

Tracking the Evolution of Multimodal Reasoning on Visual Puzzles

Vernon Y.H. Toh¹, Yew Ken Chia^{2*}, Deepanway Ghosal^{2†}, Soujanya Poria¹

¹Nanyang Technological University

²Singapore University of Technology and Design

Abstract

The releases of OpenAI’s o-[n] series, such as o1, o3, and o4-mini, mark a significant paradigm shift in Large Language Models towards advanced reasoning capabilities. Notably, models like o3 have demonstrated strong performance on benchmarks like the Abstraction and Reasoning Corpus for Artificial General Intelligence (ARC-AGI). However, this benchmark is limited to symbolic patterns, whereas humans often perceive and reason about multimodal scenarios involving both vision and language data. Thus, there is an urgent need to investigate advanced reasoning capabilities in multimodal tasks. To this end, we track the evolution of the GPT-[n] and o-[n] series models (including o1, o3, and o4-mini) and compare them against leading open-source alternatives on challenging multimodal puzzles from PUZZLEVQA and ALGOPUZZLEVQA. Our results reveal that the o-[n] series, particularly later iterations, significantly outperform both the GPT-[n] series and the evaluated open-source models, establishing clear performance tiers. *Nonetheless, despite these substantial advancements, our findings highlight that even leading models face persistent challenges. Difficulties are particularly evident in tasks requiring precise visual perception, robust compositional reasoning, and solving complex algorithmic puzzles. Our analysis further reveals that while closed-source models are primarily bottlenecked by perception, open-source models struggle with both perception and subsequent inductive reasoning, indicating critical areas for future AGI development.* We plan to continuously track new models and update our results in this paper accordingly. All resources used in this evaluation are openly available at <https://github.com/declare-lab/LLM-PuzzleTest>.

Introduction

Recent advances in large language models (LLMs) have demonstrated impressive capabilities in language understanding and generation, as seen in OpenAI’s GPT-[n] series of models (Brown et al. 2020). Yet, true artificial general intelligence (AGI) requires robust reasoning abilities across different modalities (Fei et al. 2021). Simply excelling in text-based tasks is insufficient for agents intended to understand and interact with the rich, multifaceted physical world humans navigate daily. For instance, while models such as OpenAI’s new o-[n] series demonstrate significant performance improvements, sometimes characterized as a

*Work done while at SUTD.

†Now at Deepmind.

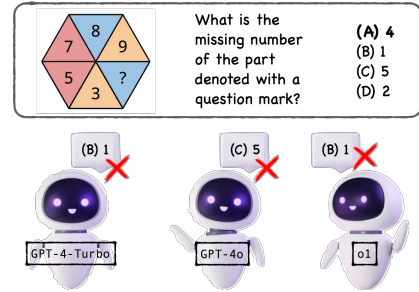


Figure 1: Puzzle from PUZZLEVQA. GPT-4-Turbo, GPT-4o, and o1 all got the puzzle incorrect.

“jumping reasoning curve” on specific benchmarks like the Abstraction and Reasoning Corpus for Artificial General Intelligence (ARC-AGI) (Chollet 2019), the landscape of multimodal reasoning is complex. The current evaluations in this area often focus on symbolic patterns, whereas humans frequently reason over complex data involving both vision and language. Thus, the ability to perceive, understand, and reason about multimodal inputs remains a crucial component of human-like intelligence, deserving urgent investigation.

To this end, puzzles often serve as effective measures of cognitive abilities such as pattern recognition and step-by-step reasoning. Notably, such measures typically do not require specific domain knowledge, allowing individuals from diverse backgrounds to engage with them. One prominent example is Raven’s Progressive Matrices (Raven 1989), a non-verbal assessment tool designed to evaluate abstract reasoning and fluid intelligence. In this test, participants are presented with abstract patterns containing a missing element and must identify the correct piece to complete the pattern.

Thus, inspired by abstract puzzles as measures of intelligence, recent multimodal benchmarks have enabled systematic evaluation across specific cognitive abilities, including visual perception, inductive reasoning, deductive reasoning, and algorithmic problem solving (Chia et al. 2024; Ghosal et al. 2024). These newer benchmarks go beyond simple object recognition or image captioning, probing deeper into how models integrate information across modalities to make inferences or devise multi-step solutions. Compared to previous measures, they require general understanding of spa-

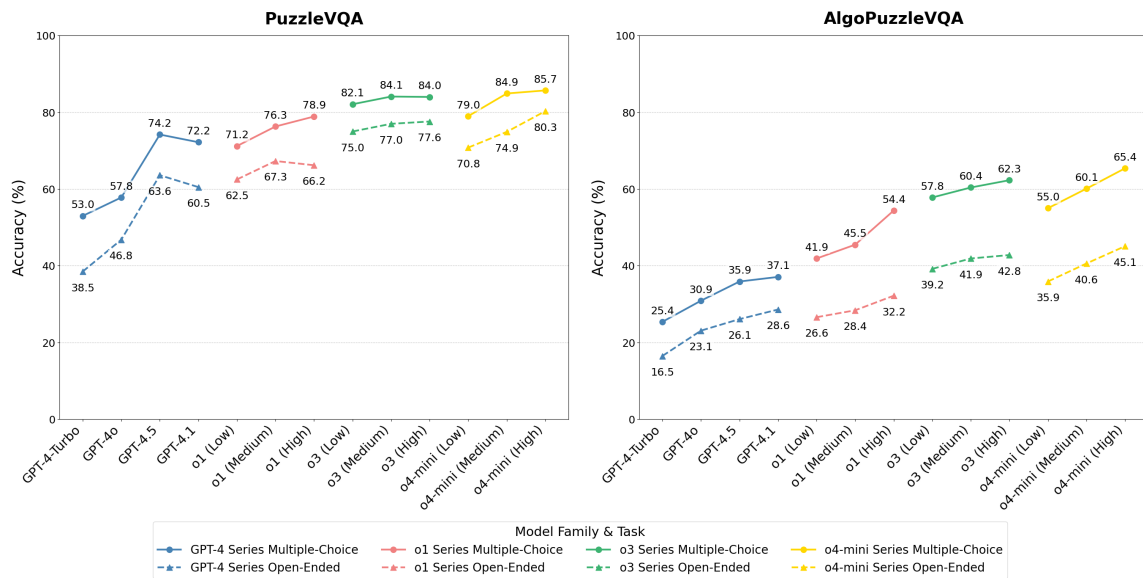


Figure 2: The performance of GPT-[n] and o-[n] series models on PUZZLEVQA and ALGOPUZZLEVQA, illustrating how multimodal reasoning evolves over time with model releases.

tial relationships, pattern recognition, and reasoning across visual and language elements, thus providing a more holistic measure of artificial general intelligence. Our research addresses several key questions: (1) How do current state-of-the-art models, specifically the GPT-[n], o-[n] series and open-source models, perform on diverse visual reasoning tasks? (2) What types of pattern recognition, compositional reasoning, and algorithmic problem-solving are particularly challenging for these models? (3) How can we systematically evaluate and compare different models' multimodal reasoning capabilities?

