# BONGARD IN WONDERLAND: VISUAL PUZZLES THAT STILL MAKE AI GO MAD?

**Antonia Wüst**[1] **Tim Tobiasch**[1] **Lukas Helff**[1,2] **Devendra S. Dhami**[3]

**Constantin A. Rothkopf**[2,4,5] **Kristian Kersting**[1,2,4,6]

[1]AI and ML Group, TU Darmstadt; [2]Hessian Center for AI (hessian.AI)
[3]Uncertainty in AI Group, TU Eindhoven; [4]Centre for Cognitive Science, TU Darmstadt
[5]Institute of Psychology, TU Darmstadt; [6]German Center for AI (DFKI)

## ABSTRACT

Recently, newly developed Vision-Language Models (VLMs), such as OpenAI's GPT-4o, have emerged, seemingly demonstrating advanced reasoning capabilities across text and image modalities. Yet, the depth of these advances in language-guided perception and abstract reasoning remains underexplored, and it is unclear whether these models can truly live up to their ambitious promises. To assess the progress and identify shortcomings, we enter the wonderland of Bongard problems, a set of classical visual reasoning puzzles that require human-like abilities of pattern recognition and abstract reasoning. While VLMs occasionally succeed in identifying discriminative concepts and solving some of the problems, they frequently falter, failing to understand and reason about visual concepts. Surprisingly, even elementary concepts that may seem trivial to humans, such as simple spirals, pose significant challenges. Moreover, even when asked to explicitly focus on and analyze these concepts, they continue to falter, suggesting not only a lack of understanding of these elementary visual concepts but also an inability to generalize to unseen concepts. These observations underscore the current limitations of VLMs, emphasize that a significant gap remains between human-like visual reasoning and machine cognition, and highlight the ongoing need for innovation in this area.[1]

## 1 Introduction

Visual reasoning, the ability to understand, interpret, and reason about the visual world, is a fundamental aspect of human intelligence [27]. It allows us to navigate our environment, interact with objects, and make sense of complex visual scenes. In recent years, the field of artificial intelligence (AI) has advanced rapidly toward replicating aspects of this visual reasoning, with significant focus placed on Vision-Language Models (VLMs) [5, 24, 25]. These models integrate visual and textual information to generate descriptive content, aiming to mimic how humans comprehend and reason about the world. Because of their human-like responses, VLMs often create the illusion of possessing human-like perception and intelligence. However, as recent work shows, VLMs and the Large Language Models (LLM) on which they are based have dramatic shortcomings in the case of reasoning [30] and visual perception [12, 13, 19, 34] or their combination [39, 47, 48].

Bongard problems (BPs), a class of visual puzzles that require the identification of underlying rules based on a limited set of images, provide a unique and challenging benchmark for assessing visual reasoning abilities in AI systems [4]. Conceived by Russian scientist Mikhail Bongard in 1967, these visual puzzles test cognitive abilities in pattern recognition and abstract reasoning, posing a formidable challenge even to advanced AI systems [15].

[1]Code available at `https://github.com/ml-research/bongard-in-wonderland`.

Figure 1: **The wonderland of Bongard problems.** The challenging puzzles of Bongard circle around simple black-and-white diagrams that picture geometric shapes, patterns, or lines arranged in various configurations. While the visualization is rather simple, the underlying concepts can be abstract and complex.

A BP consists of twelve diagrams, divided into two sides. For each side, a distinct and specific conceptual theme must be identified, which clearly differentiates it from the other side. Although the diagrams themselves are visually simple (see Figure 1), the underlying concepts that connect the images within each group can be abstract, such as *more filled objects than outlined objects* or *turning direction of spiral shape*. Thus, unlike pattern recognition in classification tasks, BPs are not about finding visual patterns in a single diagram that match a certain concept but about finding concepts that allow for the description of a set of diagrams.

While traditional machine learning approaches have made some early progress on BPs [10, 33], the potential of VLMs remains largely unexplored. Since VLMs already struggle with recognizing rather simple visual patterns [34, 47], it is expected that BPs are still a particularly hard challenge for VLMs and provide a valuable basis for exploring in more detail which patterns are more or less difficult to identify by state-of-the-art models.

In this work, we investigate the performance of VLMs in the domain of BPs. We examine how well different VLMs can discover the underlying rules in these puzzles, and identify strengths and limitations in their reasoning capabilities. For this, we consider a setting where an open-ended solution for the BPs needs to be discovered and a second multiple-choice setting in which the correct rule-pair needs to be selected from a set of possible solutions. The results from a selected set of BPs are then compared to human performance, providing insight into how well VLMs measure up to human reasoning in this domain. Further, we investigate the pattern recognition abilities of the models on four problems in more detail. Our results provide insights into the perceptual madness of VLMs and suggest opportunities for improvement.

