Core Knowledge Deficits in Multi-Modal Language Models

Yijiang Li[1] Qingying Gao[*2] Tianwei Zhao[*2] Bingyang Wang[*3] Haoran Sun[2] Haiyun Lyu[4]
Robert D. Hawkins[5] Nuno Vasconcelos[1] Tal Golan[6] Dezhi Luo[7 8] Hokin Deng[9]

Abstract

While Multi-modal Large Language Models (MLLMs) demonstrate impressive abilities over high-level perception and reasoning, their robustness in the wild still lags behind humans and exhibits diminished efficacy on simple tasks that are intuitive for humans. We examine the hypothesis that these deficiencies stem from the absence of core knowledge—rudimentary cognitive abilities innate to humans from early childhood. To probe core knowledge representation in MLLMs, we draw from developmental cognitive sciences and develop a large-scale benchmark, **CoreCognition**, encompassing 12 core cognitive concepts. We evaluate 219 models with 11 different prompts, leading to a total of 2409 data points for analysis. Our findings reveal core knowledge deficits in early-developed core abilities while models demonstrate human-comparable performance in high-level cognition. Moreover, we find that low-level abilities show little to no scaling, in stark contrast to high-level abilities. Finally, we introduce an evaluation technique "Concept Hacking", through which we demonstrate that MLLMs do not genuinely advance toward core knowledge but instead rely on illusory understanding and shortcut learning as they scale.

1. Introduction

Are human minds born with knowledge (Plato et al., 1763)? This has been the central question of Western thoughts since the ancient Greeks (Russell, 1946). Socrates and Plato both believe that humans must be born with a set of innate knowledge. In *Meno, 80d–86b*, Socrates introduces the theory of *anamnesis* (recollection), where he suggests our "soul is immortal", and "it can recollect the things it knew before" (Fowler et al., 1914). Plato further sets the distinction between innate knowledge and those we gain through experience: in *Republic VII*, the Allegory of the Cave, he suggests that our experiences are *skiés*, like shadows on the cave wall, which are contingent instantiations of the *eidos*, the knowledge born with our minds. One example of *eidos* is our understanding of a circle: while a perfect circle never exists in reality, we still understand what it means to be a perfect circle (Jowett et al., 1888). Kant's view is more intricate: he suggests we never have an innate knowledge of *noumena*, "things-in-themselves", but we have knowledge of *phenomenon*, "things-about-themselves", meaning we only are born with knowledge about the structures of our experiences, such as causality, permanence, and continuity, but never gifted with knowledge of experiences in itself (Kant, 1781). In other words, we have innate, core knowledge about basic domains of the world.

We are closer than ever to achieving human-level intelligence. By training on vast web-scale corpora and scaling to hundreds of billions of parameters, Large Language Models (LLMs) now surpass expert humans in knowledge- and reasoning-intensive tasks (Brown et al., 2020; Achiam et al., 2023; Bai et al., 2023; Touvron et al., 2023; Jaech et al., 2024). These capabilities extend beyond language: with modality alignment (Liu et al., 2024a; Li et al., 2023a; Zhu et al., 2023), Multi-modal Large Language Models (MLLMs) exhibit unprecedented high-level perception and reasoning (Gemini, 2023; Wu & Xie, 2024; Xu et al., 2024; Yang et al., 2025b; Shao et al., 2024; Yang et al., 2025a; Li et al., 2024; Fu et al., 2023), mastering tasks such as chart understanding (Masry et al., 2022), geometry and math (Lu et al., 2023), and action recognition and prediction (Ying et al., 2024; Liu et al., 2024b), often reaching or exceeding human performance (Huang & Zhang, 2024).

Despite these advances in high-level abilities, state-of-the-art MLLMs still fall short of human on simple and rudimentary tasks such as counting (Paiss et al., 2023; Chia et al., 2024), perspective taking (Tang et al., 2025b), spatial reasoning (Zhang et al., 2025; Tang et al., 2025a) and

[*]Equal contribution [1]University of California San Diego [2]Johns Hopkins University [3]Emory University [4]University of North Carolina at Chapel Hill [5]Stanford University [6]Ben-Gurion University of the Negev [7]University of Michigan [8]University College London [9]Carnegie Mellon University. Correspondence to: Yijiang Li <yijiangli@ucsd.edu>, Dezhi Luo <ihzedoul@umich.edu>, Hokin Deng <hokind@andrew.cmu.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada, PMLR 267, 2025. Copyright 2025 by the author(s).

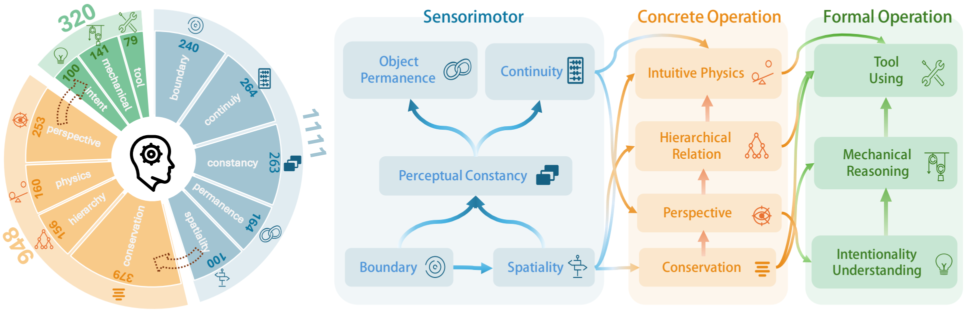
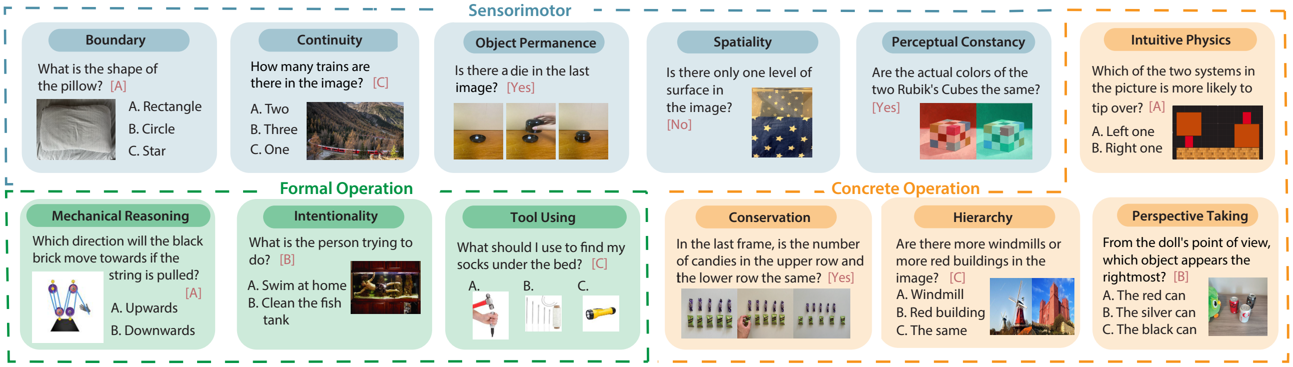


Figure 1. **Left.** Statistics of the **CoreCognition** benchmark. **Right.** Construction of taxonomy. Dependencies between abilities are indicated with arrows.



What is the shape of the pillow? [A]
A. Rectangle
B. Circle
C. Star

How many trains are there in the image? [C]
A. Two
B. Three
C. One

Is there a die in the last image? [Yes]

Is there only one level of surface in the image? [No]

Are the actual colors of the two Rubik's Cubes the same? [Yes]

Which of the two systems in the picture is more likely to tip over? [A]
A. Left one
B. Right one

Which direction will the black brick move towards if the string is pulled? [A]
A. Upwards
B. Downwards

What is the person trying to do? [B]
A. Swim at home
B. Clean the fish tank

What should I use to find my socks under the bed? [C]
A.
B.
C.

In the last frame, is the number of candies in the upper row and the lower row the same? [Yes]

Are there more windmills or more red buildings in the image? [C]
A. Windmill
B. Red building
C. The same

From the doll's point of view, which object appears the rightmost? [B]
A. The red can
B. The silver can
C. The black can

Figure 2. Examples from our CoreCognition benchmark.

Concept	Definition	Concept	Definition	Concept	Definition
Boundary	The transition from one object to another.	**Continuity**	Expecting objects to continue existing and moving predictably.	**Permanence**	Persistence of things across space and time out of perception.
Spatiality	The *a priori* understanding of the Euclidean properties of the world.	**Perceptual Constancy**	Changes in appearances don't mean changes in physical properties.	**Intuitive Physics**	Intuitions about the laws of how things interact in the physical world.
Perspective	To see what others see.	**Hierarchy**	Understanding of inclusion and exclusion of objects and categories.	**Conservation**	Invariances of properties despite transformations.
Tool Use	The capacity to manipulate specific objects to achieve goals.	**Intentionality**	To see what others want.	**Mechanical Reasoning**	Inferring actions from system states and vice versa.

Table 1. Abbreviated definitions of the 12 core abilities assessed. See Appendix A.3 for details.

compositional reasoning (Yuksekgonul et al., 2022; Sahin et al., 2024; Mitra et al., 2024) that are intuitive and easy for humans, despite their excellence at high-level reasoning tasks on similar domains (Paiss et al., 2023; Rahmanzadehgervi et al., 2024) (Moravec's Paradox (Moravec, 1988)). Excellence often does not appear to translate to more generalized and real-world contexts, where minor changes in task conditions can lead to dramatic failures (Shiffrin & Mitchell, 2023; Zhang et al., 2024a; Bai et al., 2024; Oh et al., 2025; Dong et al., 2025).

In this work, we hypothesize that the deficiencies observed in MLLMs stem from the absence of core knowledge—fundamental cognitive abilities innately present in humans from early childhood that underpin advanced reasoning. To examine this hypothesis, we explore the existence, representation, and use of core knowledge in MLLMs by introducing the first large-scale benchmark tailored for core knowledge. Drawing on insights from developmental cognitive science, we propose a taxonomy of 12 abilities encompassing the full spectrum of core knowledge, from basic cognitive skills to advanced reasoning. Based on this taxonomy, we present **CoreCognition** comprising of 2,379 samples with over 100 examples for each concept, as examplfied in Fig 2.

To provide a comprehensive evaluation of core-knowledge over the existing MLLMs, we assess a total of 219 models with 11 different prompting techniques, yielding a total of 2,409 data points. Leveraging these results, we analyze model performance across varying levels of core ability, examining the interdependencies among core knowledge and their predictive power for higher-level reasoning and perception, as well as the scaling effect (performance across different model sizes) To further ascertain core knowledge deficits in MLLMs, we design controlled experiments that manipulate causal features within images to perturb the ground-truth labels, allowing us to determine whether models genuinely possess the targeted core knowledge or merely approximate it through shortcuts and spurious correlations.

