiv.org/abs/1803.07616 (2018).
37. Schulze Buschoff, L. M., Schulz, E. & Binz, M. The acquisition of physical knowledge in generative neural networks. In *Proc. 40th International Conference on Machine Learning* (eds Krause, A. et al.) 202, 30321–30341 (JMLR, 2023).
38. Waldmann, M. *The Oxford Handbook of Causal Reasoning* (Oxford Univ. Press, 2017).
39. Cheng, P. W. From covariation to causation: a causal power theory. *Psychol. Rev.* **104**, 367 (1997).
40. Holyoak, K. J. & Cheng, P. W. Causal learning and inference as a rational process: the new synthesis. *Annu. Rev. Psychol.* **62**, 135–163 (2011).
41. Pearl, J. *Causality* (Cambridge Univ. Press, 2009).
42. Griffiths, T. L. & Tenenbaum, J. B. Theory-based causal induction. *Psychol. Rev.* **116**, 661 (2009).
43. Lagnado, D. A., Waldmann, M. R., Hagmayer, Y. & Sloman, S. A. in *Causal Learning: Psychology, Philosophy, and Computation* (eds Gopnik, A. and Schulz, L.) 154–172 (Oxford Univ. Press, 2007).
44. Carey, S. *On the Origin of Causal Understanding* (Clarendon Press/Oxford Univ. Press, 1995).
45. Gopnik, A. et al. A theory of causal learning in children: causal maps and Bayes nets. *Psychol. Rev.* **111**, 3 (2004).
46. Lucas, C. G. & Griffiths, T. L. Learning the form of causal relationships using hierarchical bayesian models. *Cogn. Sci.* **34**, 113–147 (2010).
47. Bramley, N. R., Gerstenberg, T., Mayrhofer, R. & Lagnado, D. A. Time in causal structure learning. *J. Exp. Psychol. Learn. Mem. Cogn.* **44**, 1880 (2018).
48. Griffiths, T. L. & Tenenbaum, J. B. Structure and strength in causal induction. *Cogn. Psychol.* **51**, 334–384 (2005).
49. Schulz, L., Kushnir, T. & Gopnik, A. in *Causal Learning: Psychology, Philosophy, and Computation* (eds Gopnik, A. and Schulz, L.) 67–85 (Oxford Univ. Press, 2007).
50. Bramley, N. R., Dayan, P., Griffiths, T. L. & Lagnado, D. A. Formalizing Neurath's ship: approximate algorithms for online causal learning. *Psychol. Rev.* **124**, 301 (2017).
51. Gerstenberg, T. What would have happened? Counterfactuals, hypotheticals and causal judgements. *Philos. Trans. R. Soc. B* **377**, 20210339 (2022).
52. Gerstenberg, T., Peterson, M. F., Goodman, N. D., Lagnado, D. A. & Tenenbaum, J. B. Eye-tracking causality. *Psychol. Sci.* **28**, 1731–1744 (2017).
53. Gerstenberg, T., Goodman, N. D., Lagnado, D. A. & Tenenbaum, J. B. A counterfactual simulation model of causal judgments for physical events. *Psychol. Rev.* **128**, 936 (2021).
54. Jin, Z. et al. CLadder: Assessing causal reasoning in language models. In *Proc. Advances in Neural Information Processing Systems 36 (NeurIPS 2023)* (eds Oh, A. et al.) 31038–31065 (Curran Associates, 2023).
55. Dasgupta, I. et al. Causal reasoning from meta-reinforcement learning. Preprint at https://arxiv.org/abs/1901.08162 (2019).
56. Baker, C. L. & Tenenbaum, J. B. in *Plan, Activity, and Intent Recognition: Theory and Practice* (eds Sukthankar, G. et al.) 177–204 (Morgan Kaufmann, 2014).
57. Jern, A. & Kemp, C. A decision network account of reasoning about other people's choices. *Cognition* **142**, 12–38 (2015).
58. Vélez, N. & Gweon, H. Learning from other minds: an optimistic critique of reinforcement learning models of social learning. *Curr. Opin. Behav. Sci.* **38**, 110–115 (2021).
59. Spelke, E. S., Bernier, E. P. & Skerry, A. *Core Social Cognition* (Oxford Univ. Press, 2013).
60. Baker, C., Saxe, R. & Tenenbaum, J. Bayesian theory of mind: modeling joint belief-desire attribution. *Proc. Annu. Meet. Cogn. Sci. Soc.* **33**, 2469–2474 (2011).
61. Frith, C. & Frith, U. Theory of mind. *Curr. Biol.* **15**, R644–R645 (2005).
62. Saxe, R. & Houlihan, S. D. Formalizing emotion concepts within a Bayesian model of theory of mind. *Curr. Opin. Psychol.* **17**, 15–21 (2017).
63. Baker, C. L. et al. Intuitive theories of mind: a rational approach to false belief. *Proc. Annu. Meet. Cogn. Sci. Soc.* **28**, 69k8c7v6 (2006).
64. Shum, M., Kleiman-Weiner, M., Littman, M. L. & Tenenbaum, J. B. Theory of minds: understanding behavior in groups through inverse planning. In *Proc. 33rd AAAI Conference on Artificial Intelligence* 6163–6170 (Curran Associates, 2019).
65. Baker, C. L., Jara-Ettinger, J., Saxe, R. & Tenenbaum, J. B. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nat. Hum. Behav.* **1**, 0064 (2017).
66. Zhi-Xuan, T. et al. Solving the baby intuitions benchmark with a hierarchically Bayesian theory of mind. Preprint at https://arxiv.org/abs/2208.02914 (2022).
67. Rabinowitz, N. et al. Machine theory of mind. In *Proc. 35th International Conference on Machine Learning* (eds Dy, J. & Krause, A.) 80, 4218–4227 (JMLR, 2018).
68. Kosinski, M. Theory of mind may have spontaneously emerged in large language models. Preprint at https://arxiv.org/abs/2302.02083 (2023).
69. Ullman, T. Large language models fail on trivial alterations to theory-of-mind tasks. Preprint at https://arxiv.org/abs/2302.08399 (2023).

