EVALUATING MULTI-MODAL LANGUAGE MODELS THROUGH CONCEPT HACKING

Anonymous authors
Paper under double-blind review

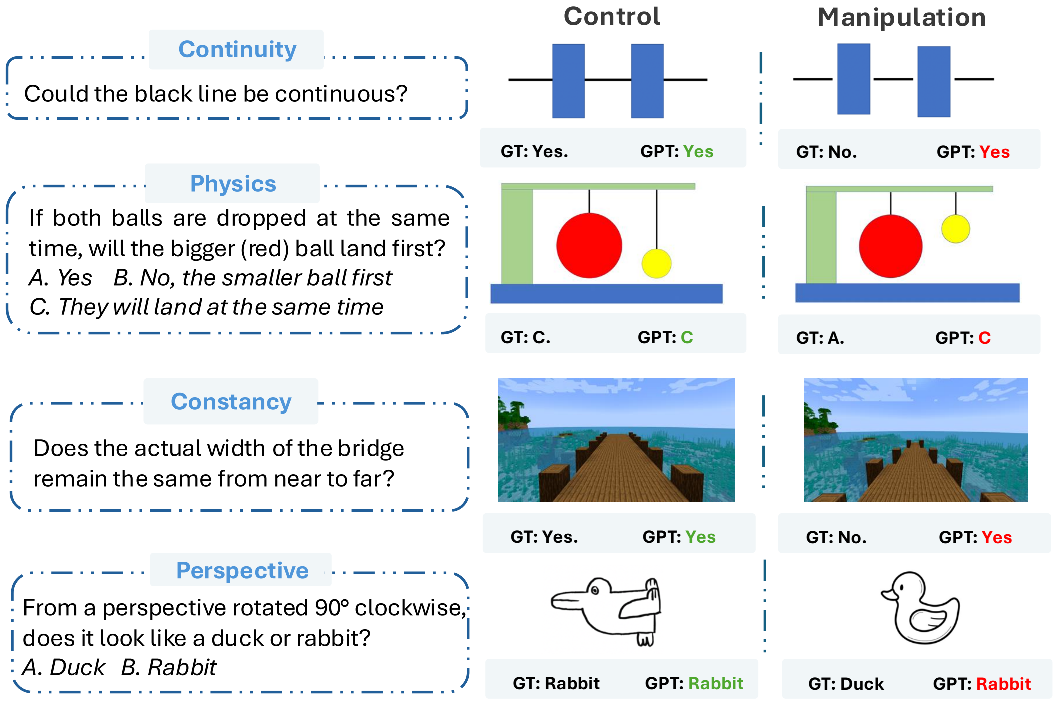
ABSTRACT

Evaluating the cognitive abilities of Multi-modal Language Models (MLLMs) is challenging due to their reliance on spurious correlations. To distinguish shortcut-taking from genuine reasoning, we introduce Concept Hacking, a paradigm manipulating concept-relevant information to flip the ground-truth but preserving concept-irrelevant confounds. For instance, in a perceptual constancy test, models must recognize that a uniformly wide bridge does not narrow in the distance; the manipulated condition using concept hacking altered the bridge to actually taper. We assessed 209 models across 45 experiment pairs spanning nine low-level cognitive abilities, encompassing all five core knowledge domains. Comparing performance on manipulated versus standard conditions revealed that models fell into shortcut-reliant or illusory understanding types, with none approaching human-level performance. Models of varying sizes appear in each category, indicating that scaling neither imparts core knowledge nor reduces shortcut reliance. These findings highlight fundamental limitations in current MLLMs, reinforcing concerns about their ability to achieve genuine understanding.

1 INTRODUCTION

Multi-modal Language Models (MLLMs) have achieved unprecedented success by leveraging vast web-scale training and modality alignment (Li et al., 2024; Fu et al., 2023; Wu & Xie, 2024; Xu et al., 2024; Shao et al., 2024; Brown et al., 2020; Achiam et al., 2023; Bai et al., 2023; Touvron et al., 2023; Jaech et al., 2024). Progressively, MLLMs have demonstrated competitive performance in complex tasks involving high-level perception and reasoning (Li et al., 2024; Liu et al., 2024; Team et al., 2023; Fu et al., 2023; OpenAI, 2023), such as spatial reasoning (Chen et al., 2024a; Cai et al., 2024), character recognition (Mori et al., 1999), scene understanding (Cordts et al., 2016; Chen et al., 2017), action recognition (Jhuang et al., 2013; Herath et al., 2017) and prediction (Lan et al., 2014; Kong & Fu, 2022), reaching near-human performance. However, recent studies have shown that even SOTA MLLMs face critical limitations as compared to human intelligence. To begin with, said excellency often does not appear to translate to more generalized and real-world contexts, with slight tweaks of the task conditions capable of causing collapses in performance (Shiffrin & Mitchell, 2023; Zhang et al., 2024), highlighting persistent robustness challenges. At the same time, they perform poorly on rudimentary reasoning tasks like counting Paiss et al. (2023) and compositional reasoning Yuksekgonul et al. (2022) despite their excellence at high-level reasoning tasks on similar domains (Paiss et al., 2023; Rahmanzadehgervi et al., 2024), demonstrating the long-standing Moravec's Paradox: tasks that are easy to humans could be extremely difficult to machines and vice versa (Moravec, 1988).