Our key findings are:

- MLLMs consistently perform worse on low-level abilities compared to high-level abilities (Section 4.2).
- Performance of MLLMs on high-level abilities does not correlate with the corresponding low-level abilities that serve as foundations for them in humans (Section 4.3).
- No observable scaling on low-level abilities with respect to increasing model parameters (Section 4.5).
- Rather than possessing and leveraging core knowledge, models rely on misleading strategies such as shortcut-taking and illusory understanding to answer questions (Section 5.2).

2. Related Works

Multi-modal Large Language Models. With the advent of large language models (LLMs), state-of-the-art (SOTA) MLLMs (Liu et al., 2024a; Li et al., 2023b) have adopted open-source LLMs (Touvron et al., 2023; Peng et al., 2023; Jiang et al., 2023) and aligned visual features to the LLM embedding space (Li et al., 2023a). To enable open-ended conversational abilities, LLaVA (Liu et al., 2024a) distills ChatGPT's conversational skills into MLLMs, resulting in substantial performance gains—a process that has become standard practice in the field (Wang et al., 2023; Bai et al., 2023; Gemini, 2023; Team, 2024; Sun et al., 2023; Li et al., 2022).

Benchmarks for Multi-modal Large Language Model.

A wide range of benchmarks has been proposed to evaluate the growing capabilities of Multi-modal Large Language Models (MLLMs), spanning vision-language perception (Antol et al., 2015; Marino et al., 2019; Xu et al., 2025), OCR and text understanding (Liu et al., 2023b), hallucination detection (Li et al., 2023c; Liang et al., 2022), and robustness to adversarial attacks (Zhao et al., 2024). Holistic evaluations such as SEED-Bench (Li et al., 2024), MM-Bench (Liu et al., 2024b), and LAMM (Yin et al., 2024) aim to provide broad coverage across modalities, tasks, and reasoning levels. Recent cognitively inspired efforts such as M3GIA (Song et al., 2024), and Marvel (Jiang et al., 2024) explore dimensions of cognitive complexity, abstraction, and multi-step reasoning. However, while these benchmarks probe various aspects of cognition, they primarily focus on task coverage or high-level general intelligence. In contrast, CoreCognition is grounded in developmental cognitive science, targeting early-emerging core knowledge abilities fundamental to human reasoning.

Core Knowledge in Humans.

The debate over core knowledge has historically framed nativist and empiricist epistemologies (Plato et al., 1763; Kant, 1781; Russell, 1946), and since the cognitive revolution, has shifted toward empirical investigation (Piaget, 1950; Fodor, 1975). Piaget's stage-based theory and subsequent research established the foundations of developmental psychology (Piaget & Inhelder, 1969; Barrouillet, 2015; Spelke et al., 1992; Rochat, 2024; Carey et al., 2015). Recent advances show that even infants exhibit rudimentary knowledge of objects (Baillargeon & Carey, 2012; Kar et al., 2019; Ullman & Tenenbaum, 2020), actions (Yang et al., 2015; Jara-Ettinger et al., 2020), numbers (Feigenson et al., 2004; Hannagan et al., 2015; Spelke, 2017), space (Newcombe & Sluzenski, 2004; Bellmund et al., 2018), and social relations (Siegal & Varley, 2002; Scott & Baillargeon, 2017; Spelke, 2022). This "developmental start-up software" enables early learning (Spelke & Kinzler, 2007; Lake et al., 2017) and serves as the foundation for complex reasoning in variable environments later in life (Barsalou, 2020; Mitchell, 2021).

3. Benchmarking Core Knowledge in Multi-modal Large Language Models

We introduce **CoreCognition**, encompassing 12 core abilities and 2,379 questions with diverse input types and formats. **CoreCognition** covers cognitive development stages from the Sensorimotor to Concrete Operational and ultimately the Formal Operational stage (Piaget, 1950; 1952; Piaget & Inhelder, 1969; 1974). An overview of the benchmark and its distribution is shown in Fig.1, with 12 representative examples in Fig.2. Section 3.1 outlines the cognitive taxonomy and theoretical framework guiding our benchmark. Section 3.2 details the curation process, while Sections 3.3 and 4.1 describe model inference and evaluation.

3.1. Cognitive Framework

We follow Jean Piaget's theory (Piaget, 1950; Piaget & Inhelder, 1969; 1974), which identifies four stages in human developmental trajectory: Sensorimotor, Preoperational, Concrete Operational, and Formal Operational. In the Sensorimotor stage, infants develop core concepts such as object permanence (Spelke et al., 1992; Bremner et al., 2015) and perceptual constancy (Green, 2023) through sensory and physical interactions. The Preoperational stage serves as a transitional phase, characterized not by distinct new abilities but by the gradual solidification of symbolic representations (Fodor, 1975). These cognitive advancements culminate in the Concrete Operational stage, where children acquire abilities for systematic reasoning about numbers, motion, and agents, including perspective-taking, conservation, intuitive physics, and hierarchical relations (Piaget & Inhelder, 1974; Moll & Meltzoff, 2011; Piloto et al., 2022; Murphy & Lassaline, 2013). The Formal Operational stage extends these abilities to abstract reasoning and complex tasks, such as understanding intentionality and mechanical reasoning (Kilner, 2011; Allen et al., 2020). Overall, we depict this developmental hierarchy in Figure 1. See Appendix A.2 for empirical support of this framework. We provide detailed description of these core abilities in Appendix A.3.

3.2. Dataset Curation

Building on the above cognitive framework, we now describe how we operationalize these theoretical concepts into concrete instances for probing specific core knowledge in MLLMs. **Prototyping** We conduct a systematic review of developmental psychology literature to identify experimental paradigms for evaluating cognitive abilities. Selections are based on construct alignment and empirical credibility. For each ability, 5–10 classical experimental prototypes were selected, each representing a distinctive operationalization.

Experiment design with toolkits To ensure alignment with developmental psychology findings while facilitating effective benchmarking, we utilize commonly available objects, online datasets, or digital modeling using software toolkits to adapt the experimental designs from the original literature.

Question design and review After the experiment was designed, two supervising researchers conducted reviews to ensure that the questions were not only pertinent to but also effectively probed the specific abilities. Example questions collected under the said paradigms are shown in Fig. 2.

Human benchmark as quality check A quality check was conducted by collecting 20 human answers for each question, proceeding only when accuracy exceeded 80%.

Further details of the process is provided in Appendix C.

3.3. Evaluation Strategy

To deal with free-form outputs from MLLMs while maintaining robustness and affordability, we employ a two-stage approach. First, the model's response is matched to one of the predefined choices or classified as a failure (FAIL). Subsequently, the matched option is compared against the ground truth to determine correctness, with FAIL automatically considered incorrect.

We employ a combination of template matching and LLM matching as the default matching method during evaluation. Specifically, a list of templates will first be used to match the model's output with one of the options. Then the failed examples will be forwarded to LLM to match the corresponding option. Please refer to Appendix D.2 for a detailed description and discussion on the five matching methods explored in this paper.

Furthermore, we adopt circular evaluation (Liu et al., 2023a) to avoid the model being lucky. Concretely, circular evaluation shifts a k-choice MCQ k times to avoid model biasing towards specific options. Only when the model answers all k questions is it determined correctly on this question.

Further details of the process is provided in Appendix D.

4. Experiments

4.1. Setup

To thoroughly assess the cognitive capabilities of multimodal language models, we meticulously selected and evaluated a diverse set of models spanning various architectures and scales. Among the 231 evaluated models, 25 are proprietary models, and 206 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 recently introduced models from the DeepSeek series that have garnered significant attention. 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. Additional details on model inference can be found in Appendix D.1.

However, out of 231 models, some models exhibited systematic failures, such as consistently producing invalid outputs. To ensure reliability and avoid our results to be contaminated, we excluded these models, retaining 219 for further analysis. The detailed filtering process is documented in Appendix D.

4.2. Main Results

In Table 2, we compare the performance of different MLLMs and human performance on the **CoreCognition** benchmarks. Specifically, we select 17 high-performance models including 8 Proprietary Models such as Qwen-VL-Max, GPT series (Hurst et al., 2024), Gemini series (Gemini, 2023) as well as Claude series (Anthropic, 2024) and 9 Open Source Models, namely Qwen2.5-VL (Team, 2025), InternVL2 (Chen et al., 2024), LLaVA-Video (Zhang et al., 2024b), NVLM-D-72B (Dai et al., 2024), mPLUG-Owl3 (Ye et al., 2024), VILA1.5 (Lin et al., 2023), Pixtral-12B (Agrawal et al., 2024), and deepseek-vl2 (Wu et al., 2024). It can clearly be inferred that even the best model such as GPT-4o and Qwen2.5-VL-72B-Instruct lag behind human performance by a large margin. Moreover, from the table, models' performance in Formal Operations is relatively

	Sensorimotor					Concrete Operation				Formal Operation		
Model	Boundary	Continuity	Permanence	Spatiality	Perceptual Constancy	Intuitive Physics	Perspective Taking	Conservation	Hierarchical Relation	Intentionality Understanding	Mechanical Reasoning	Tool Using
Human	82.45%	94.77%	88.80%	87.63%	92.92%	87.68%	97.93%	94.03%	90.21%	83.67%	87.50%	88.61%
Proprietary Models												
Qwen-VL-Max	**76.96%**	**64.57%**	50.65%	42.35%	75.79%	54.67%	18.11%	49.59%	**73.72%**	**74.00%**	58.16%	**92.41%**
GPT-4o	75.65%	62.20%	57.14%	38.82%	**76.68%**	53.33%	10.70%	**61.79%**	59.62%	70.00%	55.32%	87.34%
Gemini-1.5-Pro	74.35%	52.36%	**61.69%**	40.00%	67.59%	**56.67%**	14.81%	29.27%	72.44%	73.00%	**62.41%**	86.08%
GPT-4-Turbo	70.43%	55.91%	53.25%	32.35%	76.68%	52.00%	15.23%	58.81%	52.56%	70.00%	58.16%	89.87%
GPT-4o-Mini	70.87%	51.18%	53.90%	**43.53%**	60.08%	49.33%	**22.22%**	47.97%	53.21%	68.00%	40.43%	86.08%
Gemini-1.5-Flash	71.30%	55.91%	59.09%	41.76%	65.61%	47.33%	17.70%	34.15%	65.38%	61.00%	34.75%	84.81%
Claude-3.5-Sonnet	66.96%	52.76%	50.00%	42.35%	67.59%	48.00%	9.47%	49.05%	67.95%	54.00%	43.97%	83.54%
Gemini-1.5-Flash-8B	66.96%	48.43%	54.55%	30.59%	73.12%	40.67%	6.58%	34.69%	41.67%	62.00%	26.24%	82.28%
Open Source Models												
Qwen2.5-VL-72B-Instruct	73.48%	59.84%	47.40%	45.88%	**79.84%**	56.67%	18.93%	**71.27%**	68.59%	72.00%	**62.41%**	**91.14%**
InternVL2-76B	74.35%	**65.75%**	51.95%	44.71%	65.22%	**61.33%**	14.40%	46.61%	74.36%	76.00%	58.87%	87.34%
LLaVA-Video-72B-Qwen2	**74.78%**	62.20%	**58.44%**	**47.06%**	68.38%	53.33%	14.81%	51.76%	67.31%	72.00%	53.90%	65.82%
NVLM-D-72B	73.48%	57.87%	50.65%	34.12%	69.57%	54.67%	12.76%	39.57%	63.46%	**78.00%**	60.28%	79.75%
mPLUG-Owl3	65.22%	54.33%	53.90%	34.12%	63.24%	42.67%	**25.10%**	50.14%	**77.56%**	57.00%	37.59%	82.28%
VILA1.5-40B	67.39%	46.06%	53.25%	38.82%	65.22%	46.00%	7.82%	47.69%	48.08%	63.00%	40.43%	78.48%
Pixtral-12B-2409	65.22%	53.94%	48.05%	38.82%	62.45%	52.67%	9.05%	33.60%	63.46%	52.00%	37.59%	81.01%
deepseek-vl2	65.22%	56.30%	48.70%	34.71%	63.64%	47.33%	5.76%	45.53%	53.21%	59.00%	27.66%	83.54%

Table 2. Selected results of MLLM performances on the **CoreCognition** Dataset. The best results are bolded and the second best underlined.

higher than both Concrete Operation and Sensorimotor.