70. Schulz, L. The origins of inquiry: inductive inference and exploration in early childhood. *Trends Cogn. Sci.* **16**, 382–389 (2012).
71. Ullman, T. D. *On the Nature and Origin of Intuitive Theories: Learning, Physics and Psychology*. PhD thesis, Massachusetts Institute of Technology (2015).
72. Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. How to grow a mind: statistics, structure, and abstraction. *Science* **331**, 1279–1285 (2011).
73. Binz, M. & Schulz, E. Using cognitive psychology to understand GPT-3. *Proc. Natl Acad. Sci. USA* **120**, e2218523120 (2023).
74. Huang, J. & Chang, K. C.-C. *Towards reasoning in large language models: a survey. In Findings of the Association for Computational Linguistics: ACL 2023* (eds Rogers, A. et al.) 1049–1065 (Association for Computational Linguistics, 2023).
75. Sawada, T. et al. ARB: Advanced reasoning benchmark for large language models. Preprint at https://arxiv.org/abs/2307.13692 (2023).
76. Wei, J. et al. Chain-of-thought prompting elicits reasoning in large language models. In *Proc. Advances in Neural Information Processing Systems 35 (NeurIPS 2022)* (eds Koyejo, S. et al.) 24824–24837 (Curran Associates, 2022).
77. Webb, T., Holyoak, K. J. & Lu, H. Emergent analogical reasoning in large language models. *Nat. Hum. Behav.* **7**, 1526–1541 (2023).
78. Coda-Forno, J. et al. Inducing anxiety in large language models increases exploration and bias. Preprint at https://arxiv.org/abs/2304.11111 (2023).
79. Eisape, T. et al. A systematic comparison of syllogistic reasoning in humans and language models. In *Proc. 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (eds Duh, K. et al.) 8425–8444 (Association for Computational Linguistics, 2024).
80. Hagendorff, T., Fabi, S. & Kosinski, M. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nat. Comput. Sci.* **3**, 833–838 (2023).
81. Ettinger, A. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Trans. Assoc. Comput. Linguist.* **8**, 34–48 (2020).
82. Jones, C. R. et al. Distrubutional semantics still can’t account for affordances. *Proc. Annu. Meet. Cogn. Sci. Soc.* **44**, 482–489 (2022).
83. Rahwan, I. et al. Machine behaviour. *Nature* **568**, 477–486 (2019).
84. Schulz, E. & Dayan, P. Computational psychiatry for computers. *iScience* **23**, 12 (2020).
85. Rich, A. S. & Gureckis, T. M. Lessons for artificial intelligence from the study of natural stupidity. *Nat. Mach. Intell.* **1**, 174–180 (2019).
86. Zhang, Y., Pan, J., Zhou, Y., Pan, R. & Chai, J. Grounding visual illusions in language: do vision-language models perceive illusions like humans? In *Proc. 2023 Conference on Empirical Methods in Natural Language* (eds Bouamor, H. et al.) 5718–5728 (Association for Computational Linguistics, 2023).
87. Mitchell, M., Palmarini, A. B. & Moskvichev, A. Comparing humans, GPT-4, and GPT-4v on abstraction and reasoning tasks. Preprint at https://arxiv.org/abs/2311.09247 (2023).
88. Zečević, M., Willig, M., Dhami, D. S. & Kersting, K. Causal parrots: large language models may talk causality but are not causal. *Trans. Mach. Learn. Res.* https://openreview.net/pdf?id=tv46tCzs83 (2023).
89. Zhang, C., Wong, L., Grand, G. & Tenenbaum, J. Grounded physical language understanding with probabilistic programs and simulated worlds. *Proc. Annu. Meet. Cogn. Sci. Soc.* **45**, 3476–3483 (2023).
90. Jassim, S. et al. GRASP: a novel benchmark for evaluating language grounding and situated physics understanding in multimodal language models. Preprint at https://arxiv.org/abs/2311.09048 (2023).
