PuzzleVQA: Diagnosing Multimodal Reasoning Challenges of Language Models with Abstract Visual Patterns

Yew Ken Chia[1,2*], Vernon Toh Yan Han[1], Deepanway Ghosal[1], Lidong Bing[2], Soujanya Poria[1]

[1] Singapore University of Technology and Design, [2] DAMO Academy, Alibaba Group, Singapore

: https://github.com/declare-lab/LLM-PuzzleTest
: https://puzzlevqa.github.io/

Abstract

Large multimodal models extend the impressive capabilities of large language models by integrating multimodal understanding abilities. However, it is not clear how they can emulate the general intelligence and reasoning ability of humans. As recognizing patterns and abstracting concepts are key to general intelligence, we introduce PuzzleVQA, a collection of 2000 puzzle instances based on abstract patterns. With this dataset, we evaluate large multimodal models with abstract patterns based on fundamental concepts, including colors, numbers, sizes, and shapes. Through our experiments on state-of-the-art large multimodal models, we find that they are not able to generalize well to simple abstract patterns. Notably, GPT-4V achieves a score of 46.4% on single-concept puzzles, which shows that state-of-the-art models struggle on our dataset. To diagnose the reasoning challenges in large multimodal models, we progressively guide the models with our ground truth reasoning explanations for visual perception, inductive reasoning, and deductive reasoning. Our systematic analysis finds that the main bottlenecks of GPT-4V are weaker visual perception and inductive reasoning abilities. Through this work, we hope to shed light on the limitations of large multimodal models and how they can better emulate human cognitive processes in the future[1].

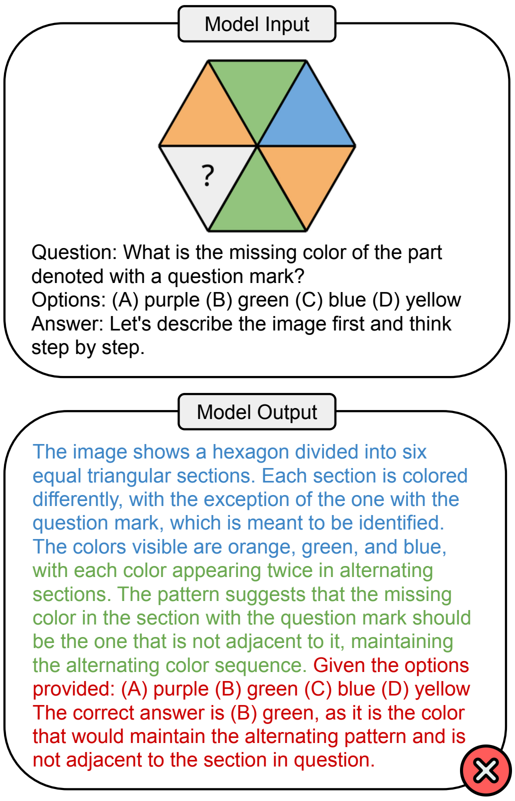

Figure 1: An example question which involves the color concept in PuzzleVQA, and an incorrect answer from GPT-4V. There are generally three stages that can be observed in the solving process: visual perception (blue), inductive reasoning (green), and deductive reasoning (red). Here, the visual perception was incomplete, causing a mistake during deductive reasoning.

1 Introduction

Rapid advances in large language models have demonstrated remarkable capabilities across diverse language tasks and applications (Bubeck et al., 2023; Brown et al., 2020; Touvron et al., 2023b). To enable more general capabilities, large multimodal models were introduced by integrating large language models with multimodal understanding (Yue et al., 2023; Yang et al., 2023; OpenAI, 2023). However, it is not clear how large multimodal models can emulate the general intelligence and reasoning ability of humans (Qiu et al., 2024). Specifically, we aim to explore how large multimodal models can emulate cognitive processes to perceive and interpret information, extrapolate from observations to broader generaliza-

[*] Yew Ken Chia is under the Joint Ph.D. Program between DAMO Academy and the Singapore University of Technology and Design.

[1]Our data and code are released at https://github.com/declare-lab/LLM-PuzzleTest.

Findings of the Association for Computational Linguistics ACL 2024, pages 16259–16273
August 11-16, 2024 ©2024 Association for Computational Linguistics

tions, and apply general principles to solve specific problems (Piaget, 1976). Furthermore, we are interested in understanding how well the models can reason about fundamental concepts such as numbers, colors, shapes, and size (Tong et al., 2024; Sharma et al., 2024).

As pattern recognition and and abstracting concepts are at the heart of general intelligence (Tenenbaum, 2018; Carey, 2000; Cole, 1996), we believe that abstract patterns are a suitable testbed for evaluating reasoning ability in large multimodal models. As shown in Figure 1, abstract patterns enable us to focus on one or more abstract concepts, and decompose the multimodal reasoning process into several stages that mimic human cognitive processes (Piaget, 1976). Firstly, the model requires visual perception to understand and interpret the abstract image in the input. Secondly, the model requires inductive reasoning to relate the observations shown and form a hypothesis for the underlying pattern. Thirdly, the model requires deductive reasoning to apply the general principle of the pattern to solve the specific problem at hand. While the abstract patterns may seem simple, we surprisingly find that even advanced large multimodal models such as Gemini Pro (Gemini Team, 2023) and GPT-4V (OpenAI, 2023) struggle to understand them.

Puzzles are problems that require ingenuity and creativity to solve, and they can serve as valuable tools for cognitive development and assessment (Zhang et al., 2019). Hence, we propose the PUZZLEVQA dataset to systematically evaluate and diagnose the reasoning challenges in large multimodal models. Our dataset consists of diverse multimodal puzzles that focus on abstract patterns with fundamental concepts including numbers, colors, shapes, and size. We design and automatically construct the dataset through multimodal templates, enabling us to generate large numbers of puzzles without costly human annotation (Ding et al., 2022). To support interpretability and systematic investigation of reasoning challenges in multimodal models, we also construct the ground truth reasoning explanations for each puzzle. Compared to existing datasets for visual question answering, PUZZLEVQA focuses specifically on how large multimodal models can mimic general cognitive processes such as inductive and deductive reasoning. As we focus on how models can generalize to novel problems, similar to fluid intelligence in humans (Cattell, 1963), our dataset in the abstract domain poses challenges for existing models without requiring extensive world knowledge.

Through our investigation of leading large multimodal models, we find that existing models are not able to generalize well to simple abstract patterns. Notably, GPT-4V achieves a score of 46.4% on single-concept puzzles, which shows that state-of-the-art models struggle on our dataset. Our analysis reveals that its main bottlenecks are weaker visual perception and inductive reasoning abilities. Hence, our main contributions include:

1. To investigate the cognitive and reasoning abilities of large multimodal models, we propose to leverage abstract patterns.
2. We introduce PUZZLEVQA, an automatically generated and diverse dataset of 2000 multimodal samples with reasoning explanations.
3. Our experiments show that even advanced large multimodal models do not generalize well to abstract patterns, and we show how to identify their reasoning bottlenecks.