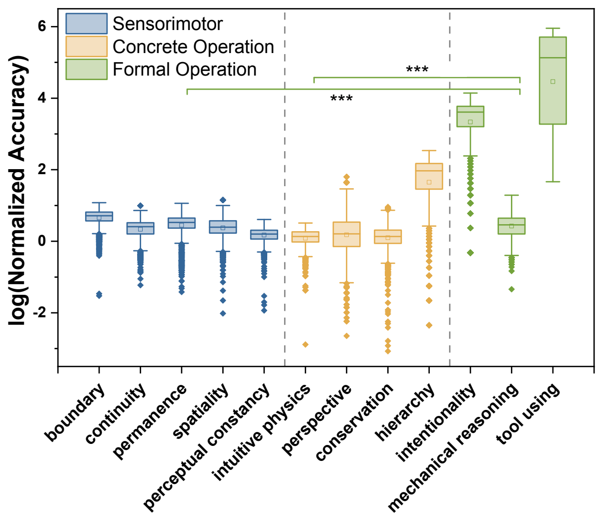


Figure 3. Log-scale accuracy by concept, normalized by chance level. Models performed better on tasks linked to later developmental stages, but struggled with those that emerge earlier in human cognition. The pairwise t-statistics is 22.2633 between Formal Operation and Concrete Operation and 28.7225 between Formal Operation and Sensorimotor.

Since different core knowledge abilities correspond to distinct question types (e.g., numerical, true/false, multiple-choice), they exhibit varying chance-level accuracy and difficulty. Thus, normalization is necessary to enable a fair comparison across these abilities. To further substantiate that MLLMs perform worse on lower-level abilities than on higher-level ones, we first normalized the accuracy of each ability by its chance-level accuracy. Subsequently, a log-scale transformation was applied to the distribution. Fig 3 presents a fair comparison between the accuracy and performance of different abilities where a clear upward trend can be identified as the concepts move from low-level to high-level. This can be concluded as MLLMs perform worse on lower-level abilities than on higher-level ones, or in other words, there exist core knowledge deficits in Multi-Modal Language Models.

4.3. Correlations Between Core Abilities within CoreCognition

To examine the relationships between models' abilities across the cognitive hierarchy, we computed Pearson correlations between performance scores for each assessed ability. Our findings revealed a distinct pattern of divergence $\rho < 0.4$ from the hierarchical structure observed in humans, alongside areas of partial alignment $\rho > 0.7$. As shown in Fig. 4, region ① shows strong correlations emerged within both the Sensorimotor Stage and the Formal Operational Stage abilities, reflecting the interdependence typically found among abilities at the same developmental level in humans. However, ② two Sensorimotor Stage abilities (Permanence and Spatiality) showed weak correlations with most higher-stage abilities, suggesting that these early competencies fail to provide the developmental scaffolding seen in human cognition, indicating potential core knowledge deficits. This pattern is reinforced by ③ the similarly weak cross-stage correlations of three Concrete Operational Stage abilities (Perspective, Conservation, and Intuitive Physics), which in human development serve as critical transitional foundations for higher-order reasoning. The absence of such structure in models highlights a key departure from the developmental trajectory observed in humans.

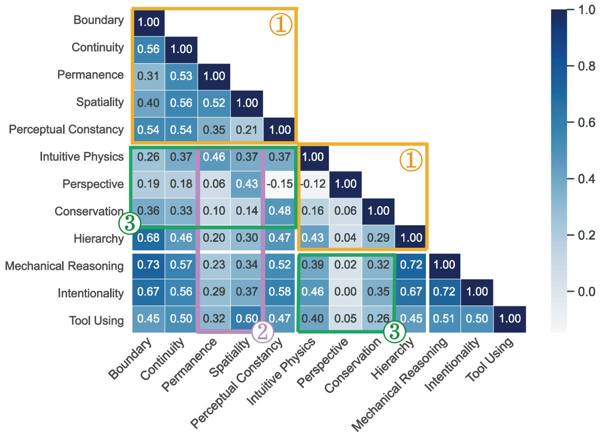


Figure 4. Pearson Correlations Between Cognitive Abilities. Correlations are strongest within developmental stages (Sensorimotor and Formal Operational), but markedly weaker across stages. In particular, lower-stage abilities show weak or absent correlations with higher-stage abilities, suggesting that early capacities fail to provide the developmental foundation observed in human cognition(Piaget & Inhelder, 1974; Spelke et al., 1992).

4.4. Correlations Between CoreCognition and other Benchmarks

We examine whether the 12 core cognitive concepts and three stages assessed in **CoreCognition** provide meaningful support for other benchmarks, such as SEED-Bench (Li et al., 2024) and mmbench (Liu et al., 2024b), as well as for high-level cognitive concepts derived from these benchmarks, such as scene understanding in SEED-Bench. To probe this relationship, we analyze the correlation heatmap. Figure 5 illustrates the correlation between the 12 **CoreCognition** concepts, its three stages, and the overall benchmark and 26 evaluation benchmarks and nine high-level cognitive abilities.

Fig 5 illustrates a highly correlated relation between **CoreCognition** and current MLLMs benchmarks with only one exception of ChartQA. It's noteworthy that SEED-Bench2 highly correlates with our benchmark and most of the core abilities except perspective and conservation. Looking at the right plot of Fig 5, we can observe also a high correlation between the core concepts and high-level abilities. The exception of text understanding is likely because assessments of core cognitive abilities are extremely reliant upon the multi-modal integration of information. Another interpretation is that the ability to do textual understanding is orthogonal to all core cognitive abilities.

4.5. Does Performance increase as Model Scales?

Not for low-level abilities. A fundamental principle in machine learning posits that increasing the scale of a large model, measured by the number of parameters, leads to systematic improvements in its reasoning capabilities (Sutton, 2019; Kaplan et al., 2020). We evaluate the extent to which this principle, commonly referred to as the scaling law, applies to low-level cognitive abilities rooted in core knowledge. As per Fig. 6, our results reveal a clear dissociation between low- and high-level abilities in terms of scaling effects. Specifically, for seven out of the nine abilities in the Sensorimotor and Concrete Operational Stages (two low-level stages), model performance demonstrated little or no improvements with increasing parameters. The two exceptions were hierarchical relation understanding, where model performance improved modestly with model size, and perspective-taking, where, unexpectedly, model performance deteriorated with model size. In contrast, all three high-level abilities within the Formal Operational Stage

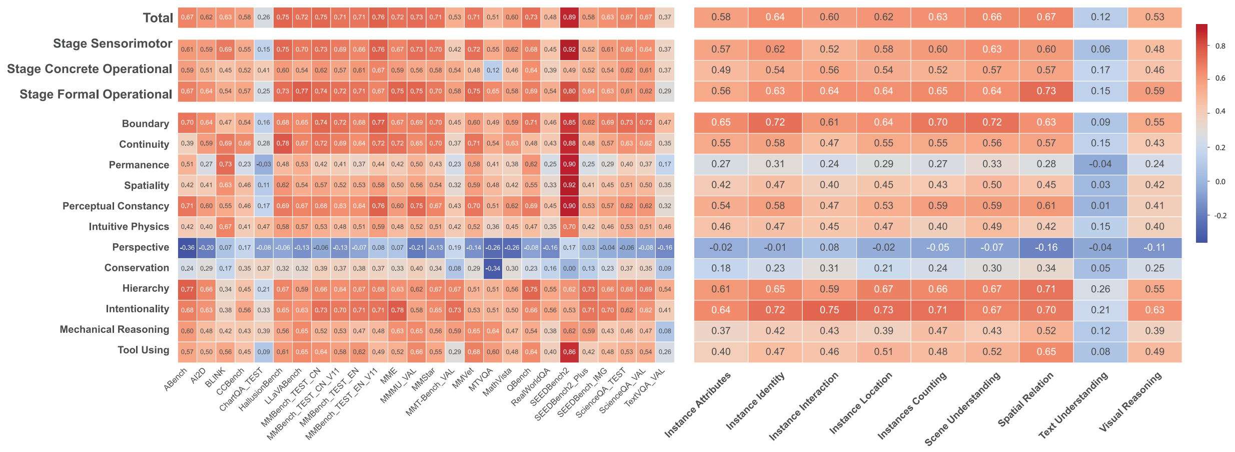


Figure 5. **Left.** Correlation Heatmap between other MLLM benchmarks and core cognitive abilities assessed in our **CoreCognition** benchmark. **Right.** Correlation Heatmap between "high-level" abilities from SEED-Bench and core cognitive abilities assessed in our **CoreCognition** benchmark. (Li et al., 2024).

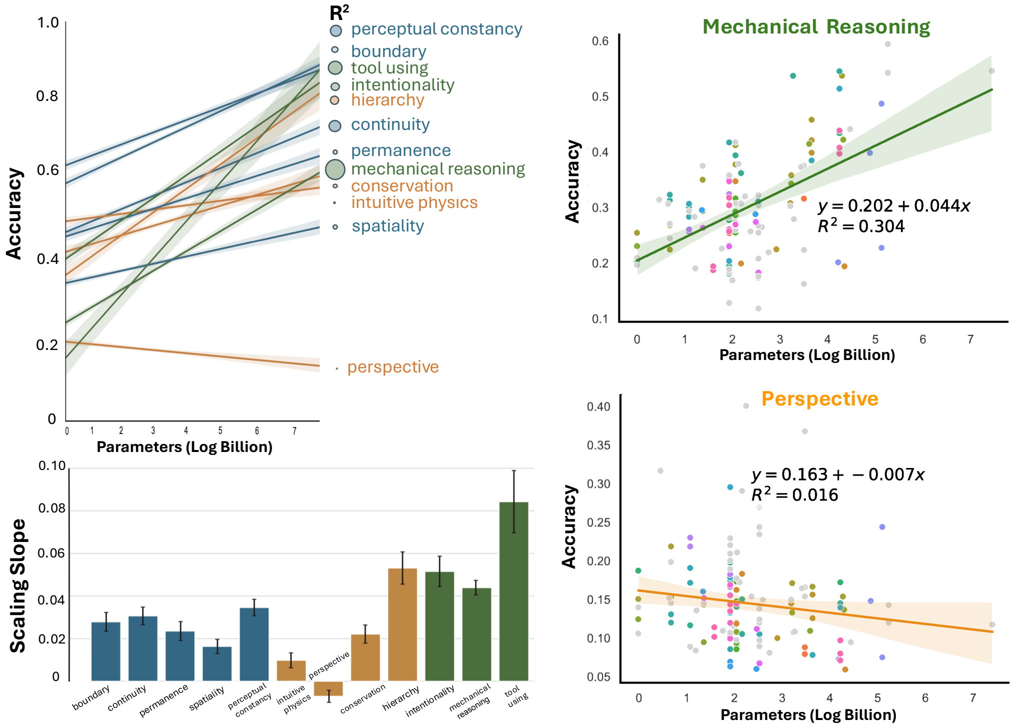


Figure 6. Relationship Between Model Performance on Each Assessed Ability and Model Size. Interestingly, scaling laws do not apply uniformly across all concepts, with some, like perspective-taking, being entirely unscalable. **Top Left.** Fitted curves for each concept using data from 219 models across 11 prompt cases. **Bottom Left.** Box plot showing the slopes of the fitted curves from the top-left plot. **Top Right.** Mechanical reasoning concept, where each dot represents a data point (shown for the empty-string case, i.e., no additional input). Dots of the same color indicate models from the same series, such as the InternVL series. **Bottom Right.** Perspective-taking concept. Dots of the same color indicate the same series.

displayed strong correlations with model size (need some values here), indicating a pronounced scaling effect. These findings indicate that while scaling improves high-level reasoning, its effect on low-level cognitive abilities is minimal and, in some cases, even detrimental. In other words, the cognitive abilities of current MLLMs exhibit varying degrees of “scalability”, with low-level cognitive abilities demonstrating weak or even no scalability. A key implication is that simply increasing model size may not be sufficient for developing core knowledge in MLLMs.

5. Concept Hacking

A key challenge in evaluating the cognitive abilities of language models is their tendency to exploit spurious correlations. In other words, their apparent proficiency in certain tasks may arise from shortcut learning rather than genuine cognitive capabilities (Bender et al., 2021). Extensive research has demonstrated such shortcut reliance in benchmarks designed to assess high-level reasoning in MLLMs. To further investigate whether evaluations of low-level cognitive abilities are similarly vulnerable to shortcut exploitation, we introduce a control experiment to rigorously examine the core knowledge possessed by the MLLMs. At the core of the control experiment lies a novel technique termed *concept-based hacking*. Concept-based 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. 7.

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. In contrast, individuals that rely on statistical correlations from their training data, rather than true conceptual understanding, will 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.7, 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.

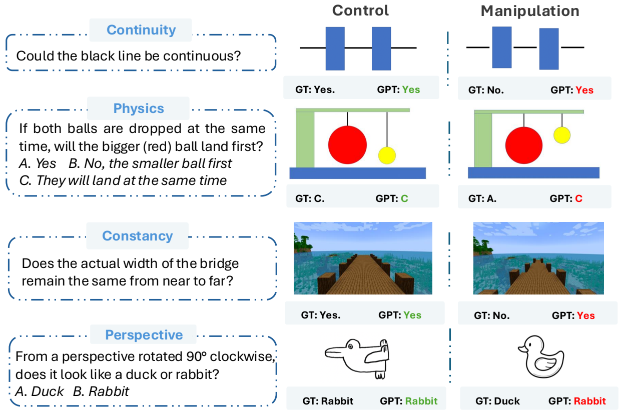

Control
Manipulation
Continuity
Could the black line be continuous?
GT: Yes.
GPT: Yes
GT: No.
GPT: Yes
Physics
If both balls are dropped at the same time, will the bigger (red) ball land first?
A. Yes B. No, the smaller ball first
C. They will land at the same time
GT: C.
GPT: C
GT: A.
GPT: C
Constancy
Does the actual width of the bridge remain the same from near to far?
GT: Yes.
GPT: Yes
GT: No.
GPT: Yes
Perspective
From a perspective rotated 90° clockwise, does it look like a duck or rabbit?
A. Duck B. Rabbit
GT: Rabbit
GPT: Rabbit
GT: Duck
GPT: Rabbit

Figure 7. Example tasks using the Concept Hacking methodology.

5.1. Objectives and Methodology

We applied the concept-based hacking method to core cognitive abilities from the Sensorimotor and Concrete Operational stages, generating 45 task pairs (each consisting of a manipulated and a corresponding standard control task). By comparing model performance on manipulated versus standard conditions, we can systematically detect shortcut-taking and illusion in core knowledge assessments.

5.2. Results: Shortcut-taking and Illusory Understanding

We have probed the models' strategies for answering the assessment of low-level abilities by assessing their performance on manipulation tasks derived from concept-based hacking and their respective controls. The results demonstrated a clear segregation of models relying on shortcut-taking and illusory understanding (Fig. 8). 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. Notably, unlike humans, essentially none of the models demonstrate roughly equal accuracy on both tasks, a sign of immunity to concept-based 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.

Most interestingly, models' susceptibility to concept-based hacking is not necessarily determined by model size or performance on the main benchmark. While models with strong shortcut-taking were mainly small, weak-performing models, the bottom right section also included some of the largest, best-performing models, such as GPT-4o. Similarly and to a larger degree, "core illusion" models in the top-left section can also be found with varied performance and size. In conjunction with the non-scaling tendency of low-level abilities noted in the above sections, the result of the concept-based hacking suggested that the increase of model size does not lead to better grasps of core knowledge, but only better shortcut-taking or illusory understanding.

6. Discussion

Our findings support the hypothesis that MLLMs lack core knowledge and that it cannot be acquired through scale alone. This presents a fundamental challenge to the current architecture of MLLMs as a pathway to human-like general intelligence (Summerfield, 2022).

One might object that human-like core knowledge is not a necessary condition for artificial general intelligence (AGI). After all, intelligence may be multiply realizable—achievable through architectures and developmental trajectories that differ from those of humans (Bechtel & Mundale, 1999). However, core knowledge may reflect foundational cognitive principles that emerge across intelligent agents, including non-human animals (Santos, 2004; Lake et al., 2017). The theory of grounded cognition fur-

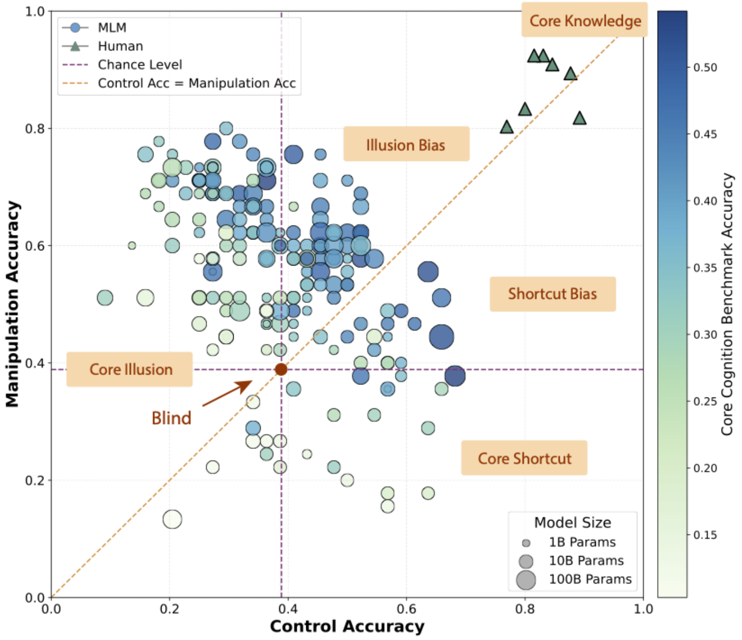


Figure 8. Accuracy results of MLLMs on Control vs. Manipulation in the Concept Hacking Evaluation. Each circle represents a model, with size indicating parameter count and color reflecting overall accuracy on the **CoreCognition** benchmark. Human performance is shown as green triangles. The red dot marks the baseline of chance-level accuracy on both tasks—models near this point are effectively "blind". As model size increases, they tend to exhibit stronger illusion or shortcut biases, rather than moving along the diagonal toward the core knowledge region occupied by humans—indicating a persistent failure to acquire genuine conceptual understanding.

ther supports this idea: it posits that high-level reasoning in real-world contexts depends on embodied interactions with the physical world (Barsalou, 2020; Pezzulo et al., 2013). If such theories hold, then the absence of core knowledge may not just limit task performance—it may fundamentally restrict an agent's ability to act robustly and flexibly in dynamic environments. Given the current lack of consensus on the path to AGI, the human developmental trajectory offers an empirically grounded reference point. Persistent model limitations—such as hallucinations, poor generalization, and brittleness—suggest that essential cognitive ingredients may still be missing. By aligning evaluation with structures known to support robust reasoning and perception in humans, our framework helps to identify and address these gaps in emerging AI systems.

We recognize that using a visual question-answering (VQA) format to probe core knowledge introduces auxiliary demands—particularly those related to language understanding. This interplay between linguistic processing and cognitive evaluation may confound the isolation of core abilities, a concern echoed by multiple reviewers. However, auxiliary task demands are an inherent challenge in evaluating AI models, regardless of the format. To mitigate such confounds, we implement three key strategies: (1) we curate questions to minimize overlap between abilities and exclude items requiring multiple competencies, (2) we manually filter ambiguous prompts and use LLMs to enhance phrasing clarity, and (3) we systematically test alternative prompt formulations to reduce susceptibility to specific wordings. While our approach does not entirely eliminate language dependencies, it offers a tractable and replicable way to evaluate core cognitive abilities in current models.