91. Kosoy, E. et al. Towards understanding how machines can learn causal overhypotheses. *Proc. Annu. Meet. Cogn. Sci. Soc.* **45**, 363–374 (2023).
92. Gandhi, K., Fränken, J.-P., Gerstenberg, T. & Goodman, N. D. Understanding social reasoning in language models with language models. In *Proc. 37th International Conference on Neural Information Processing Systems* (NIPS *’23)* (eds Oh, A. et al.) 13518–13529 (Curran Associates, 2024).
93. Srivastava, A. et al. Beyond the imitation game: quantifying and extrapolating the capabilities of language models. Preprint at https://arxiv.org/abs/2206.04615 (2022).
94. Baltrušaitis, T., Ahuja, C. & Morency, L.-P. Multimodal machine learning: a survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**, 423–443 (2018).
95. Reed, S. et al. Generative adversarial text to image synthesis. In *Proc. 33rd International Conference on Machine Learning* (eds Balcan, M. F. & Weinberger, K. Q.) 48, 1060–1069 (JMLR, 2016).
96. Wu, Q. et al. Visual question answering: a survey of methods and datasets. *Comput. Vis. Image Underst.* **163**, 21–40 (2017).
97. Manmadhan, S. & Kovoor, B. C. Visual question answering: a state-of-the-art review. *Artif. Intell. Rev.* **53**, 5705–5745 (2020).
98. Lerer, A., Gross, S. & Fergus, R. Learning physical intuition of block towers by example. In *Proc. 33rd International Conference on Machine Learning* (eds Balcan, M. F. & Weinberger, K. Q.) 48, 430–438 (JMLR, 2016).
99. Gelman, A., Goodrich, B., Gabry, J. & Vehtari, A. R-squared for Bayesian regression models. *Am. Stat.* **73**, 307–309 (2019).
100. Zhou, L., Smith, K. A., Tenenbaum, J. B. & Gerstenberg, T. Mental Jenga: a counterfactual simulation model of causal judgments about physical support. *J. Exp. Psychol. Gen.* **152**, 2237 (2023).
101. Gerstenberg, T., Zhou, L., Smith, K. A. & Tenenbaum, J. B. Faulty towers: A hypothetical simulation model of physical support. *Proc. Annu. Meet. Cogn. Sci. Soc.* **39**, 409–414 (2017).
102. Michotte, A. *The Perception of Causality* (Basic Books, 1963).
103. Jara-Ettinger, J., Schulz, L. E. & Tenenbaum, J. B. The naïve utility calculus as a unified, quantitative framework for action understanding. *Cogn. Psychol.* **123**, 101334 (2020).
104. Wu, S. A., Sridhar, S. & Gerstenberg, T. A computational model of responsibility judgments from counterfactual simulations and intention inferences. *Proc. Annu. Meet. Cogn. Sci. Soc.* **45**, 3375–3382 (2023).
105. Sutton, R. The bitter lesson. *Incomplete Ideas* http://www.incompleteideas.net/IncIdeas/BitterLesson.html (2019).
106. Kaplan, J. et al. Scaling laws for neural language models. Preprint at https://arxiv.org/abs/2001.08361 (2020).
107. Binz, M. & Schulz, E. Turning large language models into cognitive models. In *Proc. 12th International Conference on Learning Representations (ICLR)* https://openreview.net/forum?id=eiC4BKypf1 (OpenReview, 2024).
108. Rahmanzadehgervi, P., Bolton, L., Taesiri, M. R. & Nguyen, A. T. Vision language models are blind. In *Proc. Asian Conference on Computer Vision (ACCV)* 18–34 (Computer Vision Foundation, 2024).
109. Ju, C., Han, T., Zheng, K., Zhang, Y. & Xie, W. Prompting visual-language models for efficient video understanding. In *Proc. Computer Vision – ECCV 2022: 17th European Conference* (eds Avidan, S. et al.) 105–124 (Springer, 2022).
110. Meding, K., Bruijns, S. A., Schölkopf, B., Berens, P. & Wichmann, F. A. Phenomenal causality and sensory realism. *Iperception* **11**, 2041669520927038 (2020).
111. Allen, K. R. et al. Using games to understand the mind. *Nat. Hum. Behav.* **8**, 1035–1043 (2024).