In our evaluation, we assess the performance of GPT-[n], o-[n] and open-source models on abstract multimodal puzzles from PUZZLEVQA, which primarily test abstract reasoning, and on ALGOPUZZLEVQA, which requires more complex algorithmic approaches. To ensure a comprehensive evaluation, we present the puzzles in both multiple-choice and open-ended question answering formats.

Our findings reveal that while the o-[n] series demonstrates superior performance and scalability, significantly outperforming GPT-[n] models and open-source alternatives across most tasks, even these advanced models encounter substantial difficulties, particularly with complex algorithmic reasoning and fine-grained visual perception. The performance gains, while notable within the o-[n] series, highlight specific bottlenecks rather than uniform leaps across all reasoning types. For example, GPT-[n] models show iterative improvements but consistently lag, struggling with visual perception and more complex compositional tasks, while open-source models exhibit fundamental challenges in both perception and reasoning. This differentiation in capabilities and the persistent challenges, even for leading models, underscore the substantial gap between current artificial intelligence and human-like reasoning abilities. As models continue to rapidly advance

and scale, as suggested by trends like those in Figure 2, this benchmark and the detailed analysis of performance tiers and bottlenecks will serve as a critical indicator of progress toward more robust and generalized artificial intelligence.

The Path to AGI: Why Solving Puzzles and Recognizing Abstract Patterns are Key Milestones

There isn't a single universally accepted definition of Artificial General Intelligence (AGI). However, following (Xu 2024), AGI refers to a system capable of adapting to novel open environments using limited computational resources, guided by specific principles. In contrast to narrow AI, which excels in specific tasks, AGI aims to replicate the broad cognitive abilities of humans, enabling it to perform any intellectual task that a human can (Latif et al. 2024). Achieving AGI involves creating systems that can learn and reason under conditions of insufficient knowledge and resources, continuously adapting to new tasks and environments (Johansson 2024).

Solving puzzles is a significant milestone in the path to AGI because it requires the application of core cognitive skills such as abstract reasoning, problem-solving, and pattern recognition. These skills are essential for a model to generalize knowledge and adapt to new, unseen tasks. The process of puzzle-solving involves several cognitive functions central to human intelligence, including the ability to understand and manipulate abstract concepts, recognize patterns, and apply logical reasoning to arrive at a solution. By evaluating these models on puzzles, we can assess and enhance these cognitive functions within the AI, bringing it closer to human-like intelligence (Estermann et al. 2024).

Wüst et al. (2024) discuss how current models occasionally succeed in identifying discriminative concepts but often fail

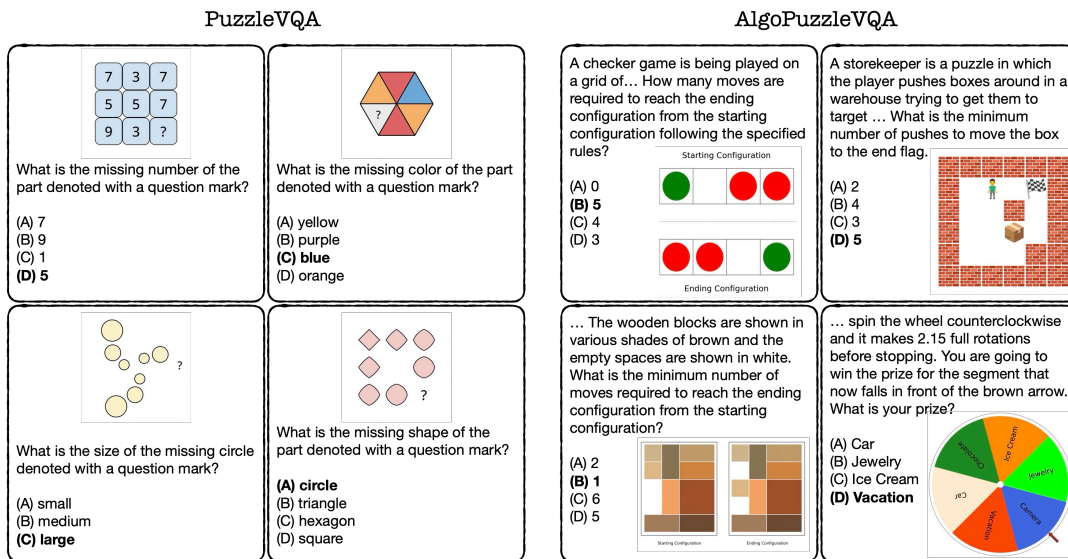


Figure 3: Examples of puzzles from PUZZLEVQA and ALGOPUZZLEVQA.

to understand and reason about visual concepts, indicating a significant limitation in their reasoning abilities. Similarly, Park et al. (2023) explore how models approach the Abstraction and Reasoning Corpus (ARC), revealing both their strengths and limitations in abstract reasoning.

The release of o-[n] models highlights the relationship between computational resources and performance. While increased computational power can enhance performance, true intelligence also requires efficient learning algorithms and the ability to generalize from limited data (Sastry et al. 2024). Simply scaling up resources does not guarantee the emergence of AGI; developing architectures that can learn abstract concepts and apply them across various domains is crucial (Mumuni and Mumuni 2025).

Performance on Existing Puzzle Benchmarks

The performance of GPT-[n] and O-[n] models across various puzzle benchmarks reveals both their strengths and limitations in reasoning and multimodal tasks.

Giadikiaroglou et al. (2024) highlights that, despite advancements, current models often struggle with complex textual rule-based and rule-less puzzles requiring advanced logical inference, revealing a significant gap between their capabilities and human-like reasoning.

On the NPR SUNDAY PUZZLE benchmark (Anderson et al. 2025), o1 achieved a 59% accuracy rate, outperforming models like R1, which demonstrates its proficiency in general reasoning. However, GPT-4o struggled with deep abductive reasoning tasks, scoring only 38% on the TRUE DETECTIVE benchmark (Del and Fishel 2023).

In the REBUS benchmark (Gritsevskiy et al. 2024), GPT-4o performed well with a 42% accuracy rate, but faced difficulties with more complex puzzles, achieving just 7% on the hardest ones. The o1 model also showed limitations on the ENIGMAEVAL benchmark (Wang et al. 2025), with only 7% accuracy on the normal split and 0% on the hard split, high-

lighting its struggle with unstructured and lateral reasoning tasks.