## 2 Related Work

**Bongard and ML.** Depeweg et al. [10] define a formal language to represent compositional visual concepts. Using this language and Bayesian inference, concepts can be induced from the examples provided in each problem. For a subset of 35 problems, there is reasonable agreement between the concepts with high posterior probability and the solutions formulated by Bongard himself [10]. Raghuraman et al. [33] explore the principles of Bongard problems on the classical and real-world image versions of them. However, they change the problem setting from an open-ended task, where a rule has to be formulated, to a setting where a subset of the puzzle's images needs to be classified correctly. Youssef et al. [41] approach Bongard problems with a reinforcement learning setting for extracting meaningful representations and counterfactual explanations.

**Benchmarks for VLMs.** Traditional visual machine learning benchmarks largely focus on straightforward machine perception tasks [9, 20, 22, 35]. In contrast, benchmarks specifically designed for VLMs often go one step further and involve more complex tasks such as image captioning, scene or diagram understanding, visual question answering (VQA), or visual-commonsense reasoning [2, 16, 17, 18, 19, 28, 34, 42, 43, 44, 45]. Yet, most of these only require simple reasoning abilities. More recent benchmarks have been introduced to probe advanced reasoning skills, e.g., logical learning [14, 38], mathematical reasoning [26] or analogy-based visual reasoning [8, 29, 46]. Although this shift towards more cognitively demanding tasks is promising, comprehensive diagnostic evaluations of VLMs' reasoning

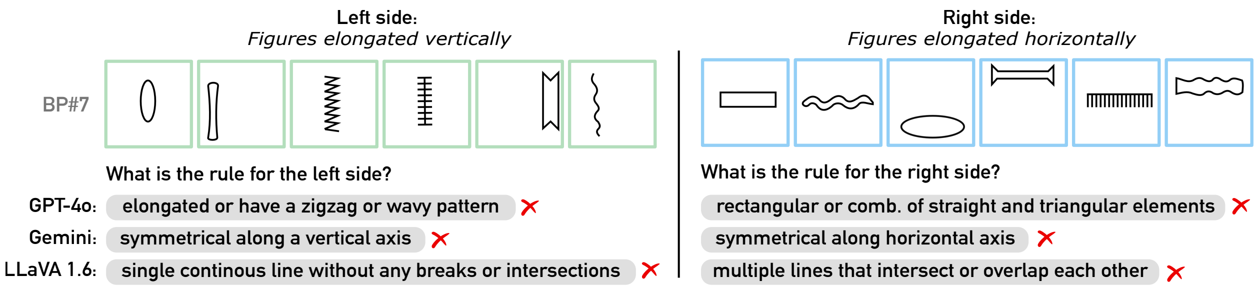


Figure 2: **VLMs struggle to solve BPs out of the box**. The images of BP#7 are depicted, where the left side has images with vertically elongated shapes and the right side has images of horizontally elongated shapes. Although the concepts *vertical* and *horizontal* may seem trivial to a human, the VLMs struggle to generate discriminative rules.

capabilities that pinpoint sources of error and model limitations remain scarce. Furthermore, the degree to which these models genuinely comprehend complex, abstract visual concepts is yet to be fully investigated.

## 3 Bongard Problems and VLMs

**Bongard problems** (BP), introduced 1970 by Mikhail Bongard [4], are classical visual puzzles that test for capabilities of pattern recognition, concept formation, and abstraction. Each BP consists of twelve simple black-and-white diagrams divided into a left and a right group. Usually, all images share some similarity, but for both sides, there is an opposing property or rule, respectively, that its six images have in common (and which is shared by no image of the other side). An example BP is shown in Figure 2 where the properties are *vertical* and *horizontal* orientation.

The task is the linguistic expression of the underlying rule that distinguishes the two groups. These rules vary in complexity, ranging from simple geometric properties like the presence of a circle to more abstract or relational concepts like symmetry or the presence of a right angle. In some cases, the rule of the right side is just the negative of the left rule, like BP#24 (*a circle present* vs. *no circle present*). Still, for the majority of BPs the second rule is a more specific opposite of the first e.g., BP#6 (*triangles* vs. *quadrangles*).

In contrast to mainstream classification tasks, BPs differ from mainstream classification tasks due to their complexity and reliance on abstract reasoning rather than direct pattern recognition. Specifically, BPs test the ability to express distinctive and common features of images, including the pattern recognition necessary to correctly associate the features with images, as well as the ability to come up with textual rules that can characterize the meta-pattern (not within each but) across all twelve diagrams that constitute a BP.

This multi-modality of BPs makes them an interesting challenge for VLMs or multimodal AI in general. In this work, we use different strategies to prompt VLMs to solve BPs which will be introduced in the following.

**Open-Ended Solving of Bongard Problems.** For each BP, the model is prompted individually by providing a text prompt together with an image showing all twelve diagrams. The prompt follows the following setup:

1. The structure of the image is described first. This includes the number of diagrams, their arrangement (left and right sides), and the fact that the diagrams are black and white with specific shapes and features.
2. The task for the model is defined by explaining that there are two rules, one for the left side and one for the right side and that the rules should not apply to any of the diagrams on the other side.
3. The model is instructed to evaluate step-by-step, beginning with a detailed analysis of the diagrams, followed by inferring the underlying rules.
4. The required response format is specified as a dictionary with two entries, one for each rule.

The complete prompt can be found in Listing 1. The answers to the reasoning task are then compared to the ground truth[2] by an LLM-Judge, as the answer setting is open-ended (cf. Listing 3 for prompt).

**Multiple Choice Setting.** In this setting, rather than having the model generate the rules itself, we provide it with a set of predefined rule pairs from the BP domain. The model is then tasked with selecting the correct rule pair from these options. The prompt follows the same structure as in the previous task but includes an additional list of available rules for the model to choose from. The expected answer is the ID corresponding to the correct rule-pair (cf. Listing 1).

[2] `https://www.foundalis.com/res/bps/bongard_problems_solutions.htm`

Table 1: **Performance of VLMs on 100 BPs (top) as well as multiple-choice BPs (bottom)**. Results depict the rounded average of solved BPs over 3 runs. All models struggled with the classical BP setup, with GPT-4o achieving the highest score, solving only 21 out of 100 BPs. Even on the multiple-choice BPs, difficulties persist. Only when the number of choices is considerably limited does the performance increase. *Context size of LLaVA 1.6 not sufficient.

| | GPT-4o | Claude | Gemini | LLaVA 1.6 | LLaVA 1.5 |
|---|---|---|---|---|---|
| **Solved BPs** (of 100) | **21** | 14 | 5 | 2 | 1 |
| Multiple Choice (100) | 23 | **28** | 16 | * | 2 |
| Multiple Choice (10) | 68 | **69** | 59 | 16 | 24 |

**Detect Specific Concepts.** To investigate the limitations of visual descriptions we create a *perception* task. Here, the relevant concepts for the BP are provided as context (e.g., *horizontal* and *vertical* orientation). Based on this, the task is to predict for every single image of the BP whether the left side or the right side concept is true. For this four specific prompts were implemented for BPs #16, #29, #36, and #55 (Listings 4, 5, 6, 7).

# 4 Experimental Evaluation

Our intention here is to investigate to what extent state-of-the-art VLMs can solve Bongard problems. At first, we evaluated the models quantitatively on all 100 puzzles. We then compared their performance against humans and investigated them qualitatively in more detail on the base of four selected BPs. Specifically, we address the following research questions:

**(Q1)** How well can state-of-the-art VLMs solve Bongard problems?

**(Q2)** How do VLMs perform in comparison to humans?

**(Q3)** How accurately can VLMs detect concepts required to solve BPs?

**Data.** For our evaluations, we considered the 100 original Bongard problems of [4]. We used the dataset variation of [10], which contains high-resolution images of the original diagrams.

**Models.** We evaluated the models GPT-4o [1], Claude 3.5 Sonnet [3], Gemini 1.5 Pro 37, and LLaVA with versions v1.6-34b [25] and v1.5-13b [24]. For simplicity, the models are referred to as GPT-4o, Claude, Gemini, LLaVA 1.6, and LLaVA 1.5 in the following. For the LLM-judge, we use GPT-4o.

**Can VLMs solve Bongard problems? (Q1)** As a first step, we aimed to assess how well current state-of-the-art VLMs can solve BPs. To do this, we tasked our selection of VLMs with solving each BP three times. The answers were evaluated by the LLM-judge, which determined whether each response correctly solved the BP. The evaluation results are shown in the top row of Table 1. Notably, GPT-4o emerged as the best performing model, with an average of 21 solved BPs. However, this performance is still surprisingly low, especially when compared to human abilities (*cf.* (Q2) and [23, 31]). A more detailed breakdown of which BPs were solved is provided in Table 3. It shows that even rather simple BPs with concepts like *small vs. large shapes* (BP#2) and *vertical vs. horizontal elongated figures* (BP#7) cannot be solved in most attempts. Some example responses from the VLMs for BP#7 are illustrated in Figure 2, highlighting that the models frequently fail to identify the correct rule and are often distracted by irrelevant features like *wavy patterns* or *rectangular shapes*.