112. Maaz, M., Rasheed, H., Khan, S. & Khan, F. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *Proc. 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1:* Long Papers*)* (eds Ku, L.-W. et al.) 12585–12602 (Association for Computational Linguistics, 2024).
113. Golan, T., Raju, P. C. & Kriegeskorte, N. Controversial stimuli: pitting neural networks against each other as models of human cognition. *Proc. Natl Acad. Sci. USA* **117**, 29330–29337 (2020).
114. Golan, T., Siegelman, M., Kriegeskorte, N. & Baldassano, C. Testing the limits of natural language models for predicting human language judgements. *Nat. Mach. Intell.* **5**, 952–964 (2023).
115. Reynolds, L. & McDonell, K. Prompt programming for large language models: beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (eds Kitamura, Y. et al.) 314 (Association for Computing Machinery, 2021).
116. Strobelt, H. et al. Interactive and visual prompt engineering for ad-hoc task adaptation with large language models. *IEEE Trans. Vis. Comput. Graph.* **29**, 1146–1156 (2022).
117. Webson, A. & Pavlick, E. Do prompt-based models really understand the meaning of their prompts? In *Proc. 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (eds Carpuat, M. et al.) 2300–2344 (Association for Computational Linguistics, 2022).
118. Liu, P. et al. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* **55**, 195 (2023).
119. Gu, J. et al. A systematic survey of prompt engineering on vision-language foundation models. Preprint at https://arxiv.org/abs/2307.12980 (2023).
120. Coda-Forno, J. et al. Meta-in-context learning in large language models. In *Proc. Advances in Neural Information Processing Systems 36 (NeurIPS 2023)* (eds Oh, A. et al.) 65189–65201 (Curran Associates, 2023).
121. Geirhos, R. et al. Partial success in closing the gap between human and machine vision. In *Proc. Advances in Neural Information Processing Systems 34 (NeurIPS 2021)* (eds Ranzato, M. et al.) 23885–23899 (Curran Associates, 2021).
122. Balestriero, R. et al. A cookbook of self-supervised learning. Preprint at https://arxiv.org/abs/2304.12210 (2023).
123. Wong, L. et al. From word models to world models: translating from natural language to the probabilistic language of thought. Preprint at https://arxiv.org/abs/2306.12672 (2023).
124. Carta, T. et al. Grounding large language models in interactive environments with online reinforcement learning. In *Proc. 40th International Conference on Machine Learning* (eds Krause, A. et al.) 202, 3676–3713 (JMLR, 2023).
125. Paszke, A. et al. Pytorch: An imperative style, high-performance deep learning library. In *Proc. Advances in Neural Information Processing Systems 32 (NeurIPS 2019)* (eds Wallach, H. et al.) 8026–8037 (Curran Associates, 2019).
126. Harris, C. R. et al. Array programming with NumPy. *Nature* **585**, 357–362 (2020).
127. Pandas Development Team. pandas-dev/pandas: Pandas. *Zenodo* https://doi.org/10.5281/zenodo.3509134 (2020).
128. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
129. Hunter, J. D. Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
130. Waskom, M. L. seaborn: statistical data visualization. *J. Open Source Softw* **6**, 3021 (2021).
131. Bürkner, P.-C. brms: an R package for Bayesian multilevel models using Stan. *J. Stat. Softw*. https://doi.org/10.18637/jss.v080.i01 (2017).
132. R Core Team. R: a language and environment for statistical computing (R Foundation for Statistical Computing, 2021).
133. Schulze Buschoff, L. M. et al. lsbuschoff/multimodal: First release. *Zenodo* https://doi.org/10.5281/zenodo.14050104 (2024).