7. Conclusion

We introduced the **CoreCognition** benchmark paired with a novel concept-based hacking method to evaluate the existence of core knowledge in MLLMs. We found that (1) they systematically perform poorly at simple, low-level cognitive abilities demanding only basic understanding of the world; (2) models' performance on high-level abilities does not correlate with the corresponding low-level abilities that ground them in humans; (3) such abilities exhibit very low scalability among models, meaning that simply raising the number of parameters could not better the models' performance on these abilities; (4) instead of core knowledge, models are biased by illusory understanding and shortcut reliance when solving low-level tasks. Taken together, our results suggest that current MLLMs exhibit core knowledge deficits—they lack a fundamental understanding of key domains such as objects, actions, numbers, space, and social relations, which humans possess from infancy.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning through a cognitively grounded evaluation of multi-modal language models (MLLMs). We provide insights into the differential emergence of low-level and high-level cognitive abilities in current models compared to human's, highlighting important limitations in the core reasoning capabilities of MLLMs and caution against overinterpreting their success on complex tasks. This work may inform the development of more robust and interpretable models, and also opens avenues for interdisciplinary dialogue between AI and cognitive science. We do not foresee any immediate societal risks arising from this research.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Agrawal, P., Antoniak, S., Hanna, E. B., Bout, B., Chaplot, D., Chudnovsky, J., Costa, D., Monicault, B. D., Garg, S., Gervet, T., Ghosh, S., Héliou, A., Jacob, P., Jiang, A. Q., Khandelwal, K., Lacroix, T., Lample, G., Casas, D. L., Lavril, T., Scao, T. L., Lo, A., Marshall, W., Martin, L., Mensch, A., Muddireddy, P., Nemychnikova, V., Pellat, M., Platen, P. V., Raghuraman, N., Rozière, B., Sablayrolles, A., Saulnier, L., Sauvestre, R., Shang, W., Soletskyi, R., Stewart, L., Stock, P., Studnia, J., Subramanian, S., Vaze, S., Wang, T., and Yang, S. Pixtral 12b, 2024. URL https://arxiv.org/abs/2410.07073.

Allen, K. R., Smith, K. A., and Tenenbaum, J. B. Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proceedings of the National Academy of Sciences*, 117(47):29302–29310, 2020.

Andrews, G. and Halford, G. S. 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.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., and Shou, M. Z. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.

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

Baillargeon, R. and Carey, S. Core cognition and beyond: The acquisition of physical and numerical knowledge. *Early childhood development and later outcome*, 1, 2012.

Baillargeon, R., Spelke, E. S., and Wasserman, S. Object permanence in five-month-old infants. *Cognition*, 20(3): 191–208, 1985.

Baker, C. L., Saxe, R., and Tenenbaum, J. B. Action understanding as inverse planning. *Cognition*, 113(3):329–349, 2009.

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

Barrouillet, P. Theories of cognitive development: From piaget to today, 2015.

Barsalou, L. W. Challenges and opportunities for grounding cognition. *Journal of Cognition*, 3(1), 2020.

Bear, D. M., Wang, E., Mrowca, D., Binder, F. J., Tung, H.-Y. F., Pramod, R., Holdaway, C., Tao, S., Smith, K., Sun, F.-Y., et al. Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*, 2021.

Bechtel, W. and Mundale, J. Multiple realizability revisited: Linking cognitive and neural states. *Philosophy of science*, 66(2):175–207, 1999.

Bell, M. A. and Adams, S. E. 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.

Bellmund, J. L., Gärdenfors, P., Moser, E. I., and Doeller, C. F. Navigating cognition: Spatial codes for human thinking. *Science*, 362(6415):eaat6766, 2018.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. 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.

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

Bertenthal, B. I., Gredebäck, G., and Boyer, T. W. Differential contributions of development and learning to infants' knowledge of object continuity and discontinuity. *Child Development*, 84(2):413–421, 2013.

Borst, G., Poirel, N., Pineau, A., Cassotti, M., and Houdé, O. 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.

Bremner, J. G., Slater, A. M., and Johnson, S. P. Perception of object persistence: The origins of object permanence in infancy. *Child Development Perspectives*, 9(1):7–13, 2015.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Byrne, R. M. Counterfactual thought. *Annual review of psychology*, 67(1):135–157, 2016.

Carey, S., Zaitchik, D., and Bascandziev, I. Theories of development: In dialog with jean piaget. *Developmental Review*, 38:36–54, 2015.

Caviola, L., Schubert, S., and Greene, J. D. The psychology of (in) effective altruism. *Trends in Cognitive Sciences*, 25(7):596–607, 2021.

Cesana-Arlotti, N., Martín, A., Téglás, E., Vorobyova, L., Cetnarski, R., and Bonatti, L. L. Precursors of logical reasoning in preverbal human infants. *Science*, 359(6381): 1263–1266, 2018.

Chalmers, D. J. Syntactic transformations on distributed representations. *Connectionist Natural Language Processing: Readings from Connection Science*, pp. 46–55, 1992.

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

Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., 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, 2024.

Chia, Y. K., Han, V. T. Y., Ghosal, D., Bing, L., and Poria, S. Puzzlevqa: Diagnosing multimodal reasoning challenges of language models with abstract visual patterns. *arXiv preprint arXiv:2403.13315*, 2024.

Church, R. B. and Goldin-Meadow, S. The mismatch between gesture and speech as an index of transitional knowledge. *Cognition*, 23:43–71, 1986.

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

Dai, W., Lee, N., Wang, B., Yang, Z., Liu, Z., Barker, J., Rintamaki, T., Shoeybi, M., Catanzaro, B., and Ping, W. Nvlm: Open frontier-class multimodal llms, 2024. URL https://arxiv.org/abs/2409.11402.

De Waal, F. B. and Preston, S. D. Mammalian empathy: behavioural manifestations and neural basis. *Nature Reviews Neuroscience*, 18(8):498–509, 2017.

Dong, H., Liu, M., Zhou, K., Chatzi, E., Kannala, J., Stachniss, C., and Fink, O. Advances in multimodal adaptation and generalization: From traditional approaches to foundation models. *arXiv preprint arXiv:2501.18592*, 2025.

Feigenson, L., Dehaene, S., and Spelke, E. Core systems of number. *Trends in cognitive sciences*, 8(7):307–314, 2004.

Fodor, J. A. *The Language of Thought*. MIT Press, 1975.

Fodor, J. A. *LOT 2: The language of thought revisited*. Oxford University Press, 2008.

Fowler, H. N., Lamb, W. R. M., et al. *Plato*, volume 5. W. Heinemann, 1914.

Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., and Ji, R. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv: 2306.13394*, 2023.

Gandhi, K., Stojnic, G., Lake, B. M., and Dillon, M. R. Baby intuitions benchmark (bib): Discerning the goals, preferences, and actions of others. *Advances in neural information processing systems*, 34:9963–9976, 2021.

Gemini. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv: 2312.11805*, 2023.

Gibson, J. J. *The Ecological Approach to Visual Perception. (1st ed.)*. Psychology Press, 1979.

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

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

Halford, G. S., Wilson, W. H., and Phillips, S. Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and brain sciences*, 21(6):803–831, 1998.

Hannagan, T., Amedi, A., Cohen, L., Dehaene-Lambertz, G., and Dehaene, S. Origins of the specialization for letters and numbers in ventral occipitotemporal cortex. *Trends in cognitive sciences*, 19(7):374–382, 2015.

Hegarty, M. Mechanical reasoning by mental simulation. *Trends in cognitive sciences*, 8(6):280–285, 2004.

Hermer, L. and Spelke, E. Modularity and development: The case of spatial reorientation. *Cognition*, 61(3):195–232, 1996.

Houdé, O. Numerical development: From the infant to the child. *Cognitive Development*, 12(3):373–391, 1997.

Houdé, O., Pineau, A., Leroux, G., Poirel, N., Perchey, G., Lanoë, C., Lubin, A., Turbelin, M.-R., Rossi, S., Simon, G., Delcroix, N., Lamberton, F., Vigneau, M., Wisniewski, G., Vicet, J.-R., and Mazoyer, B. 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.

Huang, J. and Zhang, J. A survey on evaluation of multimodal large language models, 2024. URL https://arxiv.org/abs/2408.15769.

Huitt, W. and Hummel, J. Piaget's theory of cognitive development. *Educational psychology interactive*, 3(2):1–5, 2003.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Iacoboni, M. Imitation, empathy, and mirror neurons. *Annual review of psychology*, 60(1):653–670, 2009.

Inhelder, B. and Piaget, J. *The Growth of Logical Thinking from Childhood to Adolescence*. Basic Books, 1958.

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

Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Janet, P. Mental pathology. *Psychological Review*, 12(2-3):98, 1905.

Jara-Ettinger, J., Schulz, L. E., and Tenenbaum, J. B. The naive utility calculus as a unified, quantitative framework for action understanding. *Cognitive Psychology*, 123:101334, 2020.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Jiang, Y., Zhang, J., Sun, K., Sourati, Z., Ahrabian, K., Ma, K., Ilievski, F., and Pujara, J. Marvel: Multidimensional abstraction and reasoning through visual evaluation and learning. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 46567–46592. Curran Associates, Inc., 2024.

Jowett, B. et al. *The Republic of Plato*. Clarendon press, 1888.

Kant, I. Critique of pure reason. 1781. *Modern Classical Philosophers, Cambridge, MA: Houghton Mifflin*, pp. 370–456, 1781.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., and DiCarlo, J. J. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature neuroscience*, 22(6):974–983, 2019.

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

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

Kilner, J. M. More than one pathway to action understanding. *Trends in cognitive sciences*, 15(8):352–357, 2011.

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

Kirkpatrick, E. The part played by consciousness in mental operations. *The Journal of Philosophy, Psychology and Scientific Methods*, 5(16):421–429, 1908.

Kuhn, D. and Angelev, J. An experimental study of the development of formal operational thought. *Child Development*, pp. 697–706, 1976.

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

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

Li, B., Ge, Y., Ge, Y., Wang, G., Wang, R., Zhang, R., and Shan, Y. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13299–13308, 2024.

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

Li, C., Xu, H., Tian, J., Wang, W., Yan, M., Bi, B., Ye, J., Chen, H., Xu, G., Cao, Z., Zhang, J., Huang, S., Huang, F., Zhou, J., and Si, L. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv: 2205.12005*, 2022.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023a.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *CONFERENCE*, 2023b.

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., and Wen, J.-R. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023c.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Lin, J., Yin, H., Ping, W., Lu, Y., Molchanov, P., Tao, A., Mao, H., Kautz, J., Shoeybi, M., and Han, S. Vila: On pre-training for visual language models, 2023.

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

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024a.

Liu, S., Ullman, T. D., Tenenbaum, J. B., and Spelke, E. S. Ten-month-old infants infer the value of goals from the costs of actions. *Science*, 358(6366):1038–1041, 2017.

Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023a.

Liu, Y., Li, Z., Yang, B., Li, C., Yin, X., Liu, C.-l., Jin, L., and Bai, X. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023b.

Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024b.

Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.-W., Galley, M., and Gao, J. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

Marino, K., Rastegari, M., Farhadi, A., and Mottaghi, R. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pp. 3195–3204, 2019.