GPT-4V demonstrated moderate success on the PUZZLEVQA (Chia et al. 2024) benchmark with a 46% accuracy rate, showcasing its ability in abstract pattern recognition. However, it scored only 30.3% on the ALGOPUZZLEVQA (Ghosal et al. 2024), revealing challenges in solving complex algorithmic puzzles that require both visual and algorithmic reasoning.

Unlike other puzzle benchmarks, PUZZLEVQA tests a model’s understanding of basic concepts such as colors, shapes, sizes, and spatial relationships. This sets it apart from datasets that focus purely on text or visual recognition. ALGOPUZZLEVQA increases the complexity by introducing algorithmic reasoning, requiring models to solve puzzles that combine visual comprehension with intricate algorithmic thinking. Both datasets allow for deeper analysis, as they are created with predefined ontologies, making it easier to identify reasoning gaps. These qualities make PUZZLEVQA and ALGOPUZZLEVQA the ideal benchmarks for our study, providing a comprehensive foundation for our analysis.

PUZZLEVQA & ALGOPUZZLEVQA

Understanding the capabilities and limitations of large multimodal models in reasoning tasks requires datasets that challenge their cognitive capabilities in nuanced ways. Multimodal puzzles are essential benchmarks for evaluating these models, as they require a unique combination of perception, reasoning, and abstraction. In this study, we employ PUZZLEVQA which emphasizes visual abstract reasoning and pattern recognition, alongside ALGOPUZZLEVQA, which features more complex puzzles that require algorithmic solutions.

PUZZLEVQA consists 10 puzzle categories, with a total of 2,000 test instances. Four of these categories focus on single-concept patterns, such as numbers, colors, sizes, and shapes,

while the remaining six categories emphasize dual-concept patterns, which combine two distinct concepts. ALGOPUZZLEVQA consists of 18 distinct puzzles, each with 100 test instances, resulting in a total of 1,800 test instances. These puzzles cover a wide range of topics, incorporating both visual elements like colors and positions, as well as algorithmic concepts such as Boolean logic and optimization. Figure 3 presents example puzzles from both datasets.

A key contribution of our study is expanding both datasets beyond their original multiple-choice format to include open-ended format. This enhancement enables a more comprehensive and in-depth evaluation of model reasoning capabilities. Additionally, we assess GPT-[n] and o-[n] models, which have not been examined in previous research. By doing so, we extend prior evaluations with a significantly deeper analysis, richer discussions, and extensive case study examples.

Experimental Setup

Evaluation Pipeline

For both multiple-choice and open-ended setups, we use GPT-4o for answer matching, as the varied output formats makes answer extraction a non-trivial task. In both settings, we evaluate performance based on the accuracy of predicting the correct final answer. To validate the reliability of GPT-4o for answer matching, we manually evaluated 200 random outputs with their matching scores. The results showed a classification accuracy of 99%, with the task proving straightforward due to the simplicity of the final answers.

Models

We investigate the performance of GPT-[n] and o-[n] models: (1) GPT-4-Turbo, (2) GPT-4o, (3) GPT-4.5, (4) GPT-4.1, (5) GPT-5.2, (6) o1, (7) o3, (8) o4-mini. We selected these two model series from OpenAI due to their rapid advancements and significant contributions to the field of LLMs. Each version has introduced innovative techniques that have shaped the LLM landscape. For example, GPT-4-Turbo excels in understanding visual inputs, while GPT-4o is optimized for efficient multimodal inference. The o-[n] models, a recent addition, is trained using reinforcement learning to perform complex reasoning before responding. For the o-[n] models, we conducted evaluations across all three reasoning modes: “Low”, “Medium”, “High”. To broaden the scope of our analysis and evaluate the generalizability of our findings, we consider a diverse set of models, including several leading open-source multimodal models such as LLaVA-1.5 13B, Qwen-2-VL 7B, Qwen-2.5-VL 7B, and Qwen3-VL 8B Instruct, along with the closed-source model Gemini 3 Pro Preview. Please note that our study can easily be expanded to other closed-sourced and open-sourced models.

Overall Performance Trends and Model Comparisons

The results presented in Table 1 provide a comprehensive comparison of the current capabilities and limitations of GPT-[n], o-[n], Gemini 3 Pro, and open-source models across PUZZLEVQA and ALGOPUZZLEVQA. Evaluations conducted

under both multiple-choice and open-ended settings reveal clear performance tiers, showcasing notable advancements while also highlighting persistent challenges.

Superior Performance of Reasoning Models. Across both settings, reasoning models specifically the o-[n] series and Gemini 3 Pro consistently outperform standard GPT-[n] models by a significant margin. Within the o-[n] series, performance scales reliably with increased compute: “High” variants outperform “Medium”, which in turn surpass “Low”. Additionally, Gemini 3 Pro demonstrates exceptional capability, particularly in ALGOPUZZLEVQA where it achieves state-of-the-art results (81.3% MC), surpassing the best o-[n] variants. This trend highlights the value of specialized reasoning architectures and inference-time scaling, which effectively convert additional resources into superior logic and perception performance.

GPT-[n] Models Progression and Limitations. Among the GPT-[n] series, newer models like GPT-5.2 and GPT-4.5 generally outperform earlier iterations like GPT-4o and GPT-4-Turbo. However, even these advanced standard models are consistently outperformed by the reasoning models such as o-[n] and Gemini 3 Pro. For instance, on ALGOPUZZLEVQA, GPT-5.2 (54.6%) trails significantly behind the reasoning frontiers. This persistent gap highlights the limitations of standard multimodal training for tasks demanding complex abstract or algorithmic reasoning. A case study in Figure 4 illustrates this disparity, showing how standard models fall short in visual perception compared to their reasoning counterparts.

Fundamental Challenges in Open-Source Models. Open-source models generally score lower than their closed-source counterparts, though recent iterations show progress. While Qwen-3-VL achieves 59.9% on PUZZLEVQA (MC), older models like LLaVa-1.5 lag significantly behind (27.3% vs. o4-mini’s 85.7%). Despite the improvements in the Qwen series, a distinct gap remains between open weights and proprietary frontiers, particularly in both visual perception and logical inference, which we analyze further in our bottleneck analysis.

Impact of Evaluation Settings and Dataset Complexity. The format of evaluation, whether multiple-choice or open-ended, significantly affects model performance. All models consistently perform better on multiple-choice tasks, where the limited answer space enables success through recognition or elimination rather than generative recall. In contrast, open-ended questions, which require precise, unaided responses, result in a drop in accuracy, a drop which is particularly severe for the open-source models. This gap underscores persistent challenges in achieving unconstrained, fine-grained reasoning. A case study in Figure 5 illustrates this: models often fail to accurately infer the time from a clock image, reflecting difficulties with detailed visual perception. Dataset complexity further amplifies these differences. The ALGOPUZZLEVQA dataset, which emphasizes algorithmic, spatial, and multi-step reasoning, consistently proves more difficult than the PUZZLEVQA dataset. All models perform worse on

Task	GPT-4-Turbo	GPT-4o	GPT-4.5	GPT-4.1	GPT-5.2	o1			o3			o4-mini			Gemini 3 Pro	LLaVa-1.5	Qwen-2-VL	Qwen-2.5-VL	Qwen-3-VL
						Low	Med	High	Low	Med	High	Low	Med	High					
PUZZLEVQA																			
Colors	43.0	75.0	88.5	86.0	98.0	88.5	91.5	91.5	98.5	99.0	99.5	100.0	99.5	94.0	22.0	41.5	41.5	85.5	
Numbers	83.0	85.0	97.5	98.5	69.5	96.5	98.0	99.0	99.5	99.0	98.0	98.5	99.0	99.0	20.5	45.0	45.0	80.0	
Shapes	58.5	67.5	74.5	75.0	71.5	59.5	66.5	66.5	78.5	78.0	74.0	61.0	74.5	70.5	77.0	28.0	27.0	27.0	
Size	37.0	44.0	70.5	61.0	66.0	63.5	70.5	77.5	81.0	85.5	79.0	78.5	86.0	83.5	35.5	33.5	33.5	34.5	
Colors & Numbers	62.0	52.0	86.5	82.5	68.5	92.5	98.5	99.0	95.0	97.5	97.5	99.0	100.0	99.5	92.5	26.0	31.5	31.5	70.5
Map Colour	61.5	64.0	78.5	64.0	61.0	69.5	76.5	80.0	87.5	95.5	92.0	79.5	86.5	86.5	84.5	30.0	30.0	30.0	39.0
Colors & Shapes	48.0	57.5	60.5	65.5	38.5	43.0	49.0	50.0	53.5	58.5	57.0	64.5	67.5	72.5	71.5	30.0	52.0	52.0	58.5
Numbers & Shapes	51.5	42.5	43.0	67.0	75.0	82.0	86.5	92.0	88.0	88.0	92.0	87.5	93.0	96.0	99.0	23.0	45.0	45.0	71.5
Numbers & Size	30.5	29.5	47.5	52.0	37.0	45.0	44.0	47.0	48.5	49.0	45.0	42.0	64.0	62.0	73.5	25.0	23.5	23.5	35.5
Size & Shapes	55.0	60.5	95.0	70.0	68.5	72.0	82.5	86.5	91.0	91.5	97.0	78.5	86.5	85.5	88.5	28.0	38.5	38.5	49.5
Average	53.0	57.8	74.2	72.2	67.3	71.2	76.3	78.9	82.1	84.1	84.0	79.0	84.9	85.7	86.3	27.3	36.8	36.8	59.9
ALGOPUZZLEVQA																			
Board Tiling	49.0	52.0	51.0	55.0	49.0	52.0	53.0	48.0	55.0	60.0	60.0	48.0	44.0	52.0	93.0	44.0	54.0	54.0	45.0
Calendar	55.0	61.0	67.0	68.0	70.0	84.0	90.0	90.0	93.0	96.0	96.0	94.0	98.0	99.0	99.0	23.0	25.0	25.0	70.0
Chain Link	4.0	5.0	2.0	2.0	23.0	30.0	35.0	58.0	47.0	65.0	75.0	33.0	45.0	77.0	61.0	26.0	28.0	28.0	26.0
Checker Move	15.0	18.0	36.0	34.0	56.0	41.0	46.0	52.0	61.0	59.0	56.0	82.0	94.0	97.0	94.0	24.0	23.0	23.0	64.0
Clock	19.0	14.0	56.0	68.0	76.0	65.0	76.0	81.0	96.0	97.0	98.0	85.0	92.0	91.0	98.0	19.0	25.0	25.0	55.0
Colour Hue	33.0	25.0	23.0	26.0	38.0	36.0	43.0	43.0	41.0	36.0	47.0	42.0	44.0	47.0	43.0	19.0	14.0	14.0	30.0
Map Colour	20.0	28.0	0.0	0.0	51.0	3.0	1.0	50.0	2.0	4.0	5.0	3.0	4.0	3.0	61.0	28.0	35.0	35.0	44.0
Maze Solve	32.0	39.0	30.0	27.0	38.0	41.0	44.0	50.0	36.0	39.0	39.0	32.0	42.0	49.0	49.0	20.0	27.0	27.0	26.0
Maze Box	34.0	35.0	41.0	37.0	39.0	26.0	24.0	30.0	49.0	46.0	52.0	46.0	47.0	50.0	87.0	34.0	26.0	26.0	27.0
N-Queens	18.0	14.0	3.0	3.0	98.0	18.0	13.0	15.0	34.0	38.0	35.0	39.0	52.0	56.0	96.0	21.0	18.0	18.0	36.0
Number Slide	27.0	32.0	21.0	18.0	83.0	26.0	27.0	88.0	28.0	27.0	30.0	29.0	29.0	29.0	94.0	25.0	34.0	34.0	76.0
Rotten Fruits	25.0	53.0	81.0	72.0	74.0	43.0	48.0	52.0	86.0	84.0	84.0	96.0	99.0	97.0	100.0	25.0	27.0	27.0	71.0
Rubik's Cube	40.0	31.0	61.0	47.0	72.0	64.0	48.0	74.0	78.0	75.0	85.0	77.0	80.0	92.0	88.0	34.0	31.0	32.0	50.0