2 Background: Cognitive Theories

To understand how large multimodal models can better mimic human thought processes and general intelligence, we first ground our study with relevant cognitive theories.

2.1 Fluid and Crystallized Intelligence

The Cattell-Horn theory (Cattell, 1963) of cognitive abilities distinguishes between two types of intelligence: fluid intelligence, which involves the ability to solve novel problems without relying on previously acquired knowledge, and crystallized intelligence, which involves the use of knowledge, skills, and experience. Fluid intelligence in humans could parallel large multimodal models' ability to solve new, unseen problems through pattern recognition and problem-solving strategies. On the other hand, crystallized intelligence could be akin to how the models leverage accumulated world knowledge from training data to understand and interact with the world (Sumers et al., 2023). As many works have focused on how models can leverage specialized knowledge (Yue et al., 2023), we instead focus on how they may emulate fluid intelligence to solve novel problems through abstract patterns.

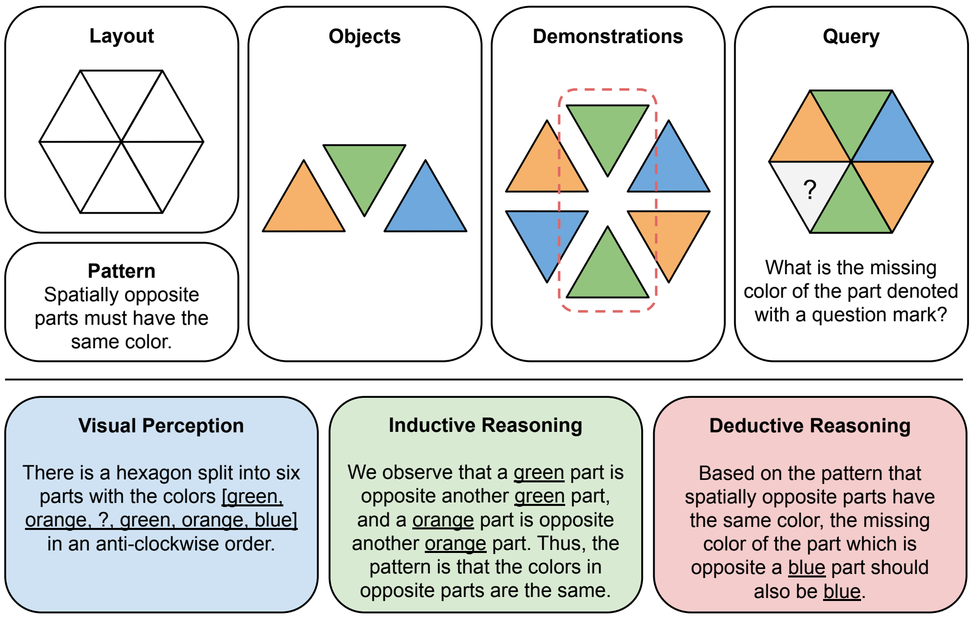

Figure 2: Illustration example of components (top) and reasoning explanations (bottom) for abstract puzzles in PUZZLEVQA. To construct each puzzle instance, we first define the layout and pattern of a multimodal template, and populate the template with suitable objects that demonstrate the underlying pattern. For interpretability, we also construct ground truth reasoning explanations to interpret the puzzle and explain the general solution stages.

2.2 Cognitive Development

Piaget's Stages of Cognitive Development (Piaget, 1976) can provide a framework for progressing from basic sensory experiences to complex abstract reasoning and problem-solving. While we note that the large multimodal models do not develop in the same organic and experiential manner as humans, we are guided to explore how the models can emulate different stages of cognitive abilities. Concretely, through abstract patterns, we can evaluate how the models perceive multimodal information, reason inductively to extrapolate from observations to broader generalizations, and apply general principles to deduce the solution for specific problems.

Sensorimotor. This stage underpins visual perception, where individuals learn to coordinate sensory experiences through interactions with the environment. To emulate this cognitive stage, we would expect models to identify simple shapes or colors but lack higher-level reasoning. Hence, we set the foundation for later stages by exploring abstract patterns based on fundamental concepts including colors, numbers, shapes, and size.

Preoperational. At this stage, individuals develop symbolic thinking, which is crucial for understanding representations in visual contexts and beginning to engage in simple, inductive reasoning processes. Models that mimic this stage should be able to perform basic reasoning about objects or concepts, but with limited understanding of abstract relationships or performing logical operations.

Concrete Operational. This stage is closely related to inductive reasoning, as individuals learn to think logically about concrete events and solve problems based on visible patterns and relationships. We would expect models that are analogous to this stage to have the ability to draw logical conclusions from specific instances and start to apply these conclusions to solve problems. Hence, we consider inductive reasoning as an integral part of understanding abstract patterns.

Formal Operational. This stage is essential for deductive reasoning and abstract thinking, allowing individuals to hypothesize and think about theoretical scenarios, which are skills necessary for solving complex problems. At this stage, we would expect comparable models to effectively induce general principles or hypotheses from observations and logically deduce specific outcomes, even in abstract or novel contexts. Thus, we consider deductive reasoning as critical to solving abstract problems.

3 PuzzleVQA Dataset

Despite the impressive capabilities of large multimodal models, we do not fully understand how they solve multimodal problems through reasoning.

Specifically, we focus on how well they can interpret multimodal inputs, form generalizations from observations, and apply the general principles to solve specific cases. Furthermore, they may reason differently about fundamental concepts such as numbers, colors, shapes, and size. Hence, we propose PUZZLEVQA, a diverse collection of abstract pattern puzzles to diagnose reasoning challenges in multimodal models. The dataset is automatically constructed through multimodal templates, and includes reasoning explanations for interpretability.

3.1 Puzzle Components

As shown in Figure 2, each puzzle in our dataset is formulated with the following main components:

1. **Objects:** The conceptual elements that interact within the puzzle, such as numbers, colors, shapes, and size.
2. **Layout:** The spatial arrangement of objects that provides visual context.
3. **Pattern:** The relationship that governs the interaction amongst objects. For example, a pattern may be that spatially opposite parts must have the same color.
4. **Demonstrations:** Multiple instances of interacting objects that collectively represent the underlying pattern. Without demonstrations, the pattern would become ambiguous.
5. **Query:** The natural language question that directs the multimodal model how to solve the puzzle by determining the missing object.

3.2 Design Considerations

We have three main design considerations for each puzzle in our dataset:

Simplicity. As the focus is on evaluating how large multimodal models reason about fundamental abstract concepts, we do not deliberately make the puzzles more complex than necessary. We also aim to make the underlying patterns straightforward, without requiring extensive world knowledge or advanced theories.

Correctness. To avoid potentially noisy annotations, we use an automatic approach with multimodal templates to generate each puzzle. For instance, given a visual layout and pattern of a template, we can automatically populate the template with the appropriate objects that demonstrate the pattern. As each puzzle instance is created based on the specific rules in the template, we can ensure that they do not contain annotation mistakes.

Diversity. To investigate the multimodal reasoning capabilities across diverse abstract concepts, we construct puzzles based on four main concepts: numbers, colors, shapes, and size. Furthermore, to evaluate how well the models can relate to multiple concepts, we design both single-concept and dual-concept puzzles.

3.3 Puzzle Construction

Multimodal Templates. To construct each abstract puzzle, we leverage multimodal templates based on fundamental concepts including numbers, colors, shapes, and size. Following the formulation and design considerations previously discussed, we first define the layout and abstract pattern for the puzzle. Each template can be randomly populated with the specific objects to represent the underlying pattern through demonstrations, forming a specific puzzle instance. For example, to construct a color-based puzzle instance shown in Figure 2, we focus on the concept of colors and define the layout as a hexagonal arrangement of six parts, with the abstract pattern that spatially opposite parts must have the same color. Thereafter, the template can be randomly populated to satisfy the pattern with colors from a predefined list of possible colors. Lastly, the query is constructed based on the fundamental concepts in the abstract pattern. To demonstrate our puzzle generation pipeline, we include a detailed implementation in Appendix A.4, based on the puzzle in Figure 1.

Reasoning Explanations. To ensure that each abstract puzzle can be easily understood, we also construct reasoning explanations based on the three problem solving stages: image descriptions for visual perception, pattern explanations for inductive reasoning, and deductive reasoning steps. Specifically, we leverage textual templates that can be populated with details from the specific puzzle instance, as shown in Figure 2. In our experiments in Section 5, this enables us to identify reasoning bottlenecks by progressively providing the explanations of each stage to the model.

3.4 Multiple-Choice Format

While we use straightforward objects in our puzzles, there may be a degree of ambiguity in the an-

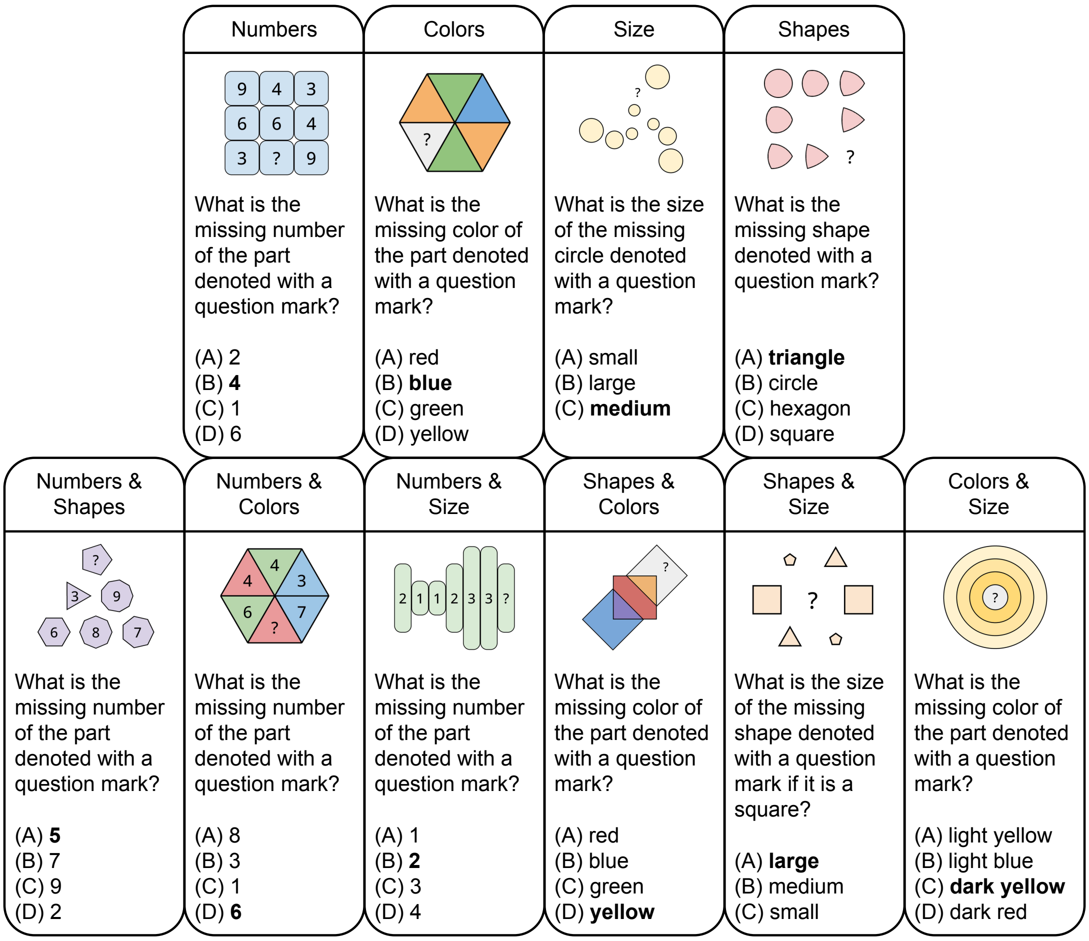

Figure 3: Taxonomy of abstract puzzles in PUZZLEVQA with sample questions, based on fundamental concepts such as colors and size. To enhance diversity, we introduce both single-concept and dual-concept puzzles.

swer regarding specific colors or sizes. Hence, we standardize the puzzle format as multiple-choice questions, where all questions are provided with four options, with the exception of three options for size (small, medium, and large). To generate the incorrect choices for each question, we use heuristics including randomly sampling numbers within the same magnitude of the answer, and further details can be found in the supplementary material. We use the standard accuracy metric for evaluation.

3.5 Dataset Analysis

To ensure that the dataset contains diverse abstract patterns, we provide a taxonomy of 10 puzzle categories based on fundamental concepts including numbers, colors, shapes, and size. As shown in Figure 3, there are four categories of single-concept patterns. To extend the depth of PUZZLEVQA, we also include dual-concept patterns, which would require models to relate two concepts in order to solve the puzzle. Within each category, we design two multimodal templates that can each be used to generate many unique puzzle instances. The full list of puzzle templates and examples of more puzzle instances can be found in the supplementary material. To maintain a reasonable dataset size for evaluating large multimodal models, we generate 100 unique puzzle instances from each template. Thus, there are 2000 test instances in PUZZLEVQA in total. We conducted an analysis in Appendix A.5 and found that the chosen dataset size is large enough to be relatively robust to experimental variance.

3.6 Implementation Details

We utilize Python code along with packages like Pillow[2] to automatically generate puzzles. Leveraging these tools, we are able to create many different unique puzzle images and text questions for each given puzzle type by augmenting the base template and objects in the image. Example code snippets to generate the puzzles are included in the supplemen-

[2] https://pypi.org/project/pillow/

	Numbers	Colors	Size	Shapes	Average
Random Baseline	25.0	25.0	33.3	25.0	27.1
Qwen-VL-Chat	25.0	22.0	32.5	34.0	28.4
LLaVA-13B	26.0	29.0	29.0	26.0	27.5
Gemini Pro	32.5	32.0	33.5	40.0	34.5
Claude 3 Opus	47.0	32.5	33.5	**44.5**	39.4
GPT-4V	**67.5**	**42.0**	**35.0**	41.0	**46.4**

Table 1: Accuracy of large multimodal models for single-concept abstract patterns in PUZZLEVQA.

tary material, and we plan to release the dataset publicly with a permissive license such as MIT license.

Each puzzle in PUZZLEVQA comprises of an image x_{image}, a natural language question $x_{question}$, an image caption that describes the image $x_{caption}$, an explanation that explains the pattern shown in the image $x_{explanation}$, a deduction statement that applies the pattern to the puzzle to derive the final answer $x_{deduction}$, a set of multiple-choice answers $x_{options}$, and the final answer x_{answer}. All of which are automatically generated during the puzzle creation process.

4 Experimental Setup

4.1 Inference Pipeline

To elicit reasoning steps from large multimodal models, we leverage zero-shot chain of thought (CoT) prompting (Kojima et al., 2022) with a prompt similar to "`Let's think step by step`". As the model may not always follow the same multiple-choice answer format, we also employ a model-based answer extraction stage. Detailed examples of the prompts can be found in the supplementary material. Please note that our main experimental setting used in Table 1 and 2 involves only the questions and images as multimodal inputs. On the other hand, we progressively provide additional ground-truth information such as image captions in Section 5.1 to diagnose the multimodal reasoning bottlenecks.

Chain of Thought Prompting. In the first prompting step, we construct a prompt $\hat{x}$ by modifying the question using a specific prompt template: "[I] Question: [X]. Options: [O]. Answer: [T]", where [I] is the input slot for x_{image}, [X] is the input slot for $x_{question}$, [O] is the input slot for $x_{options}$, and [T] is the input slot for the trigger sentence t_1. To elicit reasoning over the multimodal inputs, we use "`Let's describe the image first and think step by step`" as our trigger sentence t_1. This modified prompt $\hat{x}$ is then fed into a large multimodal model, and a greedy decoding strategy is utilized to generate the subsequent sentence y_1. If the letter answer can be extracted from y_1 with regular expressions, the prompting process terminates. However, if the letter answer cannot be extracted, we prompt the model itself to extract the answer.

Answer Extraction. In the second prompting stage, we use the generated sentence y_1 along with the modified prompt $\hat{x}$ to extract the final answer. We concatenate three elements to form "[$\hat{X}$] [Z] [A]" where [$\hat{X}$] is the input slot for $\hat{x}$, [Z] is the input slot for y_1, and [A] is the trigger sentence t_2 to extract the final answer. We defined t_2 as "`Therefore, among (A) (B) (C) (D), the answer is:`" or "`Therefore, among (A) (B) (C), the answer is:`", for puzzles with four and three multiple-choice questions respectively.

4.2 Models

To investigate the reasoning ability of large multimodal models, we select the best-performing open and closed-source models (Yue et al., 2023):

1. Qwen-VL-Chat (7B) (Bai et al., 2024) is an open-source large multimodal model designed to perceive and understand both texts and images. We use the version with open model weights and default chat template.

2. LLaVA-13B (Liu et al., 2023) is an large multimodal model which is based on the popular LLaMA (Touvron et al., 2023a) foundation language model. We use the model weights of the 1.5 version and default chat template.

3. Gemini Pro (Gemini Team, 2023) is a highly capable multimodal model released by Google, and we use their publicly available API to query the "gemini-pro-vision" version of the model.

4. Claude 3 Opus[3] is released by Anthropic and the most highest-performing multimodal model in their model family. We use their publicly available API to query the "claude-3-opus-20240229" version of the model.

5. GPT-4V (OpenAI, 2023) is released by OpenAI and widely regarded as the most capable

[3] `https://www.anthropic.com/news/claude-3-family`

	Numbers & Shapes	Colors & Numbers	Numbers & Size	Colors & Shapes	Size & Shapes	Colors & Size	Average
Random Baseline	25.0	25.0	25.0	25.0	33.3	25.0	26.4
Qwen-VL-Chat	29.5	26.5	26.5	26.5	30.0	31.0	28.3
LLaVA-13B	20.0	30.0	31.5	36.0	**39.0**	30.0	31.1
Gemini Pro	29.0	21.0	29.5	27.5	36.5	37.0	30.1
Claude 3 Opus	48.0	54.5	**34.0**	42.5	33.0	50.0	43.7
GPT-4V	**52.5**	**56.0**	31.0	**47.0**	31.5	**55.0**	**45.5**

Table 2: Accuracy of large multimodal models for dual-concept abstract patterns in PUZZLEVQA.

multimodal model based on existing benchmarks (Yue et al., 2023). We use their publicly available API to query the "gpt-4-vision-preview" version of the model.

5 Results

We report the main evaluation results on single-concept and dual-concept puzzles in Table 1 and Table 2 respectively. The evaluation results for single-concept puzzles, as shown in Table 1 reveal notable differences in performance among the open-source and closed-source models. GPT-4V stands out with the highest average score of 46.4, demonstrating superior abstract pattern reasoning on single-concept puzzles such as numbers, colors, and size. It particularly excels in the "Numbers" category with a score of 67.5, far surpassing other models, which may be due to its advantage in math reasoning tasks (Yang et al., 2023). Claude 3 Opus follows with an overall average of 39.4, showing its strength in the "Shapes" category with a top score of 44.5. The other models, including Gemini Pro and LLaVA-13B trail behind with averages of 34.5 and 27.5 respectively, performing similarly to the random baseline on several categories.

In the evaluation on dual-concept puzzles, as shown in Table 2, GPT-4V stands out again with the highest average score of 45.5. It performed particularly well in categories such as "Colors & Numbers" and "Colors & Size" with a score of 56.0 and 55.0 respectively. Claude 3 Opus closely follows with an average of 43.7, showing strong performance in "Numbers & Size" with the highest score of 34.0. Interestingly, LLaVA-13B, despite its lower overall average of 31.1, scores the highest in the "Size & Shapes" category at 39.0. Gemini Pro, on the other hand, has a more balanced performance across categories but with a slightly lower overall average of 30.1. Overall, we find that models perform similarly on average for single-concept and dual-concept patterns, which indicates that they also struggle with puzzles that require reasoning about multiple abstract concepts.

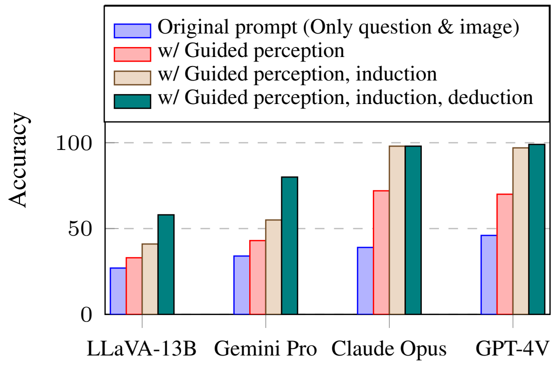

Figure 4: Analysis on multimodal reasoning bottlenecks. We progressively prompt models with ground-truth explanations for visual perception, inductive reasoning, and deductive reasoning.

5.1 Analysis of Multimodal Reasoning Bottlenecks

Given the lower performance of existing large multimodal models, this raises the natural question of why they are not able to reason well about abstract patterns. As shown in Figure 2, the stages of solving abstract puzzles can be generally decomposed into visual perception, inductive reasoning, and deductive reasoning. Hence, we analyze their reasoning bottlenecks in Figure 4 by progressively providing the ground truth explanation in their prompts. Note that we omit the final answer in the deductive reasoning explanation to avoid making the question trivial, and the detailed prompts can be found in the supplementary material. Overall, we observe that the models perform better when provided with ground truth explanations, which suggests that they are able to leverage the additional information.

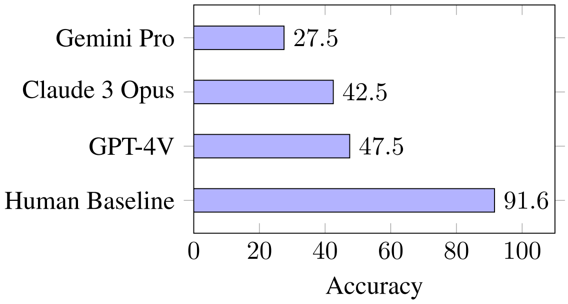

Figure 5: Comparison between average human performance and large multimodal models on a subset of PUZZLEVQA.

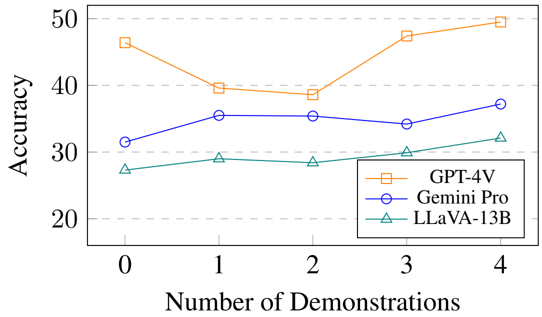

Figure 6: Analysis on the effect of few-shot demonstrations on model performance for single-concept puzzles.

Notably, GPT-4V and Claude 3 Opus is able to solve almost all cases when provided with both visual perception and inductive reasoning explanations. This suggests that the main bottlenecks for GPT-4V and Claude 3 Opus are visual perception and inductive reasoning to interpret the multimodal information and recognize the pattern from observations. However, this is not the case for LLaVA-13B and Gemini Pro, which demonstrate the largest improvement when guided by visual perception, inductive reasoning, and deductive reasoning together. This indicates that their main bottleneck is deductive reasoning to apply general principles of the pattern to solve specific cases. Note that these results are intended to serve as an **optimistic upper bound** of the model performance when provided with ground truth partial information, and may not indicate that the puzzles will become trivial.

5.2 Comparison to Human Performance

To further shed light on how the large multimodal models compare to the reasoning ability of humans, we conducted a human baseline study involving 23 university students[4]. Participants were allotted 30 minutes to solve 40 puzzle instances sampled from our 20 puzzle categories, yielding an average human baseline score of 91.6%, as shown in Figure 5. Note that the 20-24 age group of the participants correspond to the formal operational stage of cognitive development, as discussed in Section 2. In contrast, the highest-performing model, GPT-4V scored 47.5% on the same set of puzzle samples, highlighting the specific bottlenecks causing models to fall short of human cognition: primarily in visual perception and inductive reasoning, as discussed in Section 5.1.

[4]Note that the participants volunteered for the short study and we obtained prior permission from their instructor.

5.3 Effect of Few-Shot Demonstrations

While we focus on the zero-shot setting to investigate how multimodal models handle novel reasoning challenges, we also explore how models may use knowledge and strategies from other puzzles to solve a new, specific puzzle. This is akin to analogical reasoning, which involves using experience from similar scenarios to make inferences about a novel situation (Webb et al., 2022). Concretely, we run an analysis to study the effect of in-context learning (Brown et al., 2020) with few-shot demonstrations of other puzzle instances when the model is tasked to solve a specific puzzle. The demonstrations consist of interleaved instances of multimodal inputs of each puzzle image and question, as well as the ground truth reasoning explanations. To ensure that the demonstrations are sufficiently diverse, we randomly select puzzles of different categories from the given puzzle. As shown in Figure 6, we find a general trend of increasing performance with respect to the number of demonstrations. Although there are some cases of lower performance for GPT-4V, we see that models generally achieve their best performance with the most number of demonstrations. This suggests that the models are indeed capable of analogical reasoning, and in-context learning may be a promising direction to enhance the abstract reasoning abilities of multimodal models in the future.

In addition, while not the main focus of this work, there may be other methods of improving model performance, including model training or different prompting methods. Hence, we have also included preliminary studies on fine-tuning with LLaVA-13B and comparison between chain-of-thought and direct prompting in the supplementary material. To consider the effect of other factors, the supplementary material further includes an alternative setting with text-only models given the ground-truth visual perception caption, and an analysis on the effect of evaluation dataset size.

6 Related Work

The recent surge in multimodal pretraining and fine-tuning approaches (Liu et al., 2023; Bai et al., 2024) has led to the creation of various benchmarks. Benchmarks like VQA (Antol et al., 2015) and OK-VQA (Marino et al., 2019) aim at evaluating the basic perception and reasoning abilities of large multimodal models. Meanwhile, benchmarks like MMMU (Yue et al., 2023) and ScienceQA (Lu et al., 2022) offer an evaluation of LLMs' proficiency across multiple disciplines requiring domain-specific knowledge and multimodal understanding.

To investigate the fundamental challenges in multimodal perception and reasoning, we deliberately focus on the abstract domain, aiming to assess how models emulate cognitive abilities, particularly involving reasoning about abstract concepts and relationships. However, we note that existing benchmarks have limitations which make them less suitable for studying large multimodal models. The RAVEN dataset (Zhang et al., 2019) presents visual matrices with abstract patterns, challenging models to identify patterns and complete missing elements. However, we note that it has a specific spatial layout and can be solved exactly with search algorithms. Compared to CLEVR (Johnson et al., 2017) which offers synthetic visual scenarios and questions focusing on logic and commonsense, we focus on exploring how large multimodal models perceive and reason about multimodal patterns, which is more closely related to fundamental cognitive processes in humans (Mattson, 2014). While ConceptARC (Moskvichev et al., 2023) focuses on specific spatial concepts such as inside-outside and above-below, our dataset PUZZLEVQA studies how visual objects interact and relate based on broader abstract concepts such as colors, shapes, numbers, and size. Lastly, the MiniSCAN dataset (Lake et al., 2019) presents patterns that map special words to a sequence of color symbols, but is limited to color-based patterns.

What distinguishes our dataset, PUZZLEVQA, from the existing works is its systematic analysis of multimodal reasoning through abstract patterns, including perceptual, inductive, and deductive reasoning. Compared to the previous datasets, our multimodal patterns encompass broad and fundamental abstract concepts such as numbers, colors, shapes, and size. Notably, our dataset not only provides ground truth answers but also includes image captions and pattern explanations that enable more detailed and systematic diagnosis of the reasoning bottlenecks for large multimodal models.

7 Conclusion

In this work, we introduced the PUZZLEVQA dataset to investigate the reasoning challenges in large multimodal models. Our experiments demonstrated that, despite their sophistication, models such as GPT-4V exhibit substantial challenges when solving abstract pattern puzzles that require visual perception, inductive reasoning, and deductive reasoning, falling short of cognitive processes displayed by humans. Notably, our systematic analysis with ground truth explanations reveals that the main reasoning bottlenecks for GPT-4V are weaker visual perception and inductive reasoning capabilities. On the other hand, we found that other large multimodal models required more guidance with ground truth explanations, pointing to a broader range of reasoning challenges. Looking ahead, our work points to exciting avenues for advancing the reasoning abilities of large multimodal models. Future research should focus on enhancing models' understanding of multimodal information and refining their abstract reasoning faculties, in order to further enhance their general capabilities.

Acknowledgement

We are grateful to the Microsoft Foundational Model Research grant for their generous support. We are also grateful to AI Singapore Governance grant ID: AISG3-GV-2023-010, and AcRF MoE Tier-2 grant (Project no. T2MOE2008, and Grantor reference no. MOE-T2EP20220-0017) titled: "CSK NLP: Leveraging Commonsense Knowledge for NLP", for the support. This work is also supported by the Microsoft Research Accelerate Foundation Models Academic Research program. Chia Yew Ken would like to thank Tan Hui Min Grace as a source of inspiration for the fun and unique puzzles, without which this work would not have started.

Limitations

In this work, we mainly focus on the zero-shot setting to investigate how large multimodal models face reasoning challenges in novel situations. However, previous works have shown that prompting with demonstrations (Brown et al., 2020) may improve the models ability to adapt to new tasks. Hence, we also include experiments in the few-shot setting in Section 5.3, which showed inconsistent benefits, and we aim to explore this area in the future works.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond.

Tom Brown, Benjamin Mann, Nick Ryder, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, John A. Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuan-Fang Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *ArXiv*, abs/2303.12712.

Susan Carey. 2000. The origin of concepts. *Journal of Cognition and Development*, 1:37 – 41.

Raymond Bernard Cattell. 1963. Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, 54:1–22.

Charles Cole. 1996. Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought. *Journal of the Association for Information Science and Technology*, 47:403–404.

Bosheng Ding, Chengwei Qin, Linlin Liu, Lidong Bing, Shafiq R. Joty, and Boyang Albert Li. 2022. Is gpt-3 a good data annotator? In *Annual Meeting of the Association for Computational Linguistics*.

Google Gemini Team. 2023. Gemini: A family of highly capable multimodal models.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Brenden M. Lake, Tal Linzen, and Marco Baroni. 2019. Human few-shot learning of compositional instructions. In *Annual Meeting of the Cognitive Science Society*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mark P. Mattson. 2014. Superior pattern processing is the essence of the evolved human brain. *Frontiers in Neuroscience*, 8.

Arseny Moskvichev, Victor Vikram Odouard, and Melanie Mitchell. 2023. The conceptarc benchmark: Evaluating understanding and generalization in the arc domain. *ArXiv*, abs/2305.07141.

OpenAI. 2023. Gpt-4v(ision) system card.

Jean Piaget. 1976. Piaget's theory.

Linlu Qiu, Liwei Jiang, Ximing Lu, Melanie Sclar, Valentina Pyatkin, Chandra Bhagavatula, Bailin Wang, Yoon Kim, Yejin Choi, Nouha Dziri, and Xiang Ren. 2024. Phenomenal yet puzzling: Testing inductive reasoning capabilities of language models with hypothesis refinement. In *The Twelfth International Conference on Learning Representations*.

Pratyusha Sharma, Tamar Rott Shaham, Manel Baradad, Stephanie Fu, Adrian Rodriguez-Munoz, Shivam Duggal, Phillip Isola, and Antonio Torralba. 2024. A vision check-up for language models. *ArXiv*, abs/2401.01862.

Theodore R. Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L. Griffiths. 2023. Cognitive architectures for language agents. *ArXiv*, abs/2309.02427.

Joshua B. Tenenbaum. 2018. Building machines that learn and think like people. In *Adaptive Agents and Multi-Agent Systems*.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *ArXiv*, abs/2401.06209.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin R. Stone, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Taylor W. Webb, Keith J. Holyoak, and Hongjing Lu. 2022. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7:1526 – 1541.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v(ision). *ArXiv*, abs/2309.17421.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*.

Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. 2019. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Puzzle Category	Multimodal Templates	Test Instances
Numbers	2	200
Colors	2	200
Shapes	2	200
Size	2	200
Numbers & Shapes	2	200
Numbers & Colors	2	200
Numbers & Size	2	200
Shapes & Colors	2	200
Shapes & Size	2	200
Colors & Size	2	200
Total	20	2000

Table 3: Dataset statistics of PUZZLEVQA.