An influential proposal in cognitive science posits that children first acquire basic reasoning abilities about the physical world, which serve as a foundation for the development of more complex, abstract cognitive skills as they mature (Barsalou, 2008; Samuelson & Smith, 2000; Barsalou, 2010; Pezzulo et al., 2013). This grounding view of human cognitive development provides crucial insights into the limitations of MLLMs. Notably, because humans develop simple abilities before more complex ones, they do not exhibit Moravec's Paradox. Furthermore, since early, foundational abilities are causally linked to later, higher-order reasoning skills, the grounding perspective suggests that the absence of these simple abilities in early learning stages may hinder the development of complex cognitive functions. This mechanistic connection offers a potential explanation for why both Moravec's Paradox and robustness challenges are observed in MLLMs. If complex reasoning



Continuity

Could the black line be continuous?

Physics

If both balls are dropped at the same time, will the bigger (red) ball land first?
A. Yes B. No, the smaller ball first
C. They will land at the same time

Constancy

Does the actual width of the bridge remain the same from near to far?

Perspective

From a perspective rotated 90° clockwise, does it look like a duck or rabbit?
A. Duck B. Rabbit

Figure 1: Example Questions Using the Concept Hacking Manipulation

skills in MLLMs are not built upon a foundation of simpler domain-relevant abilities, their capacity to generalize across conditions may be fundamentally compromised. Evaluating and systematically implementing such foundational abilities in MLLMs is thus a critical step toward improving their robustness and reasoning capabilities.

A key challenge in assessing the cognitive abilities of language models is their tendency to exploit spurious correlations. That is, their apparent proficiency in certain tasks may stem from shortcut learning rather than genuine cognitive competence (Bender et al., 2021). Extensive research has demonstrated this reliance on superficial cues in benchmarks designed to assess high-level reasoning in MLLMs. To examine whether evaluations of lower-level cognitive abilities are similarly susceptible to such shortcuts, we introduce a control experiment designed to rigorously probe the core knowledge present in MLLMs. Central to this experiment is a novel technique termed *concept hacking*, which systematically tests whether the models genuinely understand fundamental cognitive constructs or merely leverage statistical artifacts.

2 METHODS

2.1 THE CONCEPT HACKING DESIGN

Concept hacking systematically manipulates task-relevant details in core knowledge assessments to completely invert the ground truth while preserving all task-irrelevant conditions. We illustrate four examples in Fig.1. The comparison between an individual's performance on a manipulation task and their corresponding standard control is capable of revealing three distinctive strategies for answering lower-level cognitive assessments: core knowledge understanding, shortcut-taking, and illusory understanding. Individuals that possess core knowledge of respective domains (like humans) would not be misled by the manipulation, as they will evaluate both scenarios based on a valid understanding of the world—acknowledging what information is task-relevant. In contrast, individuals that rely on statistical correlations from their training data, rather than true conceptual understanding, can be misled by the manipulations and fail the task. Finally, individuals with a strong disposition against core knowledge in specific domains would consistently fail the standard control and thereby answering the manipulation question correctly. In other words, they are "being right for the wrong reason" due to an illusory understanding of the core knowledge domain.

For example, as shown in the third case of Fig.1, a standard probe of perceptual constancy assesses whether a model understands that a bridge of uniform width extending into the ocean does not actually become narrower in the distance. In the manipulated condition, all task-irrelevant details—such as the viewing angle and environmental textures—are kept identical to the standard task, but the bridge itself is altered to genuinely taper as it extends outward. Models possessing the understanding of perceptual constancy would have no difficulty answering both the manipulation task and standard control correctly. On the contrary, a model relying on spurious correlations between the task and previous examples of similar scenarios in the data would succeed in the original task but fail the manipulated one. Finally, a model with a strong inclination toward the belief that objects extending into the horizon are actually getting thinner physically would fail the control task while correctly answering the manipulated version due to its misaligned knowledge about the world.

We applied the concept hacking method to 45 standard tasks, each designed to assess one of nine low-level cognitive abilities, with five tasks per ability. For each standard task, we created a manipulated counterpart, resulting in a total of 90 tasks (45 manipulated and 45 corresponding standard control tasks). By comparing model performance between manipulated and standard conditions, we systematically identify instances of shortcut-taking and illusory competence in core knowledge assessments.

2.2 ASSESSING LOW-LEVEL COGNITIVE ABILITIES

A large body of work in cognitive science has demonstrated that humans possess a foundational understanding of key domains of the world from a very young age, collectively referred to as core knowledge (Spelke, 2003; Spelke & Kinzler, 2007). This set of knowledge comprises fundamental principles about objects, actions, numbers, space, and social relations, including their interconnections. Core knowledge functions as children's "developmental start-up software," enabling them to navigate, interpret, and learn from the rich and dynamic environment of early life (Lake et al., 2017). To systematically investigate fundamental knowledge representations in MLLMs, we select nine low-level cognitive abilities that collectively span all five core knowledge domains. These abilities emerge at different stages of early cognitive development and serve as the building blocks for more complex reasoning processes. We design tasks to assess these abilities by adapting classic cognitive tasks from the developmental psychology literature, presenting them in a single-image question format suitable for MLLMs. Below, we provide detailed descriptions of each included low-level cognitive ability, along with an example type of classic cognitive task for assessing the respective ability, illustrating how the concept is tested in our framework.

Boundary Boundary refers to the cognitive understanding of where one object ends and another begins, an essential aspect of perceiving and understanding the physical world (Kestenbaum et al., 1987). Without understanding boundaries, it seems very hard to construct a concept of the object (Berkeley, 1709; Jackendoff, 1991).

Spatiality Spatiality, particularly demonstrated through the A-not-B task, involves a child's understanding of the location of objects in relation to their environment (Bell & Adams, 1999). In a classic A-not-B task, an object is hidden at location A (such as under a cup) and the child successfully finds it several times. Then, the object is visibly moved to a different location B (under a different cup), in full view of the child. Younger infants often make the error of searching for the object at the original location A, indicating a developmental stage where their understanding of object spatiality is still forming.