Think A Dot	37.0	42.0	40.0	49.0	56.0	53.0	57.0	61.0	66.0	74.0	69.0	67.0	82.0	74.0	74.0	36.0	34.0	34.0	56.0
Tower of Hanoi	8.0	19.0	29.0	39.0	42.0	50.0	59.0	63.0	77.0	81.0	79.0	60.0	76.0	85.0	83.0	28.0	32.0	32.0	20.0
Water Jugs	13.0	34.0	37.0	43.0	48.0	48.0	57.0	43.0	80.0	89.0	88.0	67.0	82.0	82.0	84.0	33.0	13.0	13.0	50.0
Wheel of Fortune	15.0	33.0	47.0	51.0	40.0	48.0	55.0	56.0	66.0	65.0	75.0	50.0	58.0	69.0	81.0	36.0	35.0	35.0	50.0
Wood Slide	13.0	21.0	22.0	28.0	30.0	26.0	23.0	25.0	46.0	52.0	48.0	33.0	24.0	36.0	78.0	27.0	23.0	20.0	14.0
Average	25.4	30.9	35.9	37.1	54.6	41.9	45.5	54.4	57.8	60.4	62.3	55.0	60.1	65.4	81.3	27.9	28.0	27.9	45.0
PUZZLEVQA																			
Colors	51.0	72.5	81.0	87.0	96.0	89.5	90.0	80.5	98.5	98.5	98.0	99.5	100.0	99.5	86.5	4.0	44.0	43.5	67.5
Numbers	82.5	84.5	91.0	95.5	54.0	93.0	98.0	96.5	95.0	94.5	93.0	91.5	93.0	76.0	7.5	21.0	21.0	73.5	
Shapes	32.5	51.5	66.5	54.0	47.0	45.0	48.0	54.5	63.0	62.5	57.0	55.0	57.0	48.5	27.5	23.5	23.5	51.5	
Size	19.0	39.0	53.0	59.5	59.5	48.5	55.0	54.5	72.0	75.0	72.0	57.5	69.0	79.5	79.0	15.0	21.0	21.5	32.5
Colors & Numbers	54.5	48.0	71.0	67.5	50.5	93.0	96.5	97.0	90.5	93.5	92.0	97.5	98.0	78.0	9.0	20.5	20.5	63.0	
Map Colour	30.0	45.5	75.5	48.5	40.0	67.5	74.0	75.0	89.0	90.5	90.0	74.0	84.0	88.5	88.5	8.0	15.0	15.0	18.0
Colors & Size	31.5	21.5	35.0	54.5	39.0	35.5	35.5	30.0	48.0	55.5	57.0	49.5	57.5	59.0	50.5	4.5	25.0	25.0	15.5
Numbers & Shapes	31.5	20.0	26.0	40.0	58.5	67.0	78.5	78.0	73.5	78.5	83.5	66.5	71.5	86.0	91.0	3.0	21.5	21.5	49.5
Numbers & Size	24.5	34.5	43.0	42.0	36.0	38.5	44.5	41.5	43.5	43.5	47.0	38.0	45.0	52.0	62.5	18.0	26.0	26.0	28.0
Size & Shapes	28.5	50.5	74.0	56.0	47.5	47.0	53.0	55.0	77.5	78.0	81.0	76.5	78.0	91.0	64.0	5.0	22.5	22.5	19.5
Average	38.5	46.8	63.6	60.5	52.8	62.5	67.3	66.2	75.0	77.0	77.6	70.8	74.9	80.3	72.5	10.2	24.1	24.0	41.9
ALGOPUZZLEVQA																			
Board Tiling	46.0	46.0	56.0	51.0	48.0	52.0	49.0	51.0	58.0	58.0	61.0	48.0	51.0	53.0	82.0	52.0	51.0	51.0	47.0
Calendar	43.0	52.0	64.0	63.0	58.0	71.0	80.0	83.0	85.0	95.0	94.0	90.0	93.0	97.0	96.0	8.0	13.0	13.0	45.0
Chain Link	1.0	3.0	0.0	0.0	1.0	0.0	0.0	1.0	3.0	2.0	4.0	0.0	1.0	6.0	64.0	0.0	0.0	1.0	
Checker Move	3.0	7.0	14.0	18.0	26.0	28.0	28.0	34.0	46.0	49.0	46.0	32.0	52.0	74.0	83.0	0.0	1.0	1.0	43.0
Clock	0.0	3.0	12.0	20.0	44.0	6.0	6.0	6.0	19.0	16.0	25.0	20.0	15.0	13.0	45.0	0.0	1.0	1.0	6.0
Colour Hue	5.0	10.0	15.0	14.0	26.0	14.0	15.0	15.0	20.0	17.0	23.0	12.0	18.0	20.0	28.0	3.0	5.0	5.0	23.0
Map Colour	10.0	22.0	0.0	1.0	44.0	1.0	0.0	21.0	1.0	0.0	1.0	2.0	2.0	1.0	51.0	0.0	6.0	7.0	28.0
Maze Solve	16.0	8.0	9.0	14.0	22.0	22.0	14.0	22.0	17.0	14.0	11.0	10.0	15.0	15.0	39.0	1.0	0.0	0.0	1.0
Maze Box	20.0	23.0	32.0	43.0	31.0	24.0	25.0	23.0	40.0	47.0	47.0	34.0	39.0	41.0	95.0	5.0	2.0	2.0	29.0
N-Queens	17.0	16.0	12.0	8.0	98.0	12.0	12.0	16.0	21.0	23.0	22.0	30.0	31.0	47.0	99.0	2.0	1.0	1.0	13.0
Number Slide	14.0	32.0	19.0	17.0	70.0	26.0	23.0	17.0	27.0	28.0	23.0	27.0	28.0	28.0	98.0	4.0	6.0	6.0	54.0
Rotten Fruits	32.0	53.0	69.0	71.0	59.0	41.0	41.0	43.0	75.0	76.0	77.0	95.0	95.0	96.0	95.0	0.0	3.0	3.0	66.0
Rubik's Cube	32.0	44.0	58.0	48.0	60.0	52.0	53.0	54.0	68.0	63.0	69.0	66.0	72.0	78.0	87.0	9.0	13.0	14.0	34.0
Think A Dot	36.0	41.0	41.0	44.0	51.0	40.0	45.0	32.0	38.0	48.0	38.0	55.0	56.0	55.0	57.0	9.0	8.0	8.0	38.0
Tower of Hanoi	0.0	2.0	10.0	24.0	20.0	31.0	35.0	39.0	60.0	69.0	60.0	60.0	64.0	77.0	77.0	13.0	0.0	0.0	10.0
Water Jugs	8.0	23.0	24.0	34.0	36.0	29.0	43.0	42.0	72.0	85.0	87.0	53.0	66.0	76.0	68.0	0.0	0.0	0.0	43.0
Wheel of Fortune	14.0	29.0	33.0	36.0	32.0	37.0	34.0	31.0	43.0	43.0	49.0	33.0	34.0	38.0	57.0	4.0	10.0	10.0	29.0
Wood Slide	0.0	1.0	1.0	9.0	2.0	0.0	0.0	0.0	16.0	25.0	29.0	0.0	4.0	9.0	79.0	0.0	2.0	3.0	5.0
Average	16.5																		

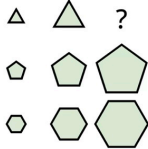
 <p>Question: What is the size of the missing part denoted by a question mark?</p> <p>Options: (A) small (B) large (C) medium</p> <p>MC Ground Truth: (B) large</p>	<p>GPT-4-Turbo MC Output: ... The image shows two rows of shapes: 1. The top row has triangles. 2. The bottom row has pentagons.</p> <p>In each row, the shapes increase in size from left to right. The first shape in each row is the smallest, and the second shape is larger.</p> <p>Given this pattern, the missing shape in the question mark's place should be a pentagon that is larger than the second pentagon but smaller than the third pentagon in the bottom row. ...(C) medium.</p>	<p>GPT-4.5 MC Output: Step-by-step reasoning: 1. Look at each row carefully. Each row has shapes of the same type (triangle, pentagon, hexagon). 2. Each row has three sizes: small, medium, and large. 3. In the first row (triangles), we have small and large shapes, so the missing one is medium. 4. Therefore, the missing triangle should be medium-sized.</p> <p>Answer: (C) medium</p>	<p>o1 (High) MC Output: A quick way to see why it must be "large" is to note that in each row we have a "small" version of the shape (left) and then a "bigger" version (middle). The missing piece at the far right continues that left-to-right increase in size, so it must be the "largest" of the three. Hence, the answer is (B) "large."</p>	<p>o4-mini (High) MC Output: We have three rows (triangles, pentagons, hexagons), and in each row the shapes go from small (left) to medium (center) to large (right). The triangle-row is missing its rightmost (large) triangle. Hence the missing piece must be LARGE. (Answer B)</p>
---	---	---	--	---

Figure 4: A case study on the *Size & Shape* puzzle in PUZZLEVQA reveals that GPT-[n] models exhibit weaker visual perception capabilities compared to o-[n] models.


 <p>Question: Daniel's event is going to start in 1 hour 40 minutes. The current time is shown on the clock. The clock is a standard analog clock without the seconds hand. What will be the time when the event starts?</p> <p>Options: (A) 1:14 (B) 8:09 (C) 4:23 (D) 2:07</p> <p>MC Ground Truth: (C) 4:23</p> <p>OE Ground Truth: 4:23</p>	<p>GPT-4.1 MC Output: Let's solve the problem step by step: ### Step 1: Read the current time on the clock - The "hour hand" is on 2. - The "minute hand" is on 43 (since it's pointing at the 43rd minute mark). So, the current time is "2:43". ### Step 2: Add 1 hour 40 minutes to the current time Start with "2:43". 1. "Add 1 hour:" 2:43 + 1 hour = "3:43" 2. "Add 40 minutes:" 3:43 + 40 minutes = "4:23" ### Step 3: Match with the options "The correct answer is (C) 4:23."</p>	<p>GPT-4.1 OE Output: Let's solve the problem step by step: **Step 1: Read the current time on the clock.** - The hour hand is on 2. - The minute hand is on 9, which means 45 minutes. So, the current time is "2:45". **Step 2: Add 1 hour 40 minutes to the current time.** - 2:45 + 1 hour = 3:45 - 3:45 + 40 minutes = 4:25 **Step 3: Write the final answer.** **When Daniel's event starts, the time will be 4:25."</p>	<p>o4-mini (High) MC Output: First, identify the hands. The longer green hand reaches right out to the minute-ticks, so that's the minute hand; the shorter brown hand is the hour hand. • The green (minute) hand sits three small ticks past the "8," i.e. at 40 + 3 = 43 minutes. • The brown (hour) hand is just past the "2," consistent with 2:43. Now add 1 hour 40 minutes to 2 : 43: 2 : 43 + 1 h = 3 : 43 3 : 43 + 40 min = 4 : 23 Answer: (C) 4:23.</p>	<p>o4-mini (High) OE Output: First, read the clock. The long (green) hand is on the 44-minute tick and the short (brown) hand is a little past the 2, so the current time is 2 : 44. Now add 1 hour 40 minutes: • 2 : 44 + 1 hour = 3 : 44 • 3 : 44 + 40 minutes = 4 : 24 So Daniel's event will begin at 4 : 24.</p>
---	--	---	--	--

Figure 5: A case study on the *Clock* puzzle in ALGOPUZZLEVQA, evaluated in both multiple-choice and open-ended settings, shows that both GPT-[n] and o-[n] models still lack fine-grained visual perception capabilities.

puzzles in the dataset, demonstrating strength not only in basic visual perception but also in understanding more complex concept combinations.

Performance on ALGOPUZZLEVQA

Significant Challenges in Algorithmic and Spatial Reasoning The results on ALGOPUZZLEVQA underscore that complex algorithmic and spatial reasoning remains a challenge for current models. Overall scores are considerably lower across all models compared to PUZZLEVQA. GPT-[n] models struggle significantly with certain tasks such as *Chain Link* and *Map Colour* in the open-ended setting, often scoring near zero, indicating a fundamental difficulty in understanding the underlying principles or performing necessary reasoning steps for these puzzle types.

Gains by o-[n] Models on Complex Algorithmic Tasks

Despite the overall difficulty, the o-[n] models demonstrate improvements on several complex algorithmic tasks within ALGOPUZZLEVQA. For instance, on *Calendar* related puzzles, o4-mini (High) achieves scores in the high 90s (MC: 99.0%, OE: 97.0%), a huge difference compared to GPT-4.5's 67.0% (MC) and 64.0% (OE). Similarly, on *Checker Move*,

GPT-4.5 scores 36.0% (MC), while o4-mini (High) reaches 97.0% (MC). These improvements suggest that the o-[n] models have developed more advanced capabilities for sequential decision-making, state tracking, and applying learned algorithmic patterns, enabling them to tackle certain classes of complex problems with significantly greater effectiveness.

Challenges in Highly Abstract or Combinatorics Puzzles Puzzles such as *Map Colour* and *Chain Link* particularly in their open-ended formats consistently produce very low scores, with o4-mini High scoring only 1.0% and GPT-4.5 0.0% on *Map Colour*, and 6.0% and 0.0% respectively on *Chain Link*. These problems typically require abstract graph reasoning, intricate constraint satisfaction, or planning over a vast combinatorial space where even the most advanced models still struggle. Similarly, *Wood Slide* remains a challenge in its open-ended format for all models.

Average Performance Summary for ALGOPUZZLEVQA

The average performance highlights both the dataset's difficulty and the varying capabilities of different models. In the multiple-choice setting, the best-performing GPT model (GPT-4.1) achieved an average accuracy of 37.1%, whereas o4-mini (High) significantly outperformed it with 65.4%. A

Task	GPT-4-Turbo	GPT-4o	GPT-4.5	GPT-4.1	o1			o3			o4-mini			Qwen-2-VL	Qwen-2.5-VL	LLaVa-1.5
					Low	Med	High	Low	Med	High	Low	Med	High			
ORIGINAL																
Colors	51.0	72.5	81.0	87.0	89.5	90.0	80.5	98.5	98.0	99.5	100.0	99.5	44.0	43.5	4.0	
Numbers	82.5	84.5	91.0	95.5	93.0	98.0	96.5	95.0	94.5	93.0	93.0	91.5	93.0	21.0	21.0	7.5
Shapes	32.5	51.5	66.5	54.0	45.0	48.0	54.5	63.0	62.0	62.5	57.0	55.0	57.0	23.5	23.5	27.5
Size	19.0	39.0	53.0	59.5	48.5	55.0	54.5	72.0	75.0	72.0	57.5	69.0	79.5	21.0	21.5	15.0
Colors & Numbers	54.5	48.0	71.0	67.5	93.0	96.5	97.0	90.5	93.5	92.0	97.0	97.5	98.0	20.5	20.5	9.0
Colors & Shapes	30.0	45.5	75.5	48.5	67.5	74.0	75.0	89.0	90.5	90.0	74.0	84.0	88.5	15.0	15.0	8.0
Colors & Size	31.5	21.5	55.0	54.5	35.5	35.5	30.0	48.0	55.5	57.0	49.5	57.5	59.0	25.5	25.0	4.5
Numbers & Shapes	31.5	20.0	26.0	40.0	67.0	78.5	78.0	73.5	78.5	83.5	66.5	71.5	86.0	21.5	21.5	3.0
Numbers & Size	24.5	34.5	43.0	42.0	38.5	44.5	41.5	43.5	43.5	47.0	38.0	45.0	52.0	26.0	26.0	18.0
Size & Shapes	28.5	50.5	74.0	56.0	47.0	53.0	55.0	77.5	78.0	81.0	76.5	78.0	91.0	22.5	22.5	5.0
Average	38.5	46.8	63.6	60.5	62.5	67.3	66.2	75.0	77.0	77.6	70.8	74.9	80.3	24.1	24.0	10.2
ORIGINAL + VISUAL PERCEPTION																
Colors	75.0	80.0	92.0	95.5	99.0	100.0	94.0	100.0	100.0	100.0	100.0	100.0	64.0	64.0	23.0	
Numbers	77.0	88.5	98.0	94.5	98.5	99.5	98.0	95.5	92.5	91.5	97.5	92.5	92.0	25.5	25.5	5.0
Shapes	71.5	63.5	66.0	63.5	45.5	50.5	55.5	68.0	63.5	65.5	54.5	54.0	54.0	52.0	52.0	27.0
Size	64.5	62.5	93.5	97.0	93.5	98.0	98.0	98.5	99.5	99.0	99.0	99.5	100.0	67.0	67.0	45.5
Colors & Numbers	67.0	52.0	94.5	83.5	97.0	95.0	95.0	95.5	96.5	94.5	96.5	98.0	98.5	32.5	32.0	13.5
Colors & Shapes	81.0	77.5	75.5	70.0	86.5	82.5	81.5	86.5	88.5	86.5	82.5	83.0	89.5	8.0	8.0	23.0
Colors & Size	53.5	78.0	84.5	80.5	96.0	99.0	99.0	96.5	99.0	97.0	93.0	98.0	99.5	20.0	20.0	12.5
Numbers & Shapes	29.5	33.5	26.5	35.0	88.5	89.0	86.0	78.0	82.0	83.0	79.5	86.5	90.5	16.5	16.5	14.5
Numbers & Size	70.0	73.0	79.0	68.5	77.0	82.0	81.5	81.0	82.5	79.5	69.0	80.0	77.5	28.5	28.5	12.5
Size & Shapes	97.5	92.5	92.5	88.0	92.0	98.0	98.0	93.5	94.5	96.5	90.0	98.5	98.5	57.5	58.0	31.5
Average	68.7	70.1	80.2	77.6	87.3	89.3	88.7	89.3	89.8	89.3	86.2	89.0	90.0	37.1	37.1	20.8
ORIGINAL + VISUAL PERCEPTION + INDUCTION																
Colors	97.0	92.0	100.0	100.0	100.0	100.0	99.0	99.5	100.0	100.0	100.0	100.0	50.5	50.5	40.5	
Numbers	98.5	99.5	98.5	100.0	98.0	96.5	97.0	100.0	100.0	99.5	99.5	99.0	98.5	85.0	85.0	31.0
Shapes	97.5	97.5	100.0	93.0	97.5	99.0	100.0	88.5	88.0	91.0	91.5	91.0	94.0	55.0	55.0	49.0
Size	95.5	96.5	99.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.5	100.0	74.0	74.0	54.0
Colors & Numbers	89.5	89.5	98.5	99.5	100.0	100.0	100.0	99.5	100.0	100.0	99.5	99.0	99.5	60.0	60.0	47.0
Colors & Shapes	64.5	77.0	88.5	87.5	88.0	87.0	89.5	95.5	93.5	97.5	90.5	94.5	98.0	17.0	17.0	30.0
Colors & Size	75.5	94.5	93.5	86.5	92.5	92.5	94.0	98.0	98.0	99.0	99.0	98.0	99.0	34.5	34.5	45.5
Numbers & Shapes	84.5	85.5	78.5	84.0	91.0	89.5	91.0	90.0	88.5	87.5	93.0	93.0	93.5	46.0	46.0	11.5
Numbers & Size	63.0	73.5	78.5	87.0	74.0	72.5	77.5	80.0	82.0	81.5	63.5	57.0	60.5	43.0	43.0	12.5
Size & Shapes	93.0	92.5	90.0	91.5	97.5	99.0	99.5	95.5	97.5	95.0	98.0	98.5	99.5	68.0	68.0	64.0
Average	85.8	89.8	92.5	92.9	93.8	93.6	94.8	94.7	94.8	95.1	93.5	93.0	94.2	53.3	53.3	38.5

Table 2: Bottleneck analysis of GPT-[n] and o-[n] models on PUZZLEVQA in the open-ended setting. **Original** refers to our default setting where only a question and an image are provided as input. To reveal the specific multimodal reasoning bottlenecks, we progressively inject ground-truth explanations in the input for visual perception and inductive reasoning.

similar pattern holds for open-ended setting where GPT-4.1 reached 28.6%, while o4-mini (High) attained 45.1%. These results show that while the o-[n] models have made notable progress in tackling algorithmic puzzles, overall performance remains modest, indicating that this domain is far from being solved.