A Appendix

A.1 Multiple-Choice Format Details

To generate multiple choice options for numeric puzzles, we use heuristics based on the range of the number. For example, if the number is less than 10, then we sample from the range 1 to 9. If the number if less than 100, we sample from the range 1 to 99, and so on. For discrete option choices, we sample from the possible objects in the image, such as the list of predefined colors or sizes or shapes.

A.2 Dataset Details

We report the dataset statistics of PUZZLEVQA in Table 3.

A.3 Prompt Examples

We show examples of the textual prompts in Figure 7. Note that the prompt examples correspond to the image and puzzle in Figure 1. We use a consistent prompt format across all abstract puzzles.

A.4 Code Implementation Example

```
import math
import random
from typing import List, Tuple, Dict

from PIL import Image, ImageDraw, ImageFont
from pydantic import BaseModel

class ColorHexagonPattern(BaseModel):
    colors: Dict[str, str] = dict(
        blue="#6fa8dc",
        green="#93c47d",
        yellow="#ffd966",
        red="#e06666",
        purple="#8e7cc3",
        orange="#f6b26b",
    )
    image_size: int = 512
    scale_factor: int = 4
    path_font: str = "fonts/OpenSans-Medium.ttf"

    @staticmethod
    def get_centroid(points: List[Tuple[float, float]]) ->
        ↪ Tuple[float, float]:
        x = sum(p[0] for p in points) / len(points)
        y = sum(p[1] for p in points) / len(points)
        return x, y

    def sample_colors(self) -> Tuple[List[str], List[str]]:
        while True:
            names = random.sample(list(self.colors), k=3)
            if "orange" in names and "yellow" in names:
                continue  # Hard to distinguish
            names = names + names
            colors = [self.colors[n] for n in names]
            return names, colors

    def make_sample(self):
        # Set the size of the image
        size = self.image_size * self.scale_factor
        image = Image.new("RGB", size=(size, size), color="
            ↪ white")
        draw = ImageDraw.Draw(image)
        center = size // 2

        # Hexagon properties
        length = size // 3  # Length of a side of the
            ↪ hexagon and triangles
        triangle_height = math.sqrt(3) / 2 * length

        # The vertices of the hexagon
        hexagon = [
            (center + length / 2, center - triangle_height),
            (center - length / 2, center - triangle_height),
            (center - length, center),
            (center - length / 2, center + triangle_height),
            (center + length / 2, center + triangle_height),
            (center + length, center),
        ]

        # Colors for the triangles
        names, colors = self.sample_colors()
        i_answer = random.randint(0, len(colors) - 1)
        answer = names[i_answer]
        colors[i_answer] = "#eeeeee"  # Grey

        # Draw the hexagon made of six triangles
        for i in range(6):
            # Coordinates of the triangle vertices
            triangle = [hexagon[i], hexagon[(i + 1) % 6], (
                ↪ center, center)]
            # Draw the triangle
            draw.polygon(triangle, fill=colors[i])
            # Draw the outline with custom width
            points = [hexagon[i], hexagon[(i + 1) % 6], (
                ↪ center, center), hexagon[i]]
            draw.line(points, fill="black", width=self.
                ↪ scale_factor * 4)
            # Draw "?" on the missing answer part
            if i == i_answer:
                draw.text(self.get_centroid(triangle),
                    text="?",
                    font=ImageFont.truetype(self.path_font,
                        ↪ size=size // 10),
                    anchor="mm",
                    fill="black",
                )

        names[i_answer] = "?"
        instances = sorted(set(n for n in names if n not in
            ↪ [answer, "?"]))
        image = image.resize((self.image_size, self.
            ↪ image_size), Image.LANCZOS)
        return (
            dict(
                question="What is the missing color of the
                    ↪ part denoted with a question mark?"
                    ↪ ,
                answer=answer,
                options=sample_options(answer, options=list(
                    ↪ self.colors), k=4),
                caption=f"There is a hexagon split into six
                    ↪ parts with the colors {names} in an
                    ↪  anti-clockwise order.",
                explanation=f"We observe that a {instances
                    ↪ [0]} part is opposite another {
                    ↪ instances[0]} part, and a {
                    ↪ instances[1]} part is opposite
                    ↪ another {instances[1]} part. Thus,
                    ↪ the pattern is that the colors in
                    ↪ opposite parts are the same.",
                deduction=f"Based on the pattern that
                    ↪ spatially opposite parts have the
                    ↪ same color, the missing color of
                    ↪ the part which is opposite a {
                    ↪ answer} part should be {answer}.",
            ),
            image,
        )
```

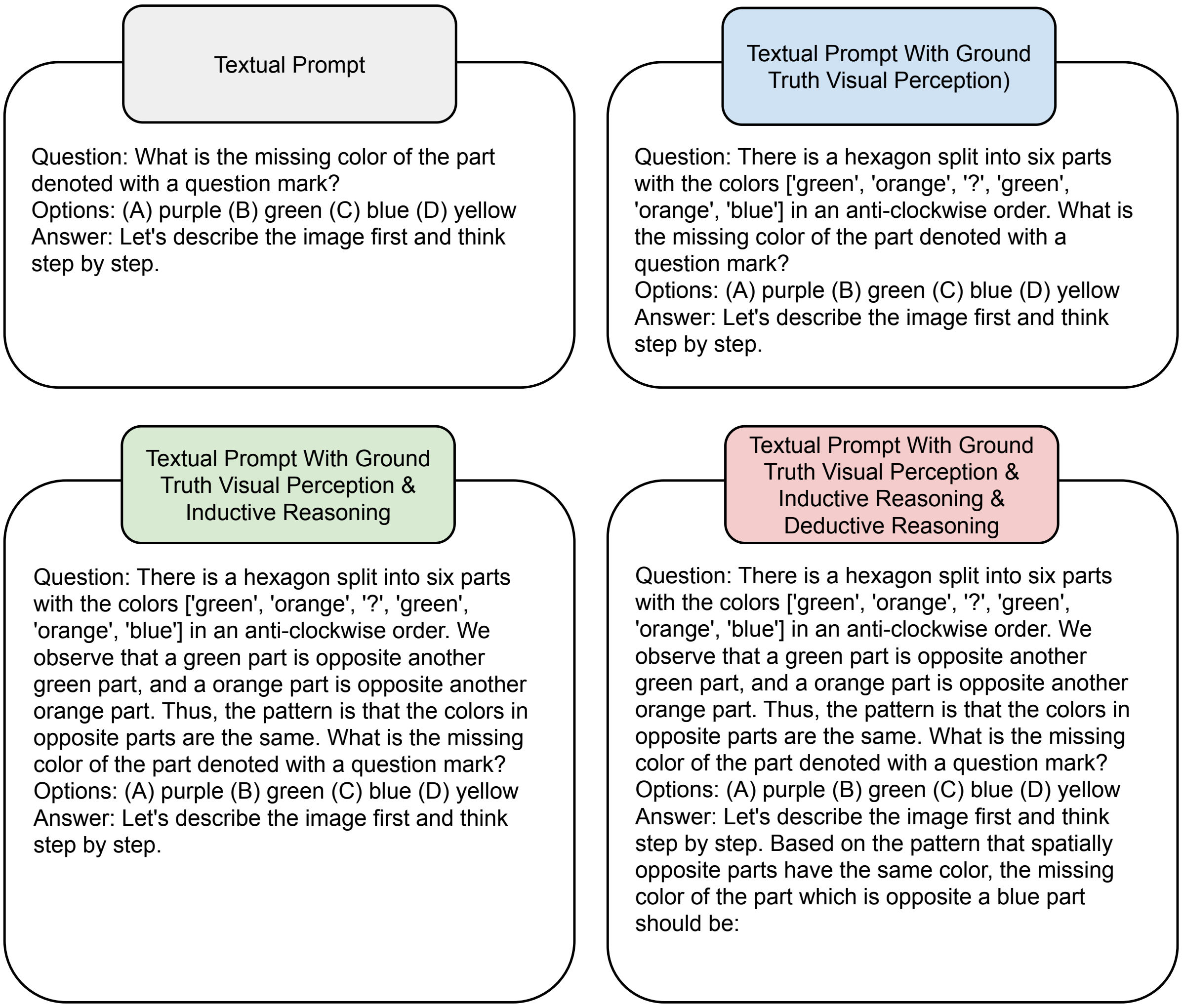

Figure 7: Textual prompt examples for eliciting reasoning steps from large multimodal models.

A.5 Dataset Size Analysis

Regarding the dataset size and diversity, we set the number of generated instances for each puzzle to 100, to reduce experimental variance and maintain a reasonable evaluation cost. Hence, the current dataset size is 2000 samples (20 templates with 100 instances each). As there are two templates per puzzle category, this means there are 200 test samples for each puzzle category. To observe the impact of the number of test samples, we evaluate the models on three different data settings as shown in Table 4: 50, 100, and 200 test samples in per puzzle, which correspond to 1000, 2000, and 4000 total samples respectively. In general, we observe some variations in the average score for single-concept puzzles, but it does not significantly affect the comparison of performance between different models. Hence, we believe that the chosen dataset size is large enough to be relatively robust to experimental variance. To investigate the multimodal reasoning capabilities across diverse abstract scenarios, we construct the puzzles based on four fundamental concepts: numbers, colors, shapes, and size. Furthermore, to evaluate how well the models can relate to multiple concepts, we design both single-concept and dual-concept puzzles, and the taxonomy of diverse puzzles is shown in Figure 3.

	1000	2000	4000
Gemini Pro Avg. Score	29.2	34.5	32.2
GPT-4V Avg. Score	48.8	46.4	48.4

Table 4: Analysis of model performance with respect to number of testing data samples.

	Numbers	Colors	Size	Shapes	Average
LLaVA-13B	25.5	29.0	37.0	39.0	32.6 (+5.1)
Gemini Pro	16.0	28.5	35.0	49.0	32.1 (-2.4)
GPT-4V	72.5	26.5	30.0	45.5	43.6 (-2.8)

Table 5: Results of direct prompting and change in average performance compared to CoT prompting.

A.6 Comparison of CoT and Direct Prompting

To investigate the effect of direct prompting without chain of thought, we evaluated the models on our main setting as shown in Table 1 and 2. The results are shown in Table 5 for single-concept puzzles and indicate that direct prompting is less effective for Gemini Pro and GPT-4V models, compared to CoT prompting in Table 1. This may be due to differences in training data and alignment methods between the models.

B Qualitative Analysis

To illustrate the reasoning bottlenecks of GPT-4V, we include two case study samples in Figure 8. For instance, the sample on the left is from the size & shapes category of puzzles, for which the model under-performed the random baseline. For visual perception, we observe that the model presents severe limitations, as it is unable to recognize simple polygon shapes and hallucinated additional shapes which are not in the image. Regarding inductive reasoning, we observe that the model was able to recognize the sizes of the different objects, but did not recognize the correct pattern that the circles directly adjacent to the center should be small in size. Hence, we believe that there is ample area for improvement for abstract reasoning ability in large multimodal models.

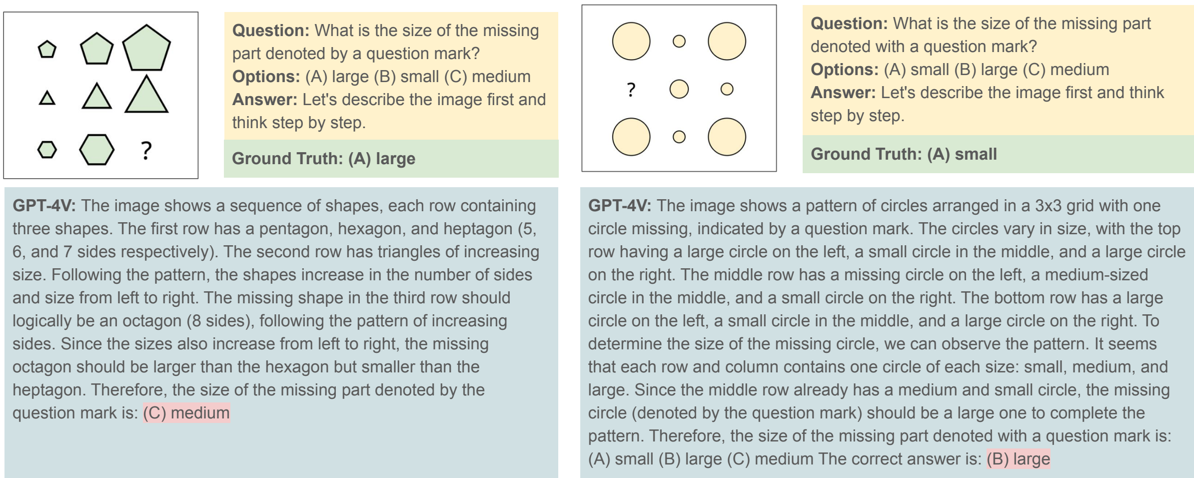

Figure 8: Case study on two sample predictions from GPT-4V. The example on the left shows visual perception failures and the example on the right shows the faulty inductive reasoning of the model which proposed a spurious pattern in the image.