Acknowledgements

We thank M. Thalmann and C. Demircan for helpful discussions on the statistical analysis. This work was supported by the Max Planck Society (L.M.S.B. and E.S.), the Volkswagen Foundation (L.M.S.B. and E.S.), the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: grant no. 01IS18039A (E.A. and M.B.), and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC 2064/1—grant no. 390727645 (E.A. and M.B.).

Author contributions

L.M.S.B. and E.S. conceived the study. L.M.S.B. and E.A. conducted the VLM experiments. E.A. conducted the human experiment. L.M.S.B. analysed the results with input from E.S. L.M.S.B., E.A. and E.S. wrote the manuscript with input from M.B.

Funding

Open access funding provided by Helmholtz Zentrum München - Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-024-00963-y.

Correspondence and requests for materials should be addressed to Luca M. Schulze Buschoff.

Peer review information *Nature Machine Intelligence* thanks Taylor Webb and Michael Frank for their contribution to the peer review of this work.

nature portfolio

Corresponding author(s): Luca M. Schulze Buschoff

Last updated by author(s): Nov 7, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
☒	□	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

The open-source models (lama_adapter_v2_multimodal7b, Fuyu-8b, OTTER-Image-MPT7B) were installed per the instructions on their related Github or Huggingface repositories and evaluated on a Slurm-based cluster with a single A100. For the results reported as GPT-4V, we used the public ChatGPT interface and the OpenAI API, specifically the November 2023 release of gpt4-vision-preview model which is not available anymore via the completions endpoint. For CLAUDE-3 we used claude-3-opus-20240229 with the Anthropic API. Code for replicating our results is available on GitHub (github.com/lsbuschoff/multimodal).

Data analysis

All models were evaluated in Python using PyTorch (2.3.1 and py3.12_cuda12.1_cudnn8.9.2_0). Additional analyses were carried out using NumPy (1.26.4), Pandas (2.2.2), and SciPy (1.13.1). Matplotlib (3.8.4) and Seaborn (0.12.2) were used for plotting. Bayesian mixed effects models were computed using brms (2.21.0) in R.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All data used in our experiments is available on GitHub (github.com/lsbuschoff/multimodal). We have used subsets of openly available data sets from Lerer et al. (https://github.com/facebookarchive/UETorch/issues/25#issuecomment-235688223), Gerstenberg et al. (https://github.com/tobiasgerstenberg/eye_tracking_causality), Zhou et al. (https://github.com/ciclstanford/mental_jenga), Wu et al. (https://github.com/cicl-stanford/counterfactual_agents), and Jara-Ettinger et al. (https://osf.io/uzs8r/).

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	107 subjects (55 female and 52 male), no sex or gender based analyses were performed
Population characteristics	native English speakers with a mean age of 27.73 (sd = 4.21)
Recruitment	Participants were recruited from Prolific with the constraint of requiring native English speakers
Ethics oversight	Ethics commission at the faculty of medicine, Eberhard Karls University Tübingen

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

□ Life sciences ☒ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Quantitative study: evaluation of vision language models on cognitive domains and comparison to human data
Research sample	Five state of the art multimodal large language models. This sample of models were chosen so that they adequately represent current SOTA models (large and small, open source and closed source).
Sampling strategy	No sample size calculation was performed. Samples were determined by the available data sets in the respective domains.
Data collection	The open-source models were evaluated on a Slurm-based cluster with a single A100. For GPT-4V, we used the public ChatGPT interface and the OpenAI API, specifically the November 2023 release of gpt4-vision-preview model. For CLAUDE-3 we used claude-3-opus-20240229 with the Anthropic API. Data collection code is available on GitHub (github.com/lsbuschoff/multimodal). Additional human data was collected on Prolific for one evaluation. Human subjects first saw an example trial, followed by 100 test images. In a 2AFC paradigm, subjects were asked if the block tower in a given image was stable or not stable. They were paid £1.5 and the median time they took to complete the experiment was 08:08 minutes, making the average base reward £11.07 per hour. Additionally, they received a bonus payment of up to £1 depending on their performance (1 cent for each correct answer). The researcher was not blinded to experimental condition or the study hypothesis.
Timing	Data collection mainly took place in October and November of 2023 and again in May and June of 2024.
Data exclusions	The raw answers from the large language models were parsed and non-intelligible answers were registered as N/A. Human subjects were excluded if they did not complete the experiment.
Non-participation	For the human experiment, 11 participants started the experiment but did not complete it.

Randomization	All models received the same stimuli in the same order. Models were input images together with each question, since gradients were enabled and the models queried in a new context for each question, randomization is not required. Human participants received the same stimuli in a randomized order.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
☒	□ Clinical data
☒	□ Dual use research of concern

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging