

COGDEVELOP2K: REVERSED COGNITIVE DEVELOPMENT IN MULTIMODAL LARGE LANGUAGE MODELS

Anonymous authors

Paper under double-blind review

ABSTRACT

Are Multi-modal Large Language Models (MLLMs) stochastic parrots? Do they genuinely understand and are capable of performing the tasks they excel at? This paper aims to explore the fundamental basis of MLLMs, i.e. core cognitive abilities that human intelligence builds upon to perceive, comprehend, and reason. To this end, we propose **CogDevelop2K**, a comprehensive benchmark that spans 12 sub-concepts from fundamental knowledge like object permanence and boundary to advanced reasoning like intentionality understanding, structured via the developmental trajectory of a human mind. We evaluate 46 MLLMs on our benchmarks. Comprehensively, we further evaluate the influence of evaluation strategies and prompting techniques. Surprisingly, we observe a **reversed** cognitive developmental trajectory compared to humans.

1 INTRODUCTION

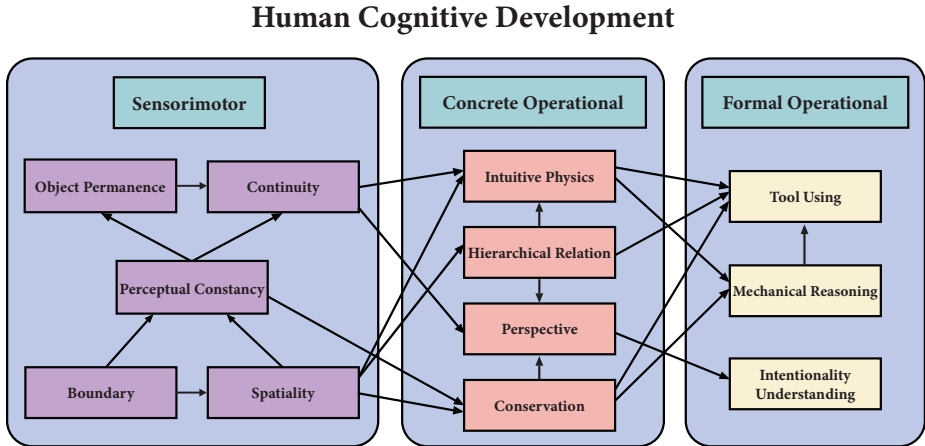
Building on the foundation of advanced large language models (LLMs), multi-modal large language models (MLLMs) have recently demonstrated human-level performance in complex tasks involving high-level reasoning, perception, and cognition Li et al. (2024a); Liu et al. (2024); Team (2023); Fu et al. (2023); OpenAI (2023) such as Spatial Reasoning Chen et al. (2024); Cai et al. (2024), OCR Mori et al. (1999), Scene Understanding Cordts et al. (2016); Chen et al. (2017), Action Recognition Jhuang et al. (2013); Herath et al. (2017) and Prediction Lan et al. (2014); Kong & Fu (2022), etc. The progress in MLLMs has reignited hopes for achieving Artificial General Intelligence (AGI). However, we pose a critical question: *Do MLLMs truly comprehend these tasks and possess the genuine capabilities to perform them, or are they merely "stochastic parrots" that rely on learning spurious correlations?* To explore this, we draw inspiration from the development of human cognition.

Extensive research in human cognitive development suggests the existence of core cognition which grounds the diversity of incredible human intelligent abilities (Spelke et al., 1992; 1994; 1995; Spelke & Kinzler, 2007; Mitchell, 2020; 2021), and such core cognition is unraveled via the developmental cascades of the human mind (Masten & Cicchetti, 2010). From infancy to early adulthood, core cognitive concepts emerge and develop 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 the ability to understand other people's perspectives (Jacoboni, 2009; De Waal & Preston, 2017; Liu et al., 2017; Caviola et al., 2021; Ninomiya et al., 2020).

As highlighted in previous research, core cognitive abilities form the foundational basis of higher-level human intelligence that existing MLLMs excel at, but such excellency do not translate into a general domain (Mitchell, 2020; 2021; Shiffrin & Mitchell, 2023). The performance of MLLMs on core cognitive tasks therefore provides a more profound insight into their true capacities for knowledge, reasoning, and perception. This, in turn, serves as a critical indicator of whether MLLMs possess genuine intelligence or if they are merely "stochastic parrots" that depend on learning spurious correlations. To this end, we draw on theoretical and empirical approaches from developmental science to create benchmarks that evaluate core cognitive abilities in large vision-language models,

054 examining the relationships between these abilities. On a high level, we follow Jean Piaget’s theory
 055 of cognitive development, which identifies four stages in children: sensorimotor, preoperational,
 056 concrete operational, and formal operational (Piaget, 1950; Piaget & Inhelder, 1969; 1974). During
 057 the sensorimotor stage, infants acquire knowledge through sensory experiences and actions, devel-
 058 oping an understanding of basic object properties, such as permanence, continuity, and boundaries.
 059 In the preoperational stage, symbolic representation emerges, along with a grasp of basic physical
 060 properties. The concrete operational stage is characterized by the development of logical thinking
 061 and an understanding of intuitive physics. Finally, the formal operational stage introduces more
 062 advanced cognitive abilities, including abstraction, hypothetical reasoning, counterfactual thinking,
 063 and tool use. The interdependence and developmental trajectories of these abilities can be mapped
 064 in terms of a tree-like structure (as illustrated in Fig. 1).

065 To evaluate the performance of MLLMs on the core cognitive abilities, we curate the first-ever vision
 066 cognitive development benchmark, termed as CogDeveop2K, which consists of a total of 2519 ques-
 067 tions with 2517 images and 455 videos. Then, we evaluate 46 MLLM models on our benchmark that
 068 spans all four stages of cognitive development. We introduce a novel multi-frame question format
 069 to evaluate models’ co-reference, cognitive reasoning and temporal understanding capability simul-
 070 taneously. Forty-seven models are compared against a human baseline under zero-shot conditions
 071 using 11 different prompts (including no prompt). Surprisingly, while prompts can boost model
 072 performance by 8.1%, models still demonstrate reversed trends in cognitive development against
 073 those observed in children. For example, GPT series perform better in formal operation stage while
 074 performing worse in concrete operation stage.



091 Figure 1: Map of core cognitive concepts during human developmental stages

092
093
094 **2 RELATED WORK**

095
096 **2.1 MULTI-MODAL LARGE LANGUAGE MODELS**

097
098 The Vision Language Model (VLM) has a long history from Convolution Neural Networks (CNN)
 099 and Recurrent Neural Networks (RNN) (Karpathy & Fei-Fei, 2014; Vinyals et al., 2015) to uni-
 100 fied modeling of visual and text modality with transformers (Li et al., 2019; Xu et al., 2023; Tan
 101 & Bansal, 2019; Alayrac et al., 2022; Radford et al., 2021). With the advancement of Large Lan-
 102 guage Models (LLMs), existing state-of-the-art MLLMs (Liu et al., 2024; Li et al., 2023) adopt
 103 open-sourced Large Language Models such as Llama (Touvron et al., 2023), Mistral (Jiang et al.,
 104 2023), etc. Instruction Tuning is also introduced to further improve the task generalization ability of
 105 MLLMs (Liu et al., 2024; Dai et al., 2023). To acquire open-ended conversation abilities, LLaVA
 106 (Liu et al., 2024) proposes to distill the conversational abilities of ChatGPT to MLLMs, boosting
 107 performance by a large margin, which becomes a defacto procedure in the area (Wang et al., 2023;
 Bai et al., 2023; Team, 2023; 2024; Sun et al., 2023; Li et al., 2022).

2.2 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 develop 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 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).

Preoperational and concrete operational stage are the second and third stage of Piaget's cognitive development. Typically spanning over 2 to 7 years of age, preoperational stage is the transitional stage to concrete operational stage, which children enters around 7 years of age. During then, 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 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 preoperational stage is mostly meaningful as the transitional period preceding concrete operational stage, we do not have evaluation dimensions specifically targeting the stage. However, tasks targeting concrete operational stage could assess presentations of knowledge associated with preoperational stage, as prominently illustrated by 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 systematic 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).

3 COGDEVELOP2K

3.1 EVALUATION DIMENSION

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

Spatiality Spatiality, particularly demonstrated through the A-not-B task, involves a child's understanding of the location of objects in relation to their environment (Bell & Adams, 1999). In a classic A-not-B task, an object is hidden at location A (such as under a cup) and the child successfully finds it several times. Then, the object is visibly moved to a different location B (under a different cup), in

162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215

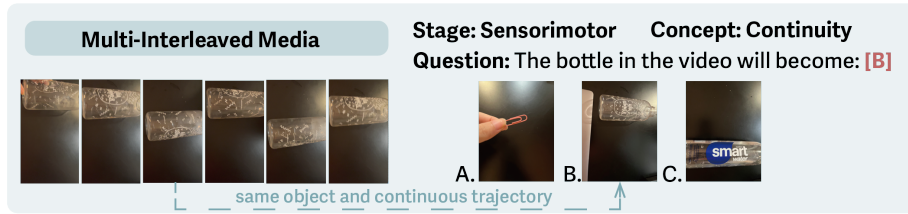


Figure 2: A video-image interleaved example of multi-frame questions. To correctly infer the answer, model needs to understand the question by mapping each image (co-reference) to its option letter, to understand correlation between frames (temporal understanding) and to infer the possible trajectory of the bottle (reasoning).












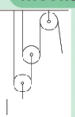
Stage: Sensorimotor		
<p style="text-align: center;">Boundary</p> <p>What is the boundary line of the pillow? [A]</p> <p>A. Rectangle B. Circle C. Star</p> 	<p style="text-align: center;">Spatiality</p> <p>Does the glass on the left in the first image contain more liquid than the glass on the right in the third image? [No]</p> 	<p style="text-align: center;">Perceptual Constancy</p> <p>If I cover <img1>, when the cover is removed, it will not become <img2>: [Yes]</p> 
<p style="text-align: center;">Continuity</p> <p>How many lines are there in the scene? [A]</p> <p>A. One B. Two C. Three</p> 	<p style="text-align: center;">Permanence</p> <p>Can the die be removed without moving the glass? [Yes]</p> 	<p style="text-align: center;">Intuitive Physics</p> <p>Which of the two horizontal rectangular bars on the left and right is less likely to fall</p> <p>A. The Left B. The Right</p> 
Stage: Concrete Operation		
<p style="text-align: center;">Perspective</p> <p>From the doll's point of view, which object appears as the leftist of the three? [C]</p> <p>A. Metal Cup B. Paper Cup C. The Can</p> 	<p style="text-align: center;">Hierarchy</p> <p>Are there more broken chairs or more brown armchairs? [A]</p> <p>A. Broken chairs B. Brown armchairs</p> 	<p style="text-align: center;">Conservation</p> <p>Is the length of the left line and the right line the same? [Yes]</p> 
Stage: Formal Operation		
<p style="text-align: center;">Tool Use</p> <p>What is the usage of the following object [B]</p> <p>A. To shed light B. To measure C. To support</p> 	<p style="text-align: center;">Intentionality</p> <p>What does the man at the bottom intend to do? [B]</p> <p>A. To take the shirt off B. To do construction</p> 	<p style="text-align: center;">Mechanical Reasoning</p> <p>In which direction will the rectangular brick go if we pull the rope on the right? [B]</p> <p>A. Going Downwards B. Going Upwards</p> 

Figure 3: We demonstrate examples of different sub-concepts from the three stages.

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 in cognitive development where an individual understands 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 the child’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.



Figure 4: Reversed Cognitive Development in Advanced Models

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 cognitive phenomena that children begin to understand hierarchical relations and be able 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

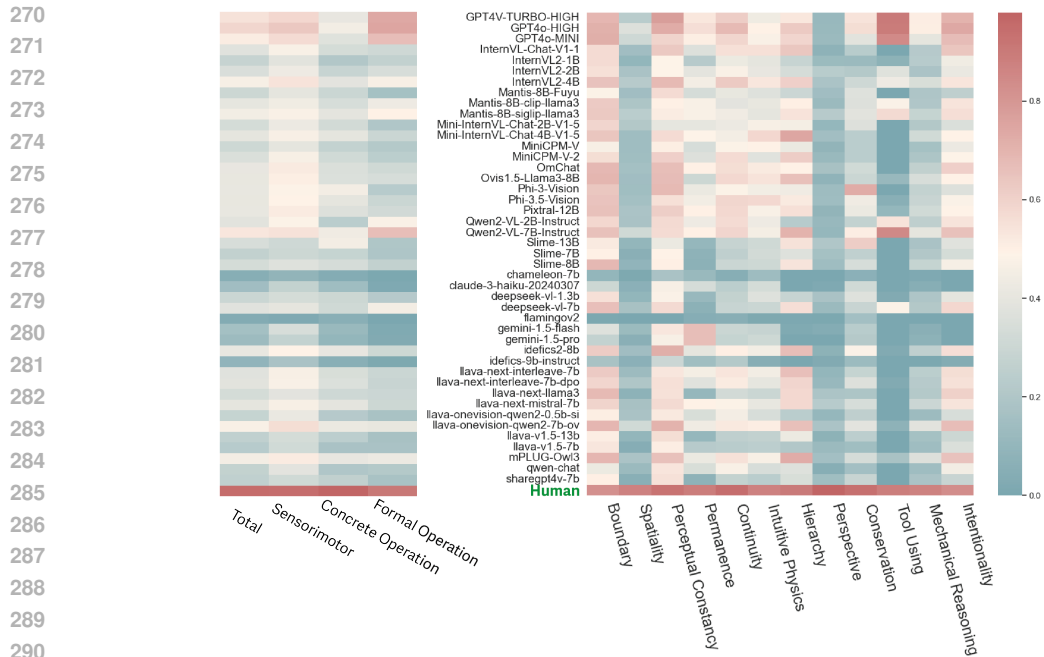


Figure 5: Multimodal Large Language Models and Human Performances

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 organizations 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).

Intention 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 (O’Brien & Shapiro, 1968; Cesana-Arlotti et al., 2018), such as induction, abduction, syllogism, Modus Ponens and Modus Tollens, and reasoning forms (Byrne, 2016), such as hypotheticals and counterfactuals, figure out how to manipulate these systems to achieve a desired outcome (Hegarty, 2004).

Tool Using Tool using refers to the ability to manipulate objects in its environment as aids in achieving a specific goal, such as obtaining food or modifying the surroundings. A lot of cognitive components 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 sensorimotor capabilities

(Gibson, 1979). For example, a door handle affords pulling or pushing, indicating how the door should be operated.

3.2 DATA SOURCE


CogDevelop2K comprises 2517 images and 445 videos with multimodal options and questions, crawled primarily from networks as well as self-recorded content. For example, concept intentionality were collected from platforms, including Wikipedia, Reddit, Twitter, Quora, and TieBa. Some of the options were adapted from user comments to ensure content diversity and relevance. Videos for intuitive physics were either self-recorded or produced using Physion¹.


All concept questions were annotated by four researchers with cognitive science and computer science background, then reviewed by two independent researchers. For a question to pass the screening stage, a minimum correctness rate of 95% was required from both reviewers.

3.3 DATASET DESIGN

Existing datasets typically support only one question-answer format or single modality type, which hinders the assessment of reasoning capabilities across different modalities within the same domain. For instance, current interleaved image understanding and video understanding models cannot be effectively compared on the same question. To address this limitation, we propose the CogDevelop2k benchmark, which includes multiple Q&A formats (e.g., multiple-choice, true/false, and numeric question-answer) and complicate question-answering by incorporating a new image-video-text interleave format as shown in Fig. 2.

To further explore the cognitive development capabilities of models across these modalities, we optimized CogDevelop2k as follows:

Addressing Weak Image-Text Correlation and Imbalance In existing interleaved image-text datasets, the correspondence between images and text is often loose, and text provides marginal information for image modeling. This imbalance can cause models to over-rely on textual information, especially when text segments are lengthy (Lin et al. (2023)). To address this issue and focus on the image understanding abilities of the model, we eliminated sentences that describe the image, such as "A. an oil painting ". This ensures that the textual information is highly relevant to but does not overlap the image content.

Testing Co-Reference, Reasoning, and Temporal Understanding with novel Multi-Frame Questions Multi-frame questions in CogDevelop can simultaneously evaluate a model's three inference ability: *Co-Reference*, *Reasoning*, and *Temporal Understanding* (Jiang et al. (2024)). Co-reference involves linking natural language descriptions with specific image inputs (e.g., "the first image" or "A. "). Reasoning requires models to make decisions based on cognitive knowledge, such as describing spatial relationships. Temporal Understanding, on the other hand, tests the model's capability to comprehend sequences of frames in terms of temporal order (multi-frame) and correlation (multi-view) (Li et al. (2024b)). Existing interleaved multi-image datasets can not adequately test all three properties simultaneously. For example, video datasets with temporal information often include only a single video, while multi-image datasets that require co-reference lack temporal dependencies. To address this, CogDevelop introduces multi-video interleaving and video-image interleave formats (multi-frame) to evaluate all three properties concurrently. The statistics of the dataset are presented in Table 2.

3.4 EVALUATION STRATEGY

We comprehensively evaluate models' capability of cognitive reasoning using **46 multi-image interleave MLLMs** with 11 different prompts. The two evaluation baselines are outlined as follows:

Human baseline We recruited 22 participants, all of whom are Chinese college students proficient in English. Each annotator was asked to label 2 to 6 concepts, with each concept being annotated

¹<https://physion.net/>

Table 1: Main statistics in CogDevelop2k. All the questions are image/video-text interleaved.

Statistic	Number
single-frame	1677
multi-frame	842
* multiple images	200
* single video	401
* multiple videos	124
* video-image-text	117
total	2519

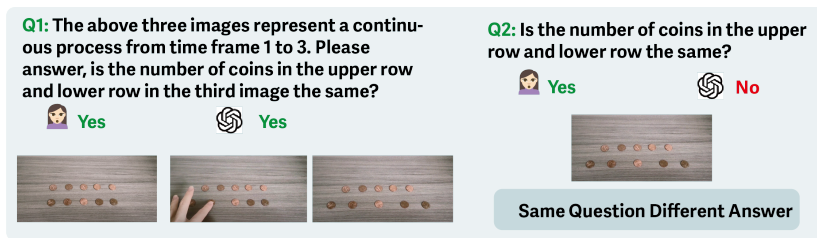


Figure 6: MLLMs’ Dissociation Between Law of Conservation and Rudimentary Quantity Understanding as Exemplified by GPT-4o

by two or more annotators. Participants were instructed to skip a question if the question is worded ambiguously or is too complicated to answer in 90 seconds.

Zero-Shot-448²-Circular Baseline Similar to Lu et al. (2022), the zero-shot 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 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. All images and videos were resized to 488². (Liu et al. (2023)).

Prompts Strategically crafted prompts can enhance model performance, regardless of whether fine-tuning is applied (Bsharat et al. (2023); Yang et al. (2023)). In contrast to Science QA datasets, where image captions are incorporated as the context in the prompts, this approach can cause models to over-rely on text rather than reasoning about the image content. To mitigate this, we use image-independent contexts, such as relevant concept introductions and character assignments, which encourage models to reason beyond the provided textual information. The prompts we used can be categorized into leading words, deeper thing, role assignment, reward/penalty, and explanation. Detailed results can be found in Sec 4.

4 RESULTS AND DISCUSSION

We systematically evaluate 48 Multi-modal Large Language Models on the CogDevelop2K benchmark, which spans 12 cognitive concepts designed to assess a broad range of the developmental trajectory of Multi-modal Large Language Models. These abilities substantiate core cognition ranging from object permanence and boundary to mechanical reasoning and intentionality understanding. The models were tested across multiple question formats and ten prompt variations, yielding a comprehensive assessment of their core cognition. For example, in the sensorimotor stage, GPT families show moderate performance, with accuracy scores between 0.4 and 0.6. In the concrete operational stage, GPT families show lower performance, with accuracy scores between 0.2 and 0.4. Nevertheless, in the formal operational stage, GPT families show stronger performance, with accuracy scores

Table 2: Evaluation of different prompting techniques. The best result is achieved when the concept explanation is provided to the model. We highlight the improvement over empty string in red.

Category	Prompt	GPT4V Turbo High	GPT4o High	GPT4o Mini
	Empty String	0.519	0.555	0.487
Leading	1. Let’s think step by step.	0.531	0.577	0.489
	2. Take a deep breath and answer this question carefully.	0.522	0.562	0.489
Word Deeper	3. Please answer the question and provide an explanation.	0.518	0.562	0.499
Thinking	4. Please answer the question and explain to me in simple terms.	0.476	0.569	0.501
	5. Please answer the question and ensure that your answer is unbiased and doesn’t rely on stereotypes.	0.522	0.575	0.478
Role Assignment	6. (Assign assistance’s role) You are an expert on cognitive science and are familiar with [Concept name]	0.565	0.617	0.545
	7. (Assign audience’s role) Please answer the question and explain it to me like I am 11 years old.	0.538	0.564	0.496
Reward & Penalty	8. Please answer the question carefully. I’m going to tip you 200 dollars for a better solution.	0.528	0.563	0.487
	9. Please answer the question carefully. You will be penalized if your answer is incorrect.	0.522	0.566	0.491
Explanation	10. Please read the concept explanation and then answer the related question. Concept: [concept description].	0.586 (+ 0.067)	0.636 (+ 0.081)	0.547 (+ 0.06)

between 0.6 and 0.8. Surprisingly, we find an inverse cognitive developmental trajectory compared to humans in more advanced models, which are typically regarded as state-of-the-art (Fig. 3 and Fig. 4).

Influence of Prompts. We investigate the influence of different prompting techniques on the performance of MLLMs on our benchmark. As illustrated in Table 2, we explore 10 different prompting techniques (divided into 5 categories). We observe that most prompts are useful on our benchmark, increasing the averaged performance by at least 1%. Concept explanation, which offers a clearer context of the question to the MLLMs, surpasses all the other prompts by at least 6%.

4.1 COGNITIVE DISCUSSIONS

We have demonstrated that MLLMs exhibit reverse cognitive development. Namely, they are systematically proficient at complex tasks that are typically understood to require abilities underlying simple tasks that they perform poorly. This surprising finding could appear as challenge to the current foundational architecture of MLLMs as a long-term solution to achieve human-like general intelligence (Summerfield, 2022).

Our finding complement earlier research which raises worry that large language models may be “stochastic parrots” that merely link words and sentences together based on probabilistics but do not understand meanings and logic (Searle, 1980; Bender et al., 2021). If an intelligent agent truly understand that changes in spatial arrangement do not affect quantity, it is logically impossible for it to count correctly the amount of coins when the transformation is shown, while count wrongly when

the transformation is not shown (Fig. 6). Contradictions like this reveal that MLLMs virtually do not understand the answers they produce when tackling cognitive reasoning tasks. If this is indeed the case, it may account for a variety of difficulties that MLLMs encounter, particularly in achieving robustness across changing task situations (Zhao et al., 2024).

The developmental trajectory of human cognition is marked by complex cognitive abilities being grounded upon extremely robust understandings of a series of foundational concepts, namely core knowledge (Spelke, 2000; Spelke & Kinzler, 2007). Through early stages of development, children exhibit rudimentary yet stable understandings of objects, actions, number, space, and social partners, each dimension lays the foundations for the acquisitions of complex abilities in later life. It has been suggested that core knowledge is precisely what supports the robustness of human cognition instantiated in commonsense reasoning (Mitchell, 2021). In reverse, the inability to implement core knowledge in artificial intelligence models prevent them from achieving human-level robustness in performances, even if such models seem to excel at certain complex cognitive reasoning tasks (Mitchell, 2020; 2021; Shiffrin & Mitchell, 2023; Palmarini & Mitchell, 2024). MLLMs’ poor performances on foundational concepts like spatiality, permanence, continuity, and perspective, those that directly reflect upon grasps of core knowledge, while achieve proficiency in complex concepts like tool using and intention understanding exactly exemplifies this concern.

To summarize, MLLMs’ performance on cognitive reasoning tasks significantly diverges from that of humans, namely in terms of having a reverse developmental trajectory between simple and complex abilities. This highlights the concerns that MLLMs do not genuinely understand meanings, which require the grounding of human singular experiences (Turing, 1950; Wittgenstein, 1958; Dennett, 1969; Searle, 1980).

5 CONCLUSION

In this paper, we explored the cognitive capabilities of Multi-modal Large Language Models (MLLMs) through the lens of core cognitive abilities that underpin human intelligence. By introducing CogDevelop2K, a novel benchmark that spans 12 subconcepts across developmental stages, we aimed to assess the fundamental understanding and reasoning capacities of MLLMs. Our evaluation of 46 models revealed intriguing insights, including a reversed cognitive developmental trajectory compared to humans. This finding raises questions about whether MLLMs truly comprehend tasks or simply exhibit performance without genuine understanding. These results underscore the need for further investigation into the cognitive foundations of MLLMs, as well as the influence of evaluation strategies and prompting techniques in shaping their outcomes. Ultimately, this study serves as a step toward unraveling the nature of MLLM intelligence and their potential limitations in mirroring human cognitive development.

REFERENCES

- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.
- Kelsey R Allen, Kevin A Smith, and Joshua B Tenenbaum. 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.
- Glenda Andrews and Graeme S Halford. Children’s ability to make transitive inferences: The importance of premise integration and structural complexity. *Cognitive Development*, 13(4):479–513, 1998.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.
- Renee Baillargeon. Representing the existence and the location of hidden objects: Object permanence in 6-and 8-month-old infants. *Cognition*, 23(1):21–41, 1986.

- 540 Renee Baillargeon, Elizabeth S Spelke, and Stanley Wasserman. Object permanence in five-month-
541 old infants. *Cognition*, 20(3):191–208, 1985.
- 542 Chris L Baker, Rebecca Saxe, and Joshua B Tenenbaum. Action understanding as inverse planning.
543 *Cognition*, 113(3):329–349, 2009.
- 544 Yvonne Barnes-Holmes, Louise McHugh, and Dermot Barnes-Holmes. Perspective-taking and the-
545 ory of mind: A relational frame account. *The Behavior Analyst Today*, 5(1):15–25, 2004.
- 546 Martha Ann Bell and Stephanie E Adams. Comparable performance on looking and reaching ver-
547 sions of the a-not-b task at 8 months of age. *Infant Behavior and Development*, 22(2):221–235,
548 1999.
- 549 Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the
550 dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM
551 conference on fairness, accountability, and transparency*, pp. 610–623, 2021.
- 552 George Berkeley. *An Essay Towards A New Theory of Vision*. Dublin, 1709.
- 553 Bennett I Bertenthal, Gustaf Gredebäck, and Ty W Boyer. Differential contributions of development
554 and learning to infants’ knowledge of object continuity and discontinuity. *Child Development*, 84
555 (2):413–421, 2013.
- 556 Grégoire Borst, Nicolas Poirel, Ariette Pineau, Mathieu Cassotti, and Olivier Houdé. Inhibitory
557 control efficiency in a piaget-like class-inclusion task in school-age children and adults: a devel-
558 opmental negative priming study. *Developmental psychology*, 49(7):1366, 2013.
- 559 Sondos Mahmoud Bsharat, Aidar Myrzakhan, and Zhiqiang Shen. Principled instructions are all
560 you need for questioning llama-1/2, gpt-3.5/4. *arXiv preprint arXiv:2312.16171*, 2023.
- 561 Ruth MJ Byrne. Counterfactual thought. *Annual review of psychology*, 67(1):135–157, 2016.
- 562 Wenxiao Cai, Yaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and
563 Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. *arXiv preprint
564 arXiv:2406.13642*, 2024.
- 565 Lucius Caviola, Stefan Schubert, and Joshua D Greene. The psychology of (in) effective altruism.
566 *Trends in Cognitive Sciences*, 25(7):596–607, 2021.
- 567 Nicolás Cesana-Arlotti, Ana Martín, Ernő Téglás, Liza Vorobyova, Ryszard Cetnarski, and Luca L
568 Bonatti. Precursors of logical reasoning in preverbal human infants. *Science*, 359(6381):1263–
569 1266, 2018.
- 570 Michael Chapman and Michelle L McBride. Beyond competence and performance: Children’s class
571 inclusion strategies, superordinate class cues, and verbal justifications. *Developmental Psychol-
572 ogy*, 28(2):319, 1992.
- 573 Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia.
574 Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings
575 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465,
576 2024.
- 577 Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille.
578 Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and
579 fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):
580 834–848, 2017.
- 581 R. Breckinridge Church and Susan Goldin-Meadow. The mismatch between gesture and speech as
582 an index of transitional knowledge. *Cognition*, 23:43–71, 1986.
- 583 Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo
584 Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban
585 scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern
586 recognition*, pp. 3213–3223, 2016.

- 594 Grace J Craig, Jean A Love, and Ellis G Olim. An experimental test of piaget’s notions concerning
595 the conservation of quantity in children. *Child Development*, 44(2):372–375, 1973.
- 596
- 597 Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng
598 Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-
599 purpose vision-language models with instruction tuning. In Alice Oh, Tristan Nau-
600 mann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances*
601 *in Neural Information Processing Systems 36: Annual Conference on Neural Informa-*
602 *tion Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16,*
603 *2023*, 2023. URL [http://papers.nips.cc/paper_files/paper/2023/hash/](http://papers.nips.cc/paper_files/paper/2023/hash/9a6a435e75419a836fe47ab6793623e6-Abstract-Conference.html)
604 [9a6a435e75419a836fe47ab6793623e6-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2023/hash/9a6a435e75419a836fe47ab6793623e6-Abstract-Conference.html).
- 605 Frans BM De Waal and Stephanie D Preston. Mammalian empathy: behavioural manifestations and
606 neural basis. *Nature Reviews Neuroscience*, 18(8):498–509, 2017.
- 607
- 608 Daniel C. Dennett. *Content and Consciousness*. Routledge, London, 1969.
- 609
- 610 Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu
611 Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation
612 benchmark for multimodal large language models. *arXiv preprint arXiv: 2306.13394*, 2023.
- 613
- 614 Kanishk Gandhi, Gala Stojnic, Brenden M Lake, and Moira R Dillon. Baby intuitions benchmark
615 (bib): Discerning the goals, preferences, and actions of others. *Advances in neural information*
processing systems, 34:9963–9976, 2021.
- 616
- 617 James J Gibson. *The Ecological Approach to Visual Perception. (1st ed.)*. Psychology Press, 1979.
- 618
- 619 EJ Green. Perceptual constancy and perceptual representation. *Analytic Philosophy*, 2023.
- 620
- 621 G S Halford. An experimental test of piaget’s notions concerning the conservation of quantity in
622 children. *Journal of experimental child psychology*, 6(1):33–43, 2011.
- 623
- 624 Mary Hegarty. Mechanical reasoning by mental simulation. *Trends in cognitive sciences*, 8(6):
625 280–285, 2004.
- 626
- 627 Samitha Herath, Mehrtash Harandi, and Fatih Porikli. Going deeper into action recognition: A
628 survey. *Image and vision computing*, 60:4–21, 2017.
- 629
- 630 Linda Hermer and Elizabeth Spelke. Modularity and development: The case of spatial reorientation.
631 *Cognition*, 61(3):195–232, 1996.
- 632
- 633 Olivier Houdé, Arlette Pineau, Gaëlle Leroux, Nicolas Poiriel, Guy Perchey, Céline Lanoë, Amélie
634 Lubin, Marie-Renée Turbelin, Sandrine Rossi, Grégory Simon, Nicolas Delcroix, Franck Lam-
635 berton, Mathieu Vigneau, Gabriel Wisniewski, Jean-René Vicet, and Bernard Mazoyer. Func-
636 tional magnetic resonance imaging study of piaget’s conservation-of-number task in preschool
637 and school-age children: a neo-piagetian approach. *Journal of experimental child psychology*,
110(3):332–346, 2011.
- 638
- 639 William Huitt and John Hummel. Piaget’s theory of cognitive development. *Educational psychology*
640 *interactive*, 3(2):1–5, 2003.
- 641
- 642 Marco Iacoboni. Imitation, empathy, and mirror neurons. *Annual review of psychology*, 60(1):
643 653–670, 2009.
- 644
- 645 Bärbel Inhelder and Jean Piaget. *The Growth of Logical Thinking from Childhood to Adolescence*.
Basic Books, 1958.
- 646
- 647 Ray Jackendoff. Parts and boundaries. *Cognition*, 41(1-3):9–45, 1991.
- Pierre Janet. Mental pathology. *Psychological Review*, 12(2-3):98, 1905.

- 648 Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards under-
649 standing action recognition. In *Proceedings of the IEEE international conference on computer*
650 *vision*, pp. 3192–3199, 2013.
- 651 Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chap-
652 lot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier,
653 L elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas
654 Wang, Timoth e Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv: 2310.06825*,
655 2023.
- 656 Dongfu Jiang, Xuan He, Huaye Zeng, Con Wei, Max Ku, Qian Liu, and Wenhua Chen. Mantis:
657 Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024.
- 658 Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descrip-
659 tions. *arXiv preprint arXiv: 1412.2306*, 2014.
- 660 Roberta Kestenbaum, Nancy Termine, and Elizabeth S Spelke. Perception of objects and object
661 boundaries by 3-month-old infants. *British journal of developmental psychology*, 5(4):367–383,
662 1987.
- 663 Byung-Geun Khang and Qasim Zaidi. Illuminant color perception of spectrally filtered spotlights.
664 *Journal of Vision*, 4(9):2–2, 2004.
- 665 In-Kyeong Kim and Elizabeth S Spelke. Perception and understanding of effects of gravity and
666 inertia on object motion. *Developmental Science*, 2(3):339–362, 1999.
- 667 EA Kirkpatrick. The part played by consciousness in mental operations. *The Journal of Philosophy,*
668 *Psychology and Scientific Methods*, 5(16):421–429, 1908.
- 669 Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *International Journal of*
670 *Computer Vision*, 130(5):1366–1401, 2022.
- 671 Deanna Kuhn and John Angelev. An experimental study of the development of formal operational
672 thought. *Child Development*, pp. 697–706, 1976.
- 673 Tian Lan, Tsung-Chuan Chen, and Silvio Savarese. A hierarchical representation for future action
674 prediction. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland,*
675 *September 6–12, 2014, Proceedings, Part III 13*, pp. 689–704. Springer, 2014.
- 676 Robin Le Poidevin. Continuants and continuity. *The Monist*, 83(3):381–398, 2000.
- 677 Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan.
678 Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF*
679 *Conference on Computer Vision and Pattern Recognition*, pp. 13299–13308, 2024a.
- 680 Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen,
681 Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou, and Luo Si. mplug:
682 Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint*
683 *arXiv: 2205.12005*, 2022.
- 684 Chieh Li, Ronald L Nuttall, and Shuwen Zhao. A test of the piagetian water-level task with chinese
685 students. *The Journal of Genetic Psychology*, 160(3):369–380, 1999.
- 686 Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li.
687 Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv*
688 *preprint arXiv:2407.07895*, 2024b.
- 689 Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image
690 pre-training with frozen image encoders and large language models. *CONFERENCE*, 2023.
- 691 Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple
692 and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.
- 693 Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz,
694 Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models, 2023.
- 695
- 696
- 697
- 698
- 699
- 700
- 701

- 702 David Liu, Henry M Wellman, Twila Tardif, and Mark A Sabbagh. Theory of mind development
703 in chinese children: a meta-analysis of false-belief understanding across cultures and languages.
704 *Developmental psychology*, 44(2):523, 2008.
- 705
706 Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances*
707 *in neural information processing systems*, 36, 2024.
- 708
709 Shari Liu, Tomer D Ullman, Joshua B Tenenbaum, and Elizabeth S Spelke. Ten-month-old infants
710 infer the value of goals from the costs of actions. *Science*, 358(6366):1038–1041, 2017.
- 711
712 Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan,
713 Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around
714 player? *arXiv preprint arXiv:2307.06281*, 2023.
- 715
716 Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord,
717 Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for
718 science question answering. In *The 36th Conference on Neural Information Processing Systems*
(*NeurIPS*), 2022.
- 719
720 Sugandha Marwaha, Mousumi Goswami, and Binny Vashist. Prevalence of principles of piaget’s
721 theory among 4-7-year-old children and their correlation with iq. *Journal of clinical and diag-*
nostic research: JCDR, 11(8):ZC111, 2017.
- 722
723 Ann S Masten and Dante Cicchetti. Developmental cascades. *Development and psychopathology*,
724 22(3):491–495, 2010.
- 725
726 Andrew N Meltzoff. Origins of theory of mind, cognition and communication. *Journal of commu-*
nication disorders, 32(4):251–269, 1999.
- 727
728 A Michotte. *The perception of causality*. Basic Books, 1963.
- 729
730 Patricia H Miller. *Theories of developmental psychology (6th ed.)*. Macmillan Higher Education,
731 2016.
- 732
733 Melanie Mitchell. On crashing the barrier of meaning in artificial intelligence. *AI magazine*, 41(2):
734 86–92, 2020.
- 735
736 Melanie Mitchell. Why ai is harder than we think. *arXiv preprint arXiv:2104.12871*, 2021.
- 737
738 Shunji Mori, Hirobumi Nishida, and Hiromitsu Yamada. *Optical character recognition*. John Wiley
& Sons, Inc., 1999.
- 739
740 Taihei Ninomiya, Atsushi Noritake, Kenta Kobayashi, and Masaki Isoda. A causal role for frontal
741 cortico-cortical coordination in social action monitoring. *Nature communications*, 11(1):5233,
2020.
- 742
743 OpenAI. Gpt-4 technical report. *arXiv preprint arXiv: 2303.08774*, 2023.
- 744
745 Thomas C O’Brien and Bernard J Shapiro. The development of logical thinking in children. *Amer-*
ican Educational Research Journal, 5(4):531–542, 1968.
- 746
747 Alessandro B Palmarini and Melanie Mitchell. Abstract understanding of core-knowledge concepts:
748 Humans vs. llms. In *ICML 2024 Workshop on LLMs and Cognition*, 2024.
- 749
750 Jean Piaget. *The Psychology of Intelligence*. Harcourt, Brace, 1950.
- 751
752 Jean Piaget. *The Origins of Intelligence in Children*. International Universities Press, 1952.
- 753
754 Jean Piaget and Bärbel Inhelder. *The Psychology of the Child*. Basic Books, New York, 1969.
- 755
756 Jean Piaget and Bärbel Inhelder. Intellectual operations and their development. In *Experimental*
Psychology Its Scope and Method: Volume VII (Psychology Revivals), pp. 144–205. Psychology
Press, 2014.

- 756 Jean Piaget and Bärbel Inhelder. *The Child's Construction of Quantities: Conservation and Atom-*
757 *ism*. Psychology Press, 1974.
- 758
- 759 Nicolas Poirel, Grégoire Borst, Grégory Simon, Sandrine Rossi, Mathieu Cassotti, Arlette Pineau,
760 and Olivier Houdé. Number conservation is related to children's prefrontal inhibitory control: an
761 fmri study of a piagetian task. *PloS one*, 7(7):e40802, 2012.
- 762 Guy Politzer. The class inclusion question: a case study in applying pragmatics to the experimental
763 study of cognition. *SpringerPlus*, 5(1):1133, 2016.
- 764
- 765 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agar-
766 wal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya
767 Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint*
768 *arXiv: 2103.00020*, 2021.
- 769 David M. Rosenthal. *The Nature of Mind*. Oxford University Press, New York, 1991.
- 770
- 771 MD Rutherford and DH Brainard. Lightness constancy: A direct test of the illumination-estimation
772 hypothesis. *Psychological Science*, 13(2):142–149, 2002.
- 773
- 774 John Searle. *Minds, brains, and programs*, 1980.
- 775
- 776 John R Searle. The intentionality of intention and action. *Inquiry*, 22(1-4):253–280, 1979.
- 777
- 778 Michael Shayer. Has piaget's construct of formal operational thinking any utility? *British Journal*
779 *of Educational Psychology*, 49(3):265–276, 1979.
- 780
- 781 Richard Shiffrin and Melanie Mitchell. Probing the psychology of ai models. *Proceedings of the*
782 *National Academy of Sciences*, 120(10):e2300963120, 2023.
- 783
- 784 Elizabeth F Shipley. The class-inclusion task: Question form and distributive comparisons. *Journal*
785 *of Psycholinguistic Research*, 8:301–331, 1979.
- 786
- 787 Elizabeth S Spelke. Core knowledge. *American psychologist*, 55(11):1233, 2000.
- 788
- 789 Elizabeth S Spelke and Katherine D Kinzler. Core knowledge. *Developmental science*, 10(1):89–96,
790 2007.
- 791
- 792 Elizabeth S Spelke, Karen Breinlinger, Janet Macomber, and Kristen Jacobson. Origins of knowl-
793 edge. *Psychological review*, 99(4):605, 1992.
- 794
- 795 Elizabeth S Spelke, Gary Katz, Susan E Purcell, Sheryl M Ehrlich, and Karen Breinlinger. Early
796 knowledge of object motion: Continuity and inertia. *Cognition*, 51(2):131–176, 1994.
- 797
- 798 Elizabeth S Spelke, Roberta Kestenbaum, Daniel J Simons, and Debra Wein. Spatiotemporal con-
799 tinuity, smoothness of motion and object identity in infancy. *British journal of developmental*
800 *psychology*, 13(2):113–142, 1995.
- 801
- 802 Christopher Summerfield. *Natural General Intelligence: How understanding the brain can help us*
803 *build AI*. Oxford university press, 2022.
- 804
- 805 Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang,
806 Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal mod-
807 els are in-context learners. *Computer Vision and Pattern Recognition*, 2023. doi: 10.1109/
808 CVPR52733.2024.01365.
- 809
- 804 Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from trans-
805 formers. *arXiv preprint arXiv:1908.07490*, 2019.
- 806
- 807 Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:*
808 *2405.09818*, 2024.
- 809
- 809 Gemini Team. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:*
810 *2312.11805*, 2023.

- 810 Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Niko-
811 lay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher,
812 Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy
813 Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn,
814 Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel
815 Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee,
816 Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra,
817 Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi,
818 Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh
819 Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen
820 Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic,
821 Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models.
822 *arXiv preprint arXiv: 2307.09288*, 2023.
- 823 Alan M Turing. *Computing machinery and intelligence*. Springer, 1950.
- 824 Ross Vasta and Lynn S Liben. The water-level task: An intriguing puzzle. *Current Directions in*
825 *Psychological Science*, 5(6):171–177, 1996.
- 826 Arnaud Viarouge, Olivier Houdé, and Grégoire Borst. The progressive 6-year-old conserver: Nu-
827 merical saliency and sensitivity as core mechanisms of numerical abstraction in a piaget-like
828 estimation task. *Cognition*, 190:137–142, 2019.
- 829 Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural
830 image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern*
831 *recognition*, pp. 3156–3164, 2015.
- 832 Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang,
833 Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv*
834 *preprint arXiv:2311.03079*, 2023.
- 835 Henry M. Wellman. *The Child’s Theory of Mind*. MIT Press, Cambridge, MA, 1992.
- 836 Henry M Wellman, David Cross, and Julianne Watson. Meta-analysis of theory-of-mind develop-
837 ment: The truth about false belief. *Child development*, 72(3):655–684, 2001.
- 838 Heinz Wimmer and Josef Perner. Beliefs about beliefs: Representation and constraining function of
839 wrong beliefs in young children’s understanding of deception. *Cognition*, 13(1):103–128, 1983.
- 840 Gerald A Winer. Class-inclusion reasoning in children: A review of the empirical literature. *Child*
841 *Development*, pp. 309–328, 1980.
- 842 Ludwig Wittgenstein. *The blue and brown books*, 1958.
- 843 Barlow C Wright and Jennifer Smailes. Factors and processes in children’s transitive deductions.
844 *Journal of Cognitive Psychology*, 27(8):967–978, 2015.
- 845 Xiao Xu, Chenfei Wu, Shachar Rosenman, Vasudev Lal, Wanxiang Che, and Nan Duan.
846 Bridgetower: Building bridges between encoders in vision-language representation learning. In
847 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10637–10647,
848 2023.
- 849 Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun
850 Chen. Large language models as optimizers. *arXiv preprint arXiv:2309.03409*, 2023.
- 851 Steven Yantis. Perceived continuity of occluded visual objects. *Psychological Science*, 6(3):182–
852 186, 1995.
- 853 Do-Joon Yi, Nicholas B Turk-Browne, Jonathan I Flombaum, Min-Shik Kim, Brian J Scholl, and
854 Marvin M Chun. Spatiotemporal object continuity in human ventral visual cortex. *Proceedings*
855 *of the National Academy of Sciences*, 105(26):8840–8845, 2008.
- 856 Yuning Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Man Cheung, and Min
857 Lin. On evaluating adversarial robustness of large vision-language models. *Advances in Neural*
858 *Information Processing Systems*, 36, 2024.