Perceptual Constancy Perceptual constancy is the cognitive ability to perceive objects as being constant in their properties, such as size, shape, and color, despite changes in perspective, distance, or lighting (Rutherford & Brainard, 2002; Khang & Zaidi, 2004; Green, 2023). For instance, consider a red ball being thrown in a park. To an observer, the ball appears smaller as it moves farther away, yet the observer understands it remains the same size throughout its trajectory.

Object Permanence Permanence, or specifically object permanence, is the idea that objects continue to exist even when they are not visible (Baillargeon, 1986; Spelke et al., 1992). Imagine a simple scene: a small child playing peek-a-boo. In the beginning, when the caregiver covers their face with their hands, the child might seem surprised or even distressed, thinking the person has disappeared. However, as children's understanding of permanence develops, they begin to realize that just because they can't see the person's face, it doesn't mean the person is gone.

Continuity Continuity is the cognitive prior in humans that in our world, objects usually exist in a consistent and continuous manner, even moving out of sight (Spelke et al., 1995; Le Poidevin, 2000; Spelke et al., 1994; Yantis, 1995; Yi et al., 2008; Bertenthal et al., 2013). Picture a train moving through a tunnel: as it enters one end, yet we naturally expect it to emerge from the other end, if the train is long enough. This expectation demonstrates our understanding of object continuity. Even though the train is not visible while it's inside the tunnel, we know it continues to exist.

Conservation Conservation refers to the ability to understand that certain properties of physical entities are conserved after an object undergoes physical transformation (Piaget & Inhelder, 1974). This is instantiated in their ability to tell that quantities of physical entities across different domains, such as number, length, solid quantity and liquid volume, will remain the same despite adjustments of their arrangement, positioning, shapes, and containers (Halford, 2011; Craig et al., 1973; Piaget & Inhelder, 1974; Houdé et al., 2011; Poirel et al., 2012; Marwaha et al., 2017; Viarouge et al., 2019). For example, when a child watches water being poured from a tall, narrow glass into a short, wide one, a grasp of liquid conservation would lead them to understand that the amount of water remains the same even though its appearance has changed.

Perspective-taking Perspective-taking is the ability to view things from another's perspective. This ability has seminal importance both to the understanding of the physical world as well as to the competence in social interactions (Wimmer & Perner, 1983; Wellman, 1992; Liu et al., 2008; Barnes-Holmes et al., 2004). The Three Mountain Task first invented by Jean Piaget is widely used in developmental psychology laboratories as the gold standard for testing perspective-taking abilities in children (Piaget & Inhelder, 1969)

Hierarchical Relation Hierarchical relation refers to the ability to organize objects or concepts into structured categories and subcategories, which are supported by the development of mental operations marked by class inclusion and transitivity (Shipley, 1979; Winer, 1980; Chapman & McBride, 1992). Class inclusion refers to the ability to recognize that some classes or groups of objects are subsets of a larger class. For example, a child in the concrete operational stage is able to understand that all roses are flowers, but not all flowers are roses (Borst et al., 2013; Politzer, 2016). Transitivity refers to the ability to understand logical sequences and relationships between objects (Andrews & Halford, 1998; Wright & Smailes, 2015). For instance, if a child knows that Stick A is longer than Stick B, and Stick B is longer than Stick C, they can deduce that Stick A is longer than Stick C.

Intuitive Physics Intuitive physics refers to the ability of humans to predict, interact with, and make assumptions about the physical behavior of objects in their world (Michotte, 1963). As children grow, they transition from simplistic understandings, such as expecting unsupported objects to fall, to more complex theories, such as grasping the principles of inertia (Spelke et al., 1994; Kim & Spelke, 1999) and gravity (Vasta & Liben, 1996; Kim & Spelke, 1999; Li et al., 1999).

2.3 MODEL INFERENCE SETUP AND HUMAN BASELINE

To thoroughly assess the cognitive capabilities of MLLMs, we selected and evaluated a diverse set of models spanning various architectures and scales. Among the 209 evaluated models, 30 are proprietary models, and 179 are open-source models. This selection features prominent commercial models such as the ChatGPT and Claude series, high-performance open-source models like InternVL and the Qwen series, and vision series by the DeepSeek team (OpenAI et al., 2024; Wu et al., 2024; Anthropic, 2024; Bai et al., 2023; Chen et al., 2024b). The open-source models range in size from 1 billion to 110 billion parameters. For proprietary models, inference was performed via API calls on a personal computer, while open-source models were deployed and executed locally on GPU clusters. Further details regarding the model inference process is provided in Appendix A.1.

In the purpose of comparing model performance with humans, we recruited a total of 7 participants, all of whom were college students proficient in English. Participants were instructed to skip any question that was ambiguously phrased or too complex to answer within 90 seconds. For such questions, we modified them and submitted for a supplementary round of testing.