Bottleneck Analysis

Our bottleneck analysis on PUZZLEVQA (Table 2) reveals that providing ground-truth visual perception shows performance gains across all models (GPT-4.5: 63.6% to 80.2%; o4-mini (High): 80.3% to 90.0%). This consistent improvement, observed across different models, clearly isolates visual perception as a primary bottleneck in the baseline. Subsequently, even with perfect visual perception assumed, injecting ground-truth inductive reasoning provides further significant performance increases (GPT-4.5: 80.2% to 92.5%; o4-mini (High): 90.0% to 94.2%), underscoring inductive reasoning as a separate and significant limiting factor. Overall, the progressive injection from the “Original” setting to “Original + Visual Perception + Induction” quantifies the substantial impact of these distinct components for closed-source models (GPT-4.5 total gain from 63.6% to 92.5%), indicating that while these models possess strong underlying reasoning, their performance is notably constrained by difficulties in both visual perception and induction. In contrast, our analysis of open-source models reveals two fundamental bottlenecks rooted in both perception and reasoning. For example, LLaVa-1.5 improves only from 10.2% to 20.8% when given perception hints. Its performance increases further to 38.5%

when induction hints are also provided. This progression shows that even with complete visual information, the model still struggles to infer the underlying pattern, indicating limitations in its core reasoning capabilities. The combination of weak perception and weak inductive reasoning helps explain the overall poor performance of these models.

Conclusion

This study assesses the multimodal reasoning capabilities of the GPT-[n], o-[n], and open-source model series using the PUZZLEVQA and ALGOPUZZLEVQA benchmarks. Our findings establish clear performance tiers, revealing a significant advancement with later o-[n] iterations, which consistently and scalably outperform GPT-[n] models, which in turn significantly outperform the open-source alternatives. Across all models, performance was notably stronger in multiple-choice settings compared to open-ended ones. The o-[n] models demonstrated distinct strengths in specific algorithmic tasks (e.g., *Calendar*) and basic concept recognition (e.g., *Colors*). However, persistent challenges remain, particularly in compositional reasoning (e.g., *Colors & Shapes*), fine-grained visual perception (especially *Size*), and complex algorithmic puzzles. Our bottleneck analysis highlights that while closed-source models are primarily limited by perception, open-source models face severe challenges in both perception and subsequent inductive reasoning. This emphasizes that, despite architectural progress across the board, substantial improvements in these core areas are essential to closing the gap toward robust, human-like multimodal intelligence.

References

- Anderson, C. J.; Biswas, J.; Boruch-Gruszecki, A.; Cassano, F.; Feldman, M. Q.; Guha, A.; Lucchetti, F.; and Wu, Z. 2025. PhD Knowledge Not Required: A Reasoning Challenge for Large Language Models. arXiv:2502.01584.
- Brown, T.; Mann, B.; Ryder, N.; et al. 2020. Language Models are Few-Shot Learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.
- Chia, Y. K.; Han, V. T. Y.; Ghosal, D.; Bing, L.; and Poria, S. 2024. PuzzleVQA: Diagnosing Multimodal Reasoning Challenges of Language Models with Abstract Visual Patterns. arXiv:2403.13315.
- Chollet, F. 2019. On the Measure of Intelligence. arXiv:1911.01547.
- Del, M.; and Fishel, M. 2023. True Detective: A Deep Abductive Reasoning Benchmark Undoable for GPT-3 and Challenging for GPT-4. arXiv:2212.10114.
- Estermann, B.; Lanzendörfer, L. A.; Niedermayr, Y.; and Wattenhofer, R. 2024. PUZZLES: A Benchmark for Neural Algorithmic Reasoning. arXiv:2407.00401.
- Fei, N.; Lu, Z.; Gao, Y.; Yang, G.; Huo, Y.; Wen, J.; Lu, H.; Song, R.; Gao, X.; Xiang, T.; Sun, H.; and Wen, J. 2021. Towards artificial general intelligence via a multimodal foundation model. *Nature Communications*, 13.
- Ghosal, D.; Han, V. T. Y.; Ken, C. Y.; and Poria, S. 2024. Are Language Models Puzzle Prodigies? Algorithmic Puzzles Unveil Serious Challenges in Multimodal Reasoning. arXiv:2403.03864.
- Giadikiaroglou, P.; Lymperaious, M.; Filandrianos, G.; and Stamou, G. 2024. Puzzle Solving using Reasoning of Large Language Models: A Survey. arXiv:2402.11291.
- Gritsevskiy, A.; Panickssery, A.; Kirtland, A.; Kauffman, D.; Gundlach, H.; Gritsevskaya, I.; Cavanagh, J.; Chiang, J.; Roux, L. L.; and Hung, M. 2024. REBUS: A Robust Evaluation Benchmark of Understanding Symbols. arXiv:2401.05604.
- Johansson, R. 2024. Machine Psychology: Integrating Operant Conditioning with the Non-Axiomatic Reasoning System for Advancing Artificial General Intelligence Research. arXiv:2405.19498.
- Latif, E.; Mai, G.; Nyaaba, M.; Wu, X.; Liu, N.; Lu, G.; Li, S.; Liu, T.; and Zhai, X. 2024. AGI: Artificial General Intelligence for Education. arXiv:2304.12479.
- Mumuni, A.; and Mumuni, F. 2025. Large language models for artificial general intelligence (AGI): A survey of foundational principles and approaches. arXiv:2501.03151.
- Park, J.; Im, J.; Hwang, S.; Lim, M.; Ualibekova, S.; Kim, S.; and Kim, S. 2023. Unraveling the ARC Puzzle: Mimicking Human Solutions with Object-Centric Decision Transformer. arXiv:2306.08204.
- Raven, J. 1989. The Raven Progressive Matrices: A Review of National Norming Studies and Ethnic and Socioeconomic Variation within the United States. *Journal of Educational Measurement*, 26(1): 1–16.
- Sastry, G.; Heim, L.; Belfield, H.; Anderljung, M.; Brundage, M.; Hazell, J.; O’Keefe, C.; Hadfield, G. K.; Ngo, R.; Pilz, K.; Gor, G.; Bluemke, E.; Shoker, S.; Egan, J.; Trager, R. F.; Avin, S.; Weller, A.; Bengio, Y.; and Coyle, D. 2024. Computing Power and the Governance of Artificial Intelligence. arXiv:2402.08797.
- Wang, C. J.; Lee, D.; Menghini, C.; Mols, J.; Doughty, J.; Khoja, A.; Lynch, J.; Hendryx, S.; Yue, S.; and Hendrycks, D. 2025. EnigmaEval: A Benchmark of Long Multimodal Reasoning Challenges. arXiv:2502.08859.
- Wüst, A.; Tobiasch, T.; Helff, L.; Dhimi, D. S.; Rothkopf, C. A.; and Kersting, K. 2024. Bongard in Wonderland: Visual Puzzles that Still Make AI Go Mad? arXiv:2410.19546.
- Xu, B. 2024. What is Meant by AGI? On the Definition of Artificial General Intelligence. arXiv:2404.10731.