Marwaha, S., Goswami, M., and Vashist, B. 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.

Masry, A., Long, D. X., Tan, J. Q., Joty, S., and Hoque, E. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177/.

Meltzoff, A. N. Origins of theory of mind, cognition and communication. *Journal of communication disorders*, 32 (4):251–269, 1999.

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

Miller, P. H. *Theories of developmental psychology (6th ed.).* Macmillan Higher Education, 2016.

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

Mitchell, M. Why ai is harder than we think. *arXiv preprint arXiv:2104.12871*, 2021.

Mitra, C., Huang, B., Darrell, T., and Herzig, R. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14420–14431, 2024.

Moll, H. and Meltzoff, A. N. Perspective-taking and its foundation in joint attention. *Perception, causation, and objectivity. Issues in philosophy and psychology*, pp. 286–304, 2011.

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

Murphy, G. L. and Lassaline, M. E. Hierarchical structure in concepts and the basic level of categorization. In *Knowledge Concepts and Categories*, pp. 93–131. Psychology Press, 2013.

Newcombe, N. S. and Sluzenski, J. Starting points and change in early spatial development. *Human Spatial Memory*, pp. 25–40, 2004.

Ninomiya, T., Noritake, A., Kobayashi, K., and Isoda, M. A causal role for frontal cortico-cortical coordination in social action monitoring. *Nature communications*, 11(1): 5233, 2020.

Oh, C., Fang, Z., Im, S., Du, X., and Li, Y. Understanding multimodal llms under distribution shifts: An information-theoretic approach. *arXiv preprint arXiv:2502.00577*, 2025.

O'Brien, T. C. and Shapiro, B. J. The development of logical thinking in children. *American Educational Research Journal*, 5(4):531–542, 1968.

Paiss, R., Ephrat, A., Tov, O., Zada, S., Mosseri, I., Irani, M., and Dekel, T. Teaching clip to count to ten. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3170–3180, 2023.

Peng, B., Li, C., He, P., Galley, M., and Gao, J. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.

Pezzulo, G., Barsalou, L. W., Cangelosi, A., Fischer, M. H., McRae, K., and Spivey, M. J. Computational grounded cognition: a new alliance between grounded cognition and computational modeling. *Frontiers in psychology*, 3: 612, 2013.

Piaget, J. *The Psychology of Intelligence*. Harcourt, Brace, 1950.

Piaget, J. *The Origins of Intelligence in Children*. International Universities Press, 1952.

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

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

Piaget, J. and Inhelder, B. Intellectual operations and their development. In *Experimental Psychology Its Scope and Method: Volume VII (Psychology Revivals)*, pp. 144–205. Psychology Press, 2014.

Piloto, L. S., Weinstein, A., Battaglia, P., and Botvinick, M. Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature human behaviour*, 6(9):1257–1267, 2022.

Plato, Spens, H., and Adams, J. *The Republic of Plato*. Printed by Robert and Andrew Foulis, Glasgow, 1763.

Poirel, N., Borst, G., Simon, G., Rossi, S., Cassotti, M., Pineau, A., and Houdé, O. Number conservation is related to children's prefrontal inhibitory control: an fmri study of a piagetian task. *PloS one*, 7(7):e40802, 2012.

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

Pylyshyn, Z. W. Computation and cognition: Issues in the foundations of cognitive science. *Behavioral and Brain sciences*, 3(1):111–132, 1980.

Rahmanzadehgervi, P., Bolton, L., Taesiri, M. R., and Nguyen, A. T. Vision language models are blind. *arXiv preprint arXiv:2407.06581*, 2024.

Rochat, P. The evolution of developmental theories since piaget: A metaview. *Perspectives on Psychological Science*, 19(6):921–930, 2024.

Rosenthal, D. M. *The Nature of Mind*. Oxford University Press, New York, 1991.

Russell, B. *History of western philosophy*. Routledge, 1946.

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

Sahin, U., Li, H., Khan, Q., Cremers, D., and Tresp, V. Enhancing multimodal compositional reasoning of visual language models with generative negative mining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5563–5573, 2024.

Santos, L. R. ‘core knowledges’: A dissociation between spatiotemporal knowledge and contact-mechanics in a non-human primate?, 2004.

Scott, R. M. and Baillargeon, R. Early false-belief understanding. *Trends in cognitive sciences*, 21(4):237–249, 2017.

Searle, J. R. The intentionality of intention and action. *Inquiry*, 22(1-4):253–280, 1979.

Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., and Li, H. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. *arXiv preprint arXiv:2403.16999*, 2024.

Shayer, M. Has piaget’s construct of formal operational thinking any utility? *British Journal of Educational Psychology*, 49(3):265–276, 1979.

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

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

Siegal, M. and Varley, R. Neural systems involved in’theory of mind’. *Nature Reviews Neuroscience*, 3(6):463–471, 2002.

Song, W., Li, Y., Xu, J., Wu, G., Ming, L., Yi, K., Luo, W., Li, H., Du, Y., Guo, F., and Yu, K. M3gia: A cognition inspired multilingual and multimodal general intelligence ability benchmark, 2024. URL https://arxiv.org/abs/2406.05343.

Spelke, E. *What babies know: Core Knowledge and Composition volume 1*, volume 1. Oxford University Press, 2022.

Spelke, E. S. Core knowledge, language, and number. *Language Learning and Development*, 13(2):147–170, 2017.

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

Spelke, E. S., Breinlinger, K., Macomber, J., and Jacobson, K. Origins of knowledge. *Psychological review*, 99(4): 605, 1992.

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

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

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

Sun, Q., Cui, Y., Zhang, X., Zhang, F., Yu, Q., Luo, Z., Wang, Y., Rao, Y., Liu, J., Huang, T., and Wang, X. Generative multimodal models are in-context learners. *Computer Vision and Pattern Recognition*, 2023. doi: 10.1109/CVPR52733.2024.01365.

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

Tang, K., Gao, J., Zeng, Y., Duan, H., Sun, Y., Xing, Z., Liu, W., Lyu, K., and Chen, K. Lego-puzzles: How good are mllms at multi-step spatial reasoning? *arXiv preprint arXiv:2503.19990*, 2025a.

Tang, Y., Liu, P., Feng, M., Tan, Z., Mao, R., Huang, C., Bi, J., Xiao, Y., Liang, S., Hua, H., et al. Mmperspective: Do mllms understand perspective? a comprehensive benchmark for perspective perception, reasoning, and robustness. *arXiv preprint arXiv:2505.20426*, 2025b.

Team, C. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv: 2405.09818*, 2024.

Team, Q. Qwen2.5-vl, January 2025. URL https://qwenlm.github.io/blog/qwen2.5-vl/.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Ullman, T. D. and Tenenbaum, J. B. Bayesian models of conceptual development: Learning as building models of the world. *Annual Review of Developmental Psychology*, 2(1):533–558, 2020.

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

Viarouge, A., Houdé, O., and Borst, G. 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.

Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.

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

Wellman, H. M., Cross, D., and Watson, J. Meta-analysis of theory-of-mind development: The truth about false belief. *Child development*, 72(3):655–684, 2001.

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

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

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

Wu, P. and Xie, S. 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.

Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., Xie, Z., Wu, Y., Hu, K., Wang, J., Sun, Y., Li, Y., Piao, Y., Guan, K., Liu, A., Xie, X., You, Y., Dong, K., Yu, X., Zhang, H., Zhao, L., Wang, Y., and Ruan, C. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024. URL https://arxiv.org/abs/2412.10302.

Xu, G., Jin, P., Hao, L., Song, Y., Sun, L., and Yuan, L. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024.

Xu, P., Shao, W., Zhang, K., Gao, P., Liu, S., Lei, M., Meng, F., Huang, S., Qiao, Y., and Luo, P. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):1877–1893, 2025. doi: 10.1109/TPAMI.2024.3507000.

Yang, D. Y.-J., Rosenblau, G., Keifer, C., and Pelphrey, K. A. An integrative neural model of social perception, action observation, and theory of mind. *Neuroscience & Biobehavioral Reviews*, 51:263–275, 2015.

Yang, S., Luo, S., and Han, S. C. Multimodal commonsense knowledge distillation for visual question answering (student abstract). In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 29545–29547, 2025a.

Yang, S., Luo, S., Han, S. C., and Hovy, E. Magic-vqa: Multimodal and grounded inference with commonsense knowledge for visual question answering. *arXiv preprint arXiv:2503.18491*, 2025b.

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

Ye, J., Xu, H., Liu, H., Hu, A., Yan, M., Qian, Q., Zhang, J., Huang, F., and Zhou, J. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models, 2024. URL https://arxiv.org/abs/2408.04840.

Yi, D.-J., Turk-Browne, N. B., Flombaum, J. I., Kim, M.-S., Scholl, B. J., and Chun, M. M. Spatiotemporal object continuity in human ventral visual cortex. *Proceedings of the National Academy of Sciences*, 105(26):8840–8845, 2008.

Yin, Z., Wang, J., Cao, J., Shi, Z., Liu, D., Li, M., Huang, X., Wang, Z., Sheng, L., Bai, L., et al. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *Advances in Neural Information Processing Systems*, 36, 2024.

Ying, K., Meng, F., Wang, J., Li, Z., Lin, H., Yang, Y., Zhang, H., Zhang, W., Lin, Y., Liu, S., jiayi lei, Lu, Q., Gao, P., Chen, R., Xu, P., Zhang, R., Zhang, H., Wang, Y., Qiao, Y., Luo, P., Zhang, K., and Shao, W. MMT-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask AGI. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=R4Ng8zYaiz.

Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., and Zou, J. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022.

Zhang, H., Li, C., Wu, W., Mao, S., Vulić, I., Zhang, Z., Wang, L., Tan, T., Wei, F., et al. A call for new recipes to enhance spatial reasoning in mllms. *arXiv preprint arXiv:2504.15037*, 2025.

Zhang, X., Li, J., Chu, W., Hai, J., Xu, R., Yang, Y., Guan, S., Xu, J., and Cui, P. On the out-of-distribution generalization of multimodal large language models. *arXiv preprint arXiv:2402.06599*, 2024a.

Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., and Li, C. Video instruction tuning with synthetic data, 2024b. URL https://arxiv.org/abs/2410.02713.

Zhao, Y., Pang, T., Du, C., Yang, X., Li, C., Cheung, N.-M. M., and Lin, M. On evaluating adversarial robustness of large vision-language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

A. Cognitive Science Framework

A.1. Core Knowledge and Human Cognitive Development

Past research has shown that humans exhibit a series of rudimentary yet robust abilities in domains such as object, number, space, action, and social cognition at a very young age. Such abilities, often known as "core" cognition, ground the set of diverse and complex abilities of human intelligence that develop later (Spelke et al., 1992; 1994; 1995; Spelke & Kinzler, 2007; Baillargeon & Carey, 2012; Mitchell, 2020; 2021). From infancy to early adulthood, human cognition develops along a structured trajectory, with interdependent relations between early, simple abilities and late, complex abilities. For instance, the ability to imagine the perspectives of others typically develops between the ages of 3 and 6 (Piaget & Inhelder, 1969), while the capacity to fully comprehend others' intentions matures around age 12 (Wimmer & Perner, 1983; Wellman et al., 2001; Liu et al., 2008). At the same time, the ability to understand other people's intentions largely depends on the ability to understand other people's perspectives (Iacoboni, 2009; De Waal & Preston, 2017; Liu et al., 2017; Caviola et al., 2021; Ninomiya et al., 2020). An influential account of human learning has suggested that cognitive development is fundamentally driven by the increase of computational/representational power of the system, which allows for more complex mental operations to be performed on external data (Fodor, 1975; Pylyshyn, 1980; Halford et al., 1998; Fodor, 2008). However, while high-level abilities might emerge directly due to enhanced operational resources, these operations are critically guided by the "core" cognition system that has enabled the system to possess a rudimentary understanding of each cognitive domain. This early-stage grounding not only empowers humans to achieve a reliable performance at basic yet widely-applicable tasks starting from very young ages but is also precisely what supports high-level abilities to robustly direct task-relevant behaviors despite the nuanced signals exist in the environment (Mitchell, 2021).

A.2. Piaget's Theory of Human Cognitive Development

The sensorimotor stage is the first stage of cognitive development proposed by Jean Piaget (Piaget, 1952; Piaget & Inhelder, 1974). Spanning from birth to approximately 2 years of age, this stage is characterized by infants' understanding of the world through their sensory experiences and motor actions. Several prominent features of human intelligence developed during this period. First, infants develop object permanence, that they realize objects and people continue to exist even when not in direct sight, or being heard or touched (Baillargeon et al., 1985). They start to understand that there is a sense of continuity for the ways that objects exist, and the inductive bias of continuity is essential, e.g., for recognizing objects when occluded or for continuously tracking objects (Spelke et al., 1995; Le Poidevin, 2000). Infants also develop the sense of boundary during this stage, namely, the ability to recognize where one object ends and another begins (Kestenbaum et al., 1987; Jackendoff, 1991). Lastly, infants develop spatial and perceptual constancy by the end of the sensorimotor stage. Spatiality refers to the ability to perceive the position and distance of objects relative to oneself and each other, and recognize the spatial invariance between them when presented by various sensory experiences (Hermer & Spelke, 1996; Bell & Adams, 1999).

The preoperational and concrete operational stages are the second and third stages of Piaget's cognitive development. Typically spanning over 2 to 7 years of age,the preoperational stage is the transitional stage to the concrete operational stage, which children enter around 7 years of age. During this period, children begin to develop internalized mental actions supported by organized structures that can be manipulated and reversed in systematic ways, known as mental operations (Janet, 1905; Kirkpatrick, 1908; Piaget, 1950; Piaget & Inhelder, 2014; Miller, 2016). Through mental operations, children are then able to rigidly perform tasks that are previously unreachable, such as thinking from other people's perspectives, understanding hierarchical relations of objects, and reasoning about physical events in the world. These tasks require not only rudimentary understandings of physical concepts, which gradually became in place during the preoperational stage but also relational and transformational reasoning that can only be done through mental operations (Piaget & Inhelder, 1974; Church & Goldin-Meadow, 1986; Houdé, 1997). Since the preoperational stage is mostly meaningful as the transitional period preceding the concrete operational stage, we do not have evaluation dimensions specifically targeting the stage. However, tasks targeting the concrete operational stage could assess the existence of knowledge associated with the preoperational stage, such as the law of conservation (Piaget, 1952; Halford, 2011; Houdé, 1997).

The formal operational stage is the fourth and final stage in Piaget's theory of cognitive development, typically emerging around 11 or 12 years of age and continuing into adulthood (Inhelder & Piaget, 1958). Starting in this stage, one is able to systematically and flexibly apply mental operations to not only concrete, physical domains but also abstract, formal domains (Kuhn & Angelev, 1976; Shayer, 1979; Huitt & Hummel, 2003). Foremost, this stage is characterized by the development of complex thinking and reasoning abilities, such as abstraction, pattern recognition, the employment of logic, and hypothetical and counterfactual reasoning (Piaget, 1950; Inhelder & Piaget, 1958). These cognitive advancements pave the way for more sophisticated abilities to interact with the physical world, marked by mechanical reasoning and tool use (O'Brien & Shapiro, 1968). Together, there is the advancement in social cognition, characterized by a deeper understanding of intentions, actions, and the reasoning behind them (Meltzoff, 1999).

A.3. Assessed Cognitive Abilities in CoreCognition and ConceptHack Datasets

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 refers to the cognitive understanding of the topological properties of our physical world (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). This concept is essential for one's systematic and logical organization of conceptual knowledge. 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).

Intentionality Understanding Intention understanding involves recognizing and interpreting the actions of others (Searle, 1979; Rosenthal, 1991). This process is not just about observing a behavior but also about understanding the goal behind it (Baker et al., 2009; Gandhi et al., 2021). For example, seeing someone reaching for a cup is not just about recognizing the physical action but understanding the intention behind it (e.g., they want to drink).

Mechanical Reasoning Mechanical reasoning refers to the ability to understand and apply mechanical concepts and logical principles to solve problems (Allen et al., 2020). This cognitive concept first involves the ability to interpret and predict the behaviors of complex physical systems and understand how different mechanisms of the systems work. Second, mechanical reasoning requires the ability to apply logic rules, such as induction, abduction, syllogism (O'Brien & Shapiro, 1968; Cesana-Arlotti et al., 2018), and reasoning forms, such as hypotheticals and counterfactual (Byrne, 2016), to figure out how to manipulate these systems to achieve a desired outcome (Hegarty, 2004).

Tool Using Tool-using refers to the ability to utilize objects (as tools) in their environment as aids in achieving a specific goal, such as obtaining food or modifying the surroundings. A lot of cognitive components are involved in tool-using ability, such as affordances, referring to computing the action possibilities offered to the agent by the tool with reference to the agent's sensorimotor capabilities (Gibson, 1979). For example, a door handle affords pulling or pushing, as how the door should be operated by a human agent.

B. Input Types and Formats

To cater to the capabilities of different models and test requirements of different core knowledge, diverse input formats are created in the **CoreCognition** benchmark.

First of all, we introduce types of questions in the **CoreCognition** benchmark. Overall, we have three types, numerical (options from 0 to 9), true/false questions, and multi-choice questions (MCQ). Table 3 is the statistics of the distribution of question types.

	MC	TF	NU
#	1045	1042	212

Table 3. Statistics of the distribution of question types.

The input to the model consists of two main components: text prompts and media information.

The text prompt provided to the model includes three elements: question, hint, and media placeholder. The question component provides background information and the task objective, ensuring that the model has all the necessary information to complete the task. The hint includes carefully designed prompts aimed at improving the model's performance. These hints vary in complexity, ranging from basic instructions specifying the response format to chain-of-thought prompts and elaborate in-context learning examples. A study of hint is provided in Appendix F. The media placeholder is embedded within the text prompt to reference media files, enabling the model to associate the textual descriptions with the corresponding media input. Depending on the task requirements, the media placeholder may appear in the question or within the provided answer choices. For example: *Question: Please count the items in the image and answer: Are there more dogs or more animals in the image <image-placeholder: h0001.png>?*. It is important to note that the format of the placeholder is not fixed; instead, it follows the conventions used during the model's training to optimize inference performance (Liu et al., 2023a).

Table 4. Statistics of CoreCognition dataset.

Statistic	**Number**
single-frame	1677
multi-frame	622
* multiple images	200
* single video	401
* multiple videos	21
total	2299

For the media information component, different types of media inputs are provided to accommodate the input constraints of models. The media inputs fall into four categories: single image, multiple images, single video, and multiple images (dataset statistics shown in Table 4). Video files include formats such as MP4, MOV, and GIF, while image files include JPG and PNG. The multi-image format is used as an alternative to video for models that do not support direct video input, in which case a video is split into multiple frames and provided as a sequence of images. Depending on the model's specific input requirements, we supply the media as file paths, preloaded media files, or preprocessed media data.

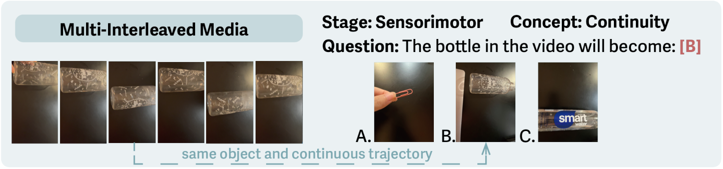

Multi-Interleaved Media
Stage: Sensorimotor
Concept: Continuity
Question: The bottle in the video will become: [B]
A.
B.
C.
smart
same object and continuous trajectory

Figure 9. A video-image interleaved example of multi-frame questions. To correctly infer the answer, model needs to understand the question by mapping each image (co-reference) to its option letter, to understand correlation between frames (temporal understanding) and to infer the possible trajectory of the bottle (reasoning).

C. Curation of CoreCognition Dataset

To select experiments from the developmental psychology literature for evaluating each cognitive ability, journal articles and conference proceedings reporting methodological and empirical studies on these abilities were systematically reviewed. The selection process involved curating a literature bank that includes studies demonstrating robust experimental paradigms. Four researchers with backgrounds in cognitive science and computer science held weekly meetings to deliberate on the suitability of various tasks, ensuring alignment with the assessed constructs, empirical credibility, and adaptability for benchmarking purposes. Upon selection of an appropriate task, a detailed protocol was developed, including operational procedures, materials and toolkits for experimental setup construction, and variability in task conditions. Each researcher was assigned approximately three cognitive abilities to propose experimental protocols, which were subsequently reviewed and approved by two supervising researchers. The following sections provide detailed descriptions of the protocol components.

C.1. Procedure for Experimental Paradigm Operationalization for the CoreCognition Dataset

Each included experiment is formally described and specified in terms of how the assessed abilities are operationally defined by relevant conditions in the experimental setup. Based on these operational definitions, the experimental designs from the original literature were adapted to utilize commonly available objects, online materials, or digital modeling using software toolkits. For each ability, 5-10 classical experimental paradigms were selected, each representing a distinctive operationalization of the assessed construct.

C.2. Materials and Toolkits for Setup Construction

The materials and toolkits for constructing the proposed adaptations were selected based on their similarity to those commonly used in laboratory assessments of respective cognitive abilities. This section details the selection criteria and adaptations made for different developmental stages. The structured adaptation of cognitive experiments using daily objects, online visual data, as well as software simulations ensures alignment with developmental psychology findings while facilitating effective benchmarking.

Abilities in the Sensorimotor Stage are typically tested using objects that children frequently encounter in their daily environments, reflecting their extensive reliance on embodied interactions with the physical world. For example, following established experimental setups in developmental psychology laboratories, benchmark tasks for key sensorimotor abilities—boundary, continuity, object permanence, spatiality, and perceptual constancy—were designed using tangible objects such as pillows, strings, cups, and sheets. One of our assessments of spatiality employed the adaptation of classic cognitive tasks such as the Visual Cliff, in which grid-patterned sheets are layered across transparent boxes to create depth that demands spatial intuition to identify. One assessment of boundary is using occlusion paradigms where daily objects of similar colors partially overlap with each other. One examination of continuity is the construction of moving objects behind blockers and testing reactions to unexpected discontinuities, which are recorded into videos featuring manual manipulation of toys and daily objects. One evaluation of object permanence is through tasks where objects are placed under a cup or cloth and the subject must anticipate their continued existence. One method to assess perceptual constancy was to make use of images of the same object under different visual presentations varied by size, shape, brightness, and color filters.

Abilities in the Concrete Operational Stage are usually tested in children who have yet to develop complex abstract thinking but can reason about physical and logical relationships within structured experimental setups. Benchmarking models for this stage involve problem trials that simulate real-life and digital constructions. For example, for assessments of perspective-taking and conservation, real-life objects such as elastic cans, coins, straws, and play doughs were typically employed to reconstruct the setup of classic Piagetian tasks using single- and multi-frame formats and are transcribed into images to accompany the questions (Piaget, 1950; Spelke et al., 1992). For assessments of intuitive physics tasks, a physics engine toolkit such as Physion was typically used to create digital simulations of hypothetical scenarios that comply to real-life physical laws, a common approach used in laboratory studies of intuitive physics (Bear et al., 2021). The rendered simulations for each scenario were subsequently transcribed via screenshots and screen recordings. We also set up the experimental scenes in real world to curate realistic data. One method for us to collect hierarchical relation understanding experimental set-ups was to leverage online data featuring images of different classes of daily items, such as chairs, cars, and animals were typically used to create questions demanding the distinguishment of inclusive relations based on class hierarchy, a common problem type used in laboratory studies of hierarchy understanding (Chapman & McBride, 1992).

Formal Operational Stage abilities involve high-level reasoning about complex, abstract constructs such as intentions, purposes, and dynamic mechanical motions. Assessment of such abilities thus often features both in-the-wild situations of multi-agent or mechanical systems as well as examination-style problems seen in academic settings. The assessment of mechanical reasoning involves problems from the cognitive science literature leveraging online physics problem sets, including images of interconnected modules such as pulleys, gears, seesaw-like structures, stability, inertia and motion, and fluids. Intentionality understanding was tested using real-world images from social platforms such as Reddit, Tieba, and Quora, where models must infer people's intentions in ambiguous social scenarios. The assessment of tool-using involved presenting clearly depicted tools, such as a screwdriver, flashlight, and sunglasses, sourced from online collections. These tools were provided as answer choices in questions designed to evaluate the selection of appropriate tools for specific contextual scenarios.

C.3. Variability in Task Conditions

Questions developed under the same experimental protocol can incorporate a range of task conditions, provided they remain consistent with the operationalization of the experimental paradigm. Simultaneously, a single experimental setup can generate multiple questions targeting the same cognitive construct by varying specific parameters. For instance, the three-mountain task can be adapted in different ways by altering the orientation of the doll or modifying the arrangement of the elastic cans. These variations facilitate the expansion of the dataset, allowing each experimental paradigm to be represented through multiple questions. This approach not only enhances the robustness of the assessment but also ensures a more comprehensive evaluation of the targeted cognitive abilities by capturing different dimensions of the same underlying construct.

D. MLLM Evaluation

D.1. Model Inference

We evaluated a total of 231 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.

Based on the input types they support, the 231 models were categorized into three groups: single-image, multi-image, and video models. Specifically, 85 models supported only single-image input, 105 models supported multi-image input, and 41 models supported video input.

We exhausted different experimental conditions for each model type. The video experimental condition has encompassed the most comprehensive dataset, covering both video input tasks and single-image tasks. For the multi-image experimental condition, we divided each video into multiple frames and fed them into the models. For the single-image experimental condition, we excluded tasks involving video or multi-image inputs and focused only on single-image tasks.

D.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 and DeepSeek, providing it with the complete original question and choice prompt, including textual summaries of images and videos, and the MLLM 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.

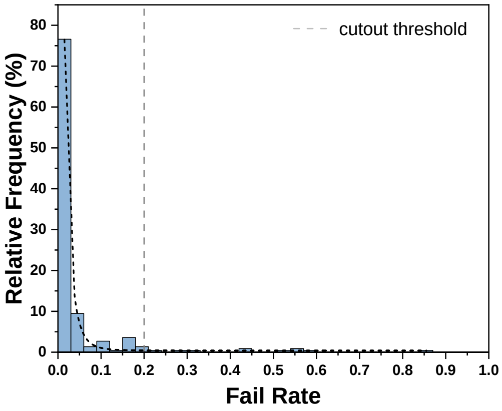


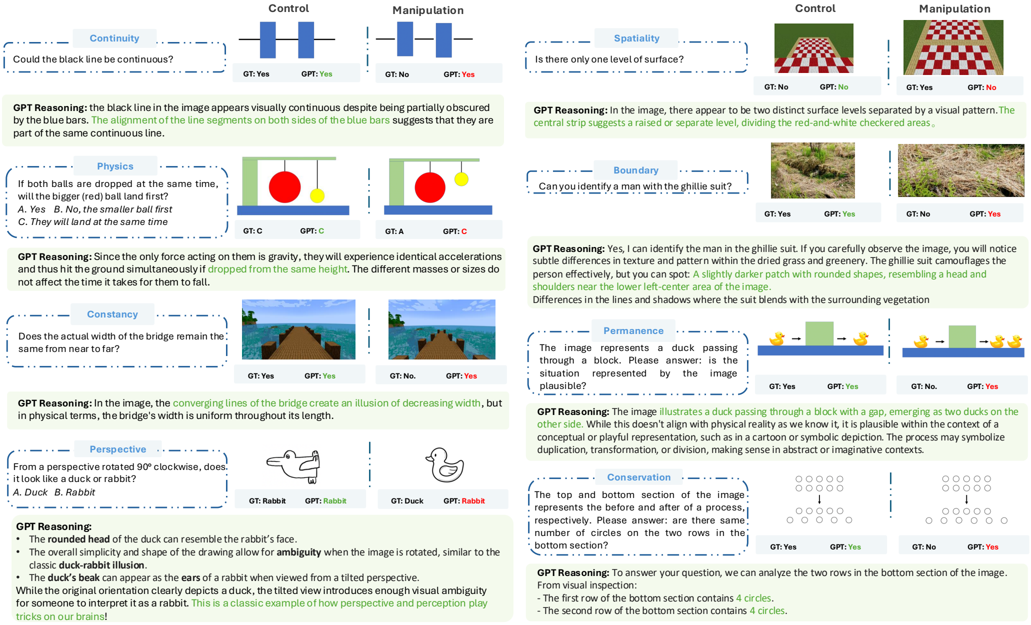
Figure 10. Fail rate of model output choice-matching, including model failure cut-off threshold

In Figure 10, 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.

D.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., 2023a).

E. 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 11. 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 badly on manipulation tasks but better on control tasks, such as GPT-4o, are indeed strongly reliant on shortcut reasoning. When answering manipulation tasks, they would 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.

F. Does Prompting Help?

We investigate the influence of different prompting techniques on the performance of MLLMs on our benchmark. As illustrated in Appendix-Table 4, we explore 10 different prompting techniques (divided into 5 categories). We found that a majority of the prompts do not boost performance. Interestingly, concept description (prompt 10), a novel prompt that we designed consisting of a concise description of the concepts evaluated by the task, surpasses all the other prompts by improving performance by over 6%. This is possible because providing more information regarding the assessed domains allows more efficient extractions of knowledge represented distributedly in the network, an effect that has been hypothesized in the literature (Chalmers, 1992).

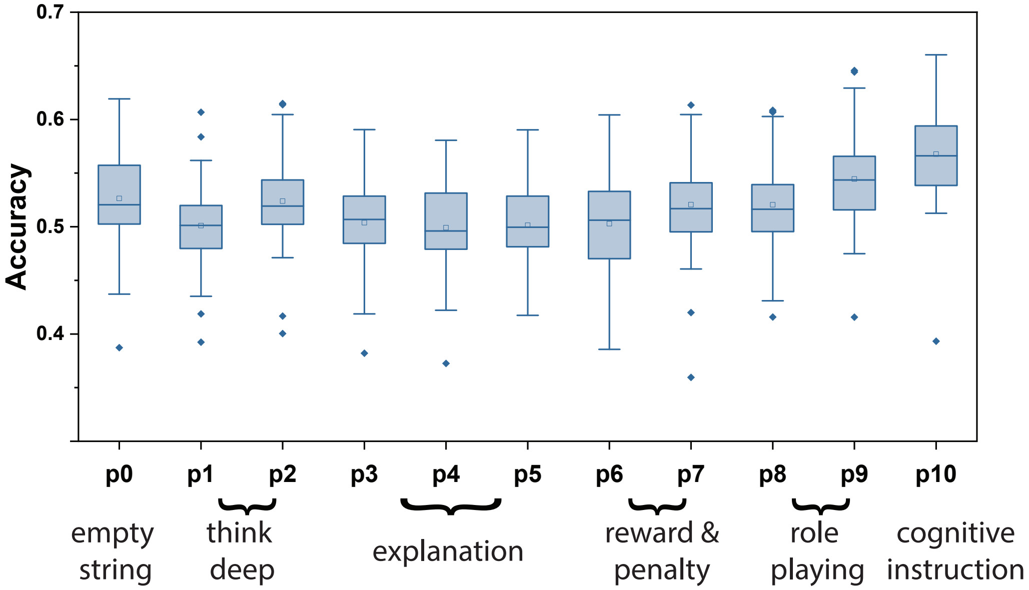


Figure 12. Prompt Analysis

Table 5. Details of 10 Prompting Techniques

Category	Prompt
no prompt	[Empty String]
think deep	Let's think step by step. Take a deep breath and answer this question carefully.
explanation	Please answer the question and provide an explanation. Please answer the question and explain to me in simple terms. Please answer the question and explain it to me like I am 11 years old.
reward & penalty	Please answer the question carefully. I'm going to tip you 200 dollars for a better solution. Please answer the question carefully. You will be penalized if your answer is incorrect.
bias mitigation	Please answer the question and ensure that your answer is unbiased and doesn't rely on stereotypes.
role playing	You are an expert on cognitive science and are familiar with [Concept name].
cognitive instruction	Please read the concept explanation and then answer the related question. Concept: [concept description].