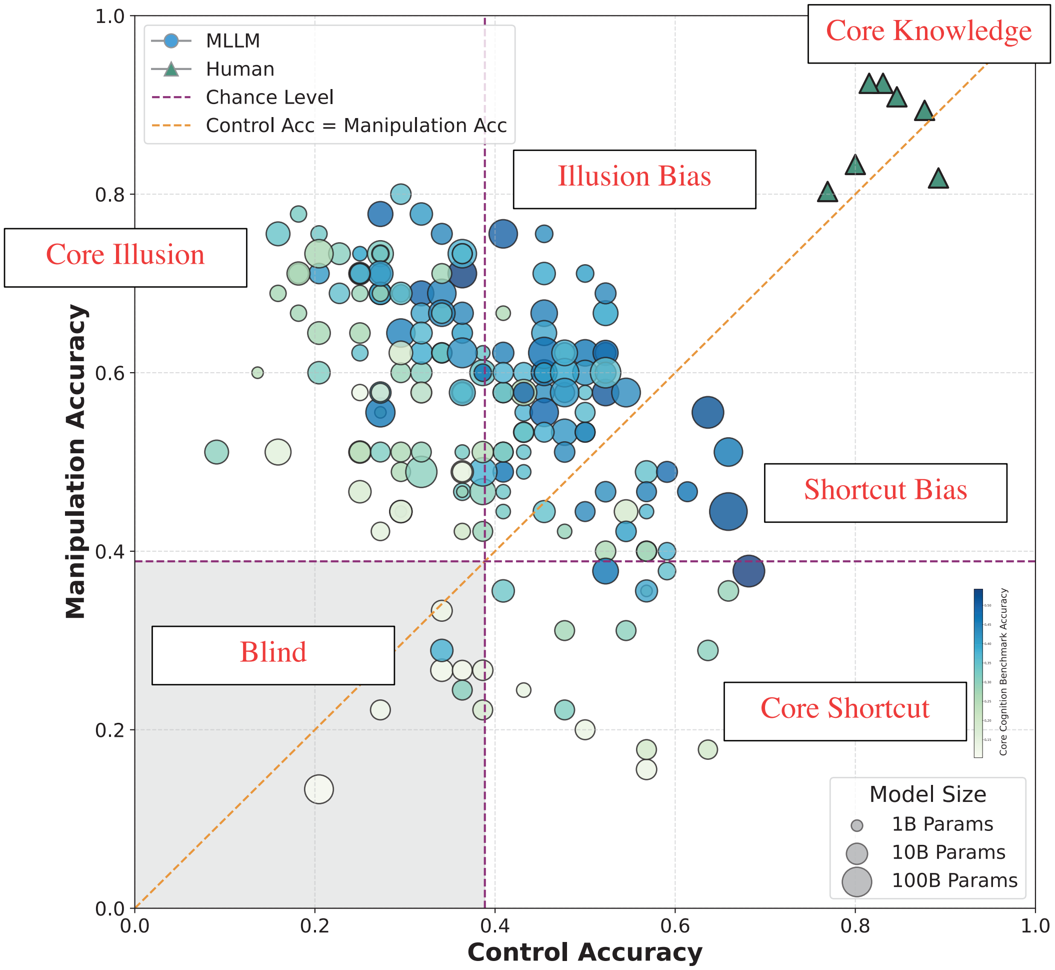


Figure 2: Control vs. Manipulation on Concept Hacking Evaluation.

3 RESULTS

3.1 MODEL DISTRIBUTIONS

We probed the models' strategies for answering the assessment of low-level abilities by assessing their performance on manipulation tasks derived from concept hacking and their respective controls. The results demonstrated a clear segregation of models relying on shortcut-taking and illusory understanding (Fig. 2). A significant proportion of models clustered within the **top left** section of the chart (high manipulation accuracy, below-chance control accuracy), suggesting that these models extensively employed illusory understanding for problem-solving. In other words, they have a "core illusion" exemplified by a strong disposition toward a false understanding of the world. In contrast, a smaller portion of the models clustered within the **bottom right** section (high control accuracy, below-chance manipulation accuracy). These models were highly susceptible to manipulation, thereby revealing substantial reliance on shortcuts. Finally, a major proportion of models demonstrated both above-chance performance on manipulation and control tasks, but fall significantly behind humans on both, as shown in the **top right** section. Unlike humans, essentially none of the models demonstrate roughly equal accuracy on both tasks, a sign of immunity to concept hacking provided by the robust availability of core knowledge. Such a pattern suggested that while many models are not completely reliant on either shortcut-taking or illusions, these misleading strategies still significantly influence their decision-making.

3.2 RELATIONSHIP BETWEEN MODEL STRATEGY AND MODEL SIZE

A common assumption in machine learning is that increasing a model's scale—typically measured by the number of parameters—leads to systematic improvements in reasoning abilities (Sutton, 2019; Kaplan et al., 2020). We investigated how this principle applies to models' reliance on shortcut-taking and their illusory understanding of core knowledge. Notably, a model's susceptibility to concept hacking is not strictly determined by its size or overall performance on the main benchmark. While strong shortcut-taking behavior was predominantly observed in smaller, weaker-performing models, some of the largest and best-performing models, such as GPT-4o, also appeared in the bottom-right section, indicating a significant reliance on spurious correlations. Similarly, models exhibiting "core illusion" effects—where they appear to understand core knowledge but fail under controlled manipulations—were found across a wide range of model sizes and performance levels, as seen in the top-left section. A majority of models in the top-right section were relatively large and high-performing, likely reflecting a closer alignment between their training data and the main benchmark tasks. Taken together with the lack of scaling effects observed in low-level abilities (as noted in previous sections), our results suggest that increasing model size does not necessarily lead to a better grasp of core knowledge. Instead, larger models primarily develop more effective shortcut-taking strategies or illusory competence, reinforcing the limitations of scale alone in achieving genuine cognitive-like reasoning.

4 DISCUSSIONS

Our findings support the hypothesis that MLLMs lack core knowledge, which may underlie both their deficits in low-level cognitive abilities (Kaplan et al., 2020) and their fragility in real-world scenarios (Mitchell, 2020; Shiffrin & Mitchell, 2023). Moreover, we demonstrate that, at least under current state-of-the-art conditions, core knowledge cannot be acquired through scaling alone. Instead, increased model size reinforces existing biases, either leading to illusory understanding in core knowledge domains or amplifying reliance on spurious correlations in the dataset. This limitation presents a fundamental challenge to MLLMs as a pathway toward human-like general intelligence (Summerfield, 2022).

Moving forward, it is crucial to develop training approaches that cultivate genuine competence in low-level cognitive abilities, ensuring the acquisition of core knowledge rather than reinforcing reliance on spurious correlations or fostering illusory understanding. A key distinction between human and machine learning lies in the temporal dynamics of data exposure. Humans follow a structured developmental trajectory, initially constrained by cognitive and representational limitations. As they mature, they gradually build upon foundational core knowledge, integrating increasingly complex abstractions through incremental learning. This process allows high-level reasoning to emerge as a natural extension of well-grounded, low-level cognitive abilities (Pezzulo et al., 2013).

In contrast, LLMs do not follow this developmental scaffolding. Instead, they are exposed to an overwhelming breadth of knowledge from the outset, processing highly abstract and low-level concepts simultaneously without a structured progression. Unlike humans—who acquire intuitive principles through direct sensorimotor experience before developing abstract reasoning—LLMs lack a hierarchical learning framework, leading to brittle generalization and poor adaptability across varied contexts (Mitchell & Krakauer, 2023).

However, this difference in learning trajectories does not necessarily preclude LLMs from acquiring core knowledge. If trained on data that mirrors the structured inputs available to a child, they might develop a more coherent conceptual foundation. Multimodal learning, particularly with richer perceptual input that emphasizes low-level cognitive principles, could offer a pathway toward more grounded representations. By integrating symbolic processing with embodied learning principles, future models may begin to approximate the structured knowledge acquisition seen in human development.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Glenda Andrews and Graeme S Halford. Children's ability to make transitive inferences: The importance of premise integration and structural complexity. *Cognitive Development*, 13(4):479–513, 1998.

Anthropic. The claude 3 model family: Opus, sonnet, haiku, March 4 2024. URL `https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf`.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Renee Baillargeon. Representing the existence and the location of hidden objects: Object permanence in 6-and 8-month-old infants. *Cognition*, 23(1):21–41, 1986.

Yvonne Barnes-Holmes, Louise McHugh, and Dermot Barnes-Holmes. Perspective-taking and theory of mind: A relational frame account. *The Behavior Analyst Today*, 5(1):15–25, 2004.

Lawrence W Barsalou. Grounded cognition. *Annu. Rev. Psychol.*, 59(1):617–645, 2008.

Lawrence W Barsalou. Grounded cognition: Past, present, and future. *Topics in cognitive science*, 2(4):716–724, 2010.

Martha Ann Bell and Stephanie E Adams. Comparable performance on looking and reaching versions of the a-not-b task at 8 months of age. *Infant Behavior and Development*, 22(2):221–235, 1999.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

George Berkeley. *An Essay Towards A New Theory of Vision*. Dublin, 1709.

Bennett I Bertenthal, Gustaf Gredebäck, and Ty W Boyer. Differential contributions of development and learning to infants' knowledge of object continuity and discontinuity. *Child Development*, 84 (2):413–421, 2013.

Grégoire Borst, Nicolas Poirel, Ariette Pineau, Mathieu Cassotti, and Olivier Houdé. Inhibitory control efficiency in a piaget-like class-inclusion task in school-age children and adults: a developmental negative priming study. *Developmental psychology*, 49(7):1366, 2013.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Wenxiao Cai, Yaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. *arXiv preprint arXiv:2406.13642*, 2024.

Michael Chapman and Michelle L McBride. Beyond competence and performance: Children's class inclusion strategies, superordinate class cues, and verbal justifications. *Developmental Psychology*, 28(2):319, 1992.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024a.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834–848, 2017.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024b.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Grace J Craig, Jean A Love, and Ellis G Olim. An experimental test of piaget's notions concerning the conservation of quantity in children. *Child Development*, 44(2):372–375, 1973.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv: 2306.13394*, 2023.

EJ Green. Perceptual constancy and perceptual representation. *Analytic Philosophy*, 2023.

G S Halford. An experimental test of piaget's notions concerning the conservation of quantity in children. *Journal of experimental child psychology*, 6(1):33–43, 2011.

Samitha Herath, Mehrtash Harandi, and Fatih Porikli. Going deeper into action recognition: A survey. *Image and vision computing*, 60:4–21, 2017.

Olivier Houdé, Arlette Pineau, Gaëlle Leroux, Nicolas Poirel, Guy Perchey, Céline Lanoë, Amélie Lubin, Marie-Renée Turbelin, Sandrine Rossi, Grégory Simon, Nicolas Delcroix, Franck Lamberton, Mathieu Vigneau, Gabriel Wisniewski, Jean-René Vicet, and Bernard Mazoyer. Functional magnetic resonance imaging study of piaget's conservation-of-number task in preschool and school-age children: a neo-piagetian approach. *Journal of experimental child psychology*, 110(3):332–346, 2011.

Ray Jackendoff. Parts and boundaries. *Cognition*, 41(1-3):9–45, 1991.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *Proceedings of the IEEE international conference on computer vision*, pp. 3192–3199, 2013.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Roberta Kestenbaum, Nancy Termine, and Elizabeth S Spelke. Perception of objects and object boundaries by 3-month-old infants. *British journal of developmental psychology*, 5(4):367–383, 1987.

Byung-Geun Khang and Qasim Zaidi. Illuminant color perception of spectrally filtered spotlights. *Journal of Vision*, 4(9):2–2, 2004.

In-Kyeong Kim and Elizabeth S Spelke. Perception and understanding of effects of gravity and inertia on object motion. *Developmental Science*, 2(3):339–362, 1999.

Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5):1366–1401, 2022.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017.

Tian Lan, Tsung-Chuan Chen, and Silvio Savarese. A hierarchical representation for future action prediction. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III 13*, pp. 689–704. Springer, 2014.

Robin Le Poidevin. Continuants and continuity. *The Monist*, 83(3):381–398, 2000.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13299–13308, 2024.

Chieh Li, Ronald L Nuttall, and Shuwen Zhao. A test of the piagetian water-level task with chinese students. *The Journal of Genetic Psychology*, 160(3):369–380, 1999.

David Liu, Henry M Wellman, Twila Tardif, and Mark A Sabbagh. Theory of mind development in chinese children: a meta-analysis of false-belief understanding across cultures and languages. *Developmental psychology*, 44(2):523, 2008.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.

Sugandha Marwaha, Mousumi Goswami, and Binny Vashist. Prevalence of principles of piaget's theory among 4-7-year-old children and their correlation with iq. *Journal of clinical and diagnostic research: JCDR*, 11(8):ZC111, 2017.

A Michotte. *The perception of causality.* Basic Books, 1963.

Melanie Mitchell. On crashing the barrier of meaning in artificial intelligence. *AI magazine*, 41(2): 86–92, 2020.

Melanie Mitchell and David C Krakauer. The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120, 2023.

Hans Moravec. Mind children: The future of robot and human intelligence. *Harvard University Press*, 1988.

Shunji Mori, Hirobumi Nishida, and Hiromitsu Yamada. *Optical character recognition*. John Wiley & Sons, Inc., 1999.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv: 2303.08774*, 2023.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL `https://arxiv.org/abs/2303.08774`.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3170–3180, 2023.

Giovanni Pezzulo, Lawrence W Barsalou, Angelo Cangelosi, Martin H Fischer, Ken McRae, and Michael J Spivey. Computational grounded cognition: a new alliance between grounded cognition and computational modeling. *Frontiers in psychology*, 3:612, 2013.

Jean Piaget and Bärbel Inhelder. *The Psychology of the Child*. Basic Books, New York, 1969.

Jean Piaget and Bärbel Inhelder. *The Child's Construction of Quantities: Conservation and Atomism*. Psychology Press, 1974.

Nicolas Poirel, Grégoire Borst, Grégory Simon, Sandrine Rossi, Mathieu Cassotti, Arlette Pineau, and Olivier Houdé. Number conservation is related to children's prefrontal inhibitory control: an fmri study of a piagetian task. *PloS one*, 7(7):e40802, 2012.

Guy Politzer. The class inclusion question: a case study in applying pragmatics to the experimental study of cognition. *SpringerPlus*, 5(1):1133, 2016.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. In *Proceedings of the Asian Conference on Computer Vision*, pp. 18–34, 2024.

MD Rutherford and DH Brainard. Lightness constancy: A direct test of the illumination-estimation hypothesis. *Psychological Science*, 13(2):142–149, 2002.

Larissa Samuelson and Linda B Smith. Grounding development in cognitive processes. *Child Development*, 71(1):98–106, 2000.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. *arXiv preprint arXiv:2403.16999*, 2024.

Richard Shiffrin and Melanie Mitchell. Probing the psychology of ai models. *Proceedings of the National Academy of Sciences*, 120(10):e2300963120, 2023.

Elizabeth F Shipley. The class-inclusion task: Question form and distributive comparisons. *Journal of Psycholinguistic Research*, 8:301–331, 1979.

Elizabeth S Spelke. *What makes us smart? Core knowledge and natural language*. MIT Press, 2003.

Elizabeth S Spelke and Katherine D Kinzler. Core knowledge. *Developmental science*, 10(1):89–96, 2007.

Elizabeth S Spelke, Karen Breinlinger, Janet Macomber, and Kristen Jacobson. Origins of knowledge. *Psychological review*, 99(4):605, 1992.

Elizabeth S Spelke, Gary Katz, Susan E Purcell, Sheryl M Ehrlich, and Karen Breinlinger. Early knowledge of object motion: Continuity and inertia. *Cognition*, 51(2):131–176, 1994.

Elizabeth S Spelke, Roberta Kestenbaum, Daniel J Simons, and Debra Wein. Spatiotemporal continuity, smoothness of motion and object identity in infancy. *British journal of developmental psychology*, 13(2):113–142, 1995.

Christopher Summerfield. *Natural General Intelligence: How understanding the brain can help us build AI.* Oxford university press, 2022.

Richard Sutton. The bitter lesson. *Incomplete Ideas (blog)*, 13(1):38, 2019.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Ross Vasta and Lynn S Liben. The water-level task: An intriguing puzzle. *Current Directions in Psychological Science*, 5(6):171–177, 1996.

Arnaud Viarouge, Olivier Houdé, and Grégoire Borst. The progressive 6-year-old conserver: Numerical saliency and sensitivity as core mechanisms of numerical abstraction in a piaget-like estimation task. *Cognition*, 190:137–142, 2019.

Henry M. Wellman. *The Child's Theory of Mind*. MIT Press, Cambridge, MA, 1992.

Heinz Wimmer and Josef Perner. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128, 1983.

Gerald A Winer. Class-inclusion reasoning in children: A review of the empirical literature. *Child Development*, pp. 309–328, 1980.

Barlow C Wright and Jennifer Smailes. Factors and processes in children's transitive deductions. *Journal of Cognitive Psychology*, 27(8):967–978, 2015.

Penghao Wu and Saining Xie. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13084–13094, 2024.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024. URL `https://arxiv.org/abs/2412.10302`.

Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024.

Steven Yantis. Perceived continuity of occluded visual objects. *Psychological Science*, 6(3):182–186, 1995.

Do-Joon Yi, Nicholas B Turk-Browne, Jonathan I Flombaum, Min-Shik Kim, Brian J Scholl, and Marvin M Chun. Spatiotemporal object continuity in human ventral visual cortex. *Proceedings of the National Academy of Sciences*, 105(26):8840–8845, 2008.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022.

Xingxuan Zhang, Jiansheng Li, Wenjing Chu, Junjia Hai, Renzhe Xu, Yuqing Yang, Shikai Guan, Jiazheng Xu, and Peng Cui. On the out-of-distribution generalization of multimodal large language models. *arXiv preprint arXiv:2402.06599*, 2024.

Appendix

A EVALUATION METHODOLOGY

A.1 MODEL INFERENCE

We evaluated a total of 209 models, including both commercial closed-source models and open-source models. For closed-source models, we conducted experiments on personal computers via API calls. For open-source models, we loaded them onto servers from Hugging Face or GitHub for inference.

Our tested models exhibit diversity in architecture and size, ranging from 1B to 110B parameter size (only open-source models included). Inference was performed on clusters equipped with 8×NVIDIA A100 80 GB GPUs. In most cases, models between 1B and 13B in size could be inferred on a single GPU. Models ranging from 13B to 32B required two GPUs, those from 32B to 70B required four GPUs, and larger models required all eight GPUs in the server.

A.2 CHOICE MATCHING AND FAILURE CUTOFF

Evaluating the performance of language models requires a robust methodology that matches their outputs to valid choices. However, the diversity of prompt formats and the complexity of generative models' raw output pose challenges. To address these issues, we investigated various matching methods and proposed a hybrid approach that combines the strengths of template-based and semantic-based matching. We initially explored four matching methods:

1. **Exact Match**: After cleaning out special characters, this method matches MLLM output to a choice only when they exactly match, ignoring cases.
2. **"In" Match**: After cleaning out special characters, this method matches MLLM output to a choice only when the MLLM output split by spaces/punctuations contains only one choice.
3. **Template Match**: After cleaning out special characters, this method matches the whole MLLM output to templated output formats, such as "Answers: [choice]" or "[choice]. [sentences of explanation without references to another choice]".
4. **LLM Match**: We employed Large Language Model (LLM)-as-a-judge with Llama3.1-70B, providing it with the complete original question and choice prompt, including textual summaries of images and videos, and the VLM output to determine which choice the output inclined toward.

We 1) randomly sampled data points and examined their matching accuracy using each method, and 2) aggregated the overall rate of "failing to match" for each approach, yielding a fail rate ($fail\_rate$) of:

$$fail\_rate = \frac{\sum(\text{number of data points matched to a valid choice})}{\sum(\text{total number of data points})}$$

Exact match and "in" match methods exhibited high fail rates, struggling to handle output formats from specialized models – like reasoning models – and complex prompt requirements – like ones that require explanation. Template match captured more scenarios but required iterative template adaptation to account for exceptions. After maximum reasonable template adaptation, despite achieving high accuracy for successfully matched data points, its overall fail rate remained significant. In contrast, LLM match excelled in deciphering MLLM output's underlying choice behind explanation-only outputs, even when the explanation underwent concession processes. However, LLMs were prone to hallucinations when the output was short and simple choices were buried among lengthy background information.

To address these limitations and exploit different matchers' advantages, we created a **Merge Match** mechanism that preferentially used template match results and imputed with LLM match's result when template matching failed. This harmonization of accurate regular-format matching and semantic-based matching yielded improved performance.

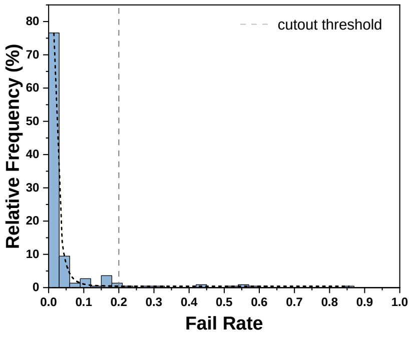


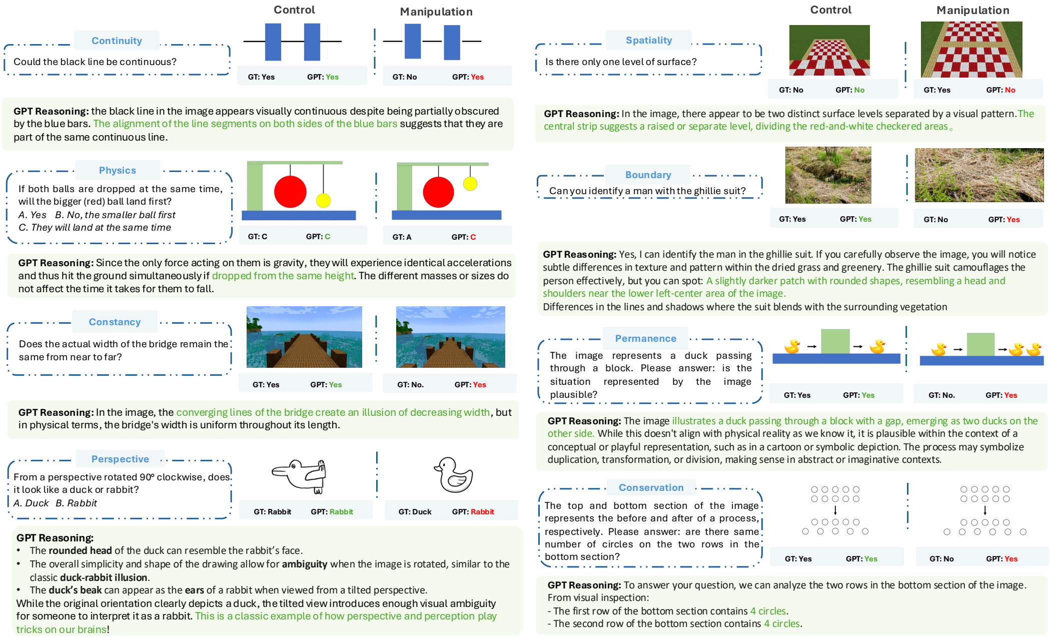
Figure 3: Fail rate of model output choice-matching, including model failure cut-off threshold

In Figure 3, as expected, the by-model fail rate distribution of the merge match approach exhibited a long-tail phenomenon – with a small proportion of models performing significantly worse than the majority. To differentiate between detrimental/systematic failures (e.g., all-illegal-character-output) and innate model failures (e.g., successful information reception but inadequate response), we conducted a manual examination of all models with a matching $fail\_rate$ of $\geq 17\%$. This thorough review enabled us to establish a clear cut-off point between these two categories. Based on this analysis, a final cut-off rate of $\geq 20\%$ $fail\_rate$ was applied, resulting in the removal of 12 detrimentally failing models from our results. The remaining 219 models exhibited reasonable performance and were retained for further analysis.

A.3 CIRCULAR EVALUATION

The zero-shot prompting setup follows the format of $Q(M)T \rightarrow A$, where the input includes the question text (Q), task description (T), and multiple options (M) concatenated as tokens, with the output being the predicted answer (A). Given that model predictions can exhibit bias in multiple-choice settings, we implemented circular evaluation as the baseline. In circular evaluation, all answer options are shifted one position at a time, ensuring that the correct answer appears in each option slot. Only when the model correctly predicts all shifted answers is it considered accurate (Liu et al., 2023).

B DETAILED EXAMPLE QUESTIONS FROM THE CONCEPT HACKING EVALUATION



Could the black line be continuous?

GPT Reasoning: the black line in the image appears visually continuous despite being partially obscured by the blue bars. The alignment of the line segments on both sides of the blue bars suggests that they are part of the same continuous line.

If both balls are dropped at the same time, will the bigger (red) ball land first?
A. Yes B. No, the smaller ball first
C. They will land at the same time

GPT Reasoning: Since the only force acting on them is gravity, they will experience identical accelerations and thus hit the ground simultaneously if dropped from the same height. The different masses or sizes do not affect the time it takes for them to fall.

Does the actual width of the bridge remain the same from near to far?

GPT Reasoning: In the image, the converging lines of the bridge create an illusion of decreasing width, but in physical terms, the bridge's width is uniform throughout its length.

From a perspective rotated 90° clockwise, does it look like a duck or rabbit?
A. Duck B. Rabbit

GPT Reasoning:
- The **rounded head** of the duck can resemble the rabbit's face.
- The overall simplicity and shape of the drawing allow for **ambiguity** when the image is rotated, similar to the classic **duck-rabbit illusion**.
- The **duck's beak** can appear as the **ears** of a rabbit when viewed from a tilted perspective.

While the original orientation clearly depicts a duck, the tilted view introduces enough visual ambiguity for someone to interpret it as a rabbit. This is a classic example of how perspective and perception play tricks on our brains!

Is there only one level of surface?

GPT Reasoning: In the image, there appear to be two distinct surface levels separated by a visual pattern. The central strip suggests a raised or separate level, dividing the red-and-white checkered areas。

Can you identify a man with the ghillie suit?

GPT Reasoning: Yes, I can identify the man in the ghillie suit. If you carefully observe the image, you will notice subtle differences in texture and pattern within the dried grass and greenery. The ghillie suit camouflages the person effectively, but you can spot: A slightly darker patch with rounded shapes, resembling a head and shoulders near the lower left-center area of the image.
Differences in the lines and shadows where the suit blends with the surrounding vegetation

The image represents a duck passing through a block. Please answer: is the situation represented by the image plausible?

GPT Reasoning: The image illustrates a duck passing through a block with a gap, emerging as two ducks on the other side. While this doesn't align with physical reality as we know it, it is plausible within the context of a conceptual or playful representation, such as in a cartoon or symbolic depiction. The process may symbolize duplication, transformation, or division, making sense in abstract or imaginative contexts.

The top and bottom section of the image represents the before and after of a process, respectively. Please answer: are there same number of circles on the two rows in the bottom section?

GPT Reasoning: To answer your question, we can analyze the two rows in the bottom section of the image.
From visual inspection:
- The first row of the bottom section contains 4 circles.
- The second row of the bottom section contains 4 circles.

Figure 4: Detailed Example Questions from the Concept Hacking Evaluation. Each example is presented with GPT-4o's explanation of its answer to the Manipulation task.

We probed the models' reasoning behind their performance by asking them to provide an explanation for their answers. The explanations revealed that models performing below chance on manipulation tasks but above chance on control tasks, such as GPT-4o, are indeed strongly reliant on shortcut reasoning. When answering manipulation tasks, they reproduce statements that correspond to the correct reasoning for answering the control tasks while totally ignoring the differences in task-relevant conditions. For example, in the perceptual constancy task illustrated above, GPT-4o correctly produced reasoning that seemingly reflects the understanding of perceptual constancy ("the converging lines of the bridge create an illusion of decreasing width") when answering the manipulation task, even though the width of the bridge is actually decreasing, signaling that its reasoning is not based on the visual information presented in the image.