In a further setting, we analyzed how the results change when providing the models with all existing rule pairs of the BPs and asking them to select the correct one (*cf.* Table 1 *Multiple Choice (100)*). Interestingly, this does not change the performance for GPT-4 and LLaVA-v1.5 significantly. However, Gemini and Claude's performance is better in this setting; Claude can even solve 28 BPs on average.

To further simplify the task, we reduced the number of options to 10 possible rule pairs and repeated the procedure. Now, the models perform better, reaching up to 69 solved BPs (*cf.* Table 1 *Multiple Choice (10)*). This is interesting since before the correct solution was present as well, but the models could not select it correctly. Therefore, it is unclear whether the models actually caught the concepts or if merely the exclusion procedure was easier. The question remains whether, with specialized context, it is possible to solve a BP if it has not been solved before. We investigated this in more detail in the context of (Q3). First, let us compare the results of the VLMs to human reasoning abilities.

**How do VLMs perform in comparison to humans? (Q2)** To compare the performance of the VLMs to human performance we considered the results from the user study of [10]. Here, 13 participants were asked to solve the first 50 BPs within 45 minutes. We aggregated the mean of the number of solved BPs for the humans and the VLMs on the

Table 2: **Humans, unlike VLMs, are capable of solving the majority of BPs.** Results of each VLM on the individual Bongard Problems compared to the results of 13 humans on 31 BPs from [10]. Each model was prompted 13 times and the mean of the correct responses per BP is reported. *Cf.* Table 6 for performance on single BPs.

| BP Type | Human | GPT-4o | Claude | Gemini | LLaVA 1.6 | LLaVA 1.5 |
|---|---|---|---|---|---|---|
| existence | **93.64** | 33.33 | 38.46 | 30.77 | 0.00 | 0.00 |
| size | **86.92** | 1.54 | 21.54 | 0.00 | 0.00 | 0.00 |
| concept | **72.13** | 39.56 | 25.27 | 6.59 | 3.30 | 4.40 |
| number | **72.50** | 7.69 | 25.00 | 5.77 | 3.85 | 0.00 |
| spatial | **93.75** | 10.00 | 8.46 | 7.69 | 0.00 | 0.00 |
| All | **84.41** | 17.77 | 20.16 | 8.22 | 1.33 | 1.06 |

specific subset of the study. The results are presented in Table 2 (All). Human performance far surpasses that of the VLMs, with humans achieving an average accuracy of 84.41% across the BPs, while the best-performing VLM, Claude, only reaches 20.16%.

For a more detailed comparison, we group the BPs into five categories, based on the nature of their ground truth rules, the categorization can be found in Table 6 and the results for the respective groups are reported in Table 2. Humans perform the best for *existence* and *spatial* BPs with over 93% accuracy. Overall, *existence* is also the best category for VLMs, however *spatial* is one of the worst with all models only reaching 10% or less. This is in line with recent works suggesting that VLMs struggle with spatial relations [19, 34, 47].

In the *concept* category, GPT-4o scored the highest accuracy with almost 40% solved BPs. This is a category where humans struggled a bit more, i.e., solved around 72% of the BPs. A reason for this could be that VLMs benefit here from their extensive world knowledge while not all humans might be familiar with geometrical visual concepts, such as *convex* vs. *concave* (BP#4) and *thin elongated convex hull* vs. *compact convex hull* (BP#12).

Overall, there was only one BP, BP#6, where a VLM outperformed the humans. GPT-4o solved BP#6 correctly 13 times, while one of the thirteen human participants made an error, likely due to a careless mistake, highlighting the inherent fallibility of humans. Despite this, in response to (Q2), we must conclude that VLMs still fall significantly short of matching the visual pattern recognition capabilities of humans. To gain deeper insight into the shortcomings of the VLMs, we will now examine in detail whether they can recognize the core concepts of four example BPs when explicitly prompted for them in the course of Q3.

**How accurately can VLMs detect concepts required to solve BPs? (Q3)** We saw that the VLMs show poor performance on the BP dataset. This could stem from difficulties in perceiving the diagrams properly but also from reasoning failures, such as incorrectly formulating rules that apply distinctly to each side. None of the models was able to solve BP#16, BP#29, BP#36 and BP#55 correctly even though the conceptual complexity of the rules is rather small. This raises the question on whether the perception of the models is flawed. We investigated this in more detail by providing the individual diagrams of the BPs to the models and asking them directly for the relevant concepts. An excerpt of the responses is displayed in Figure 3 (*cf.* Table 7, 8, 9, 10 for all responses). When all images from the BP are correctly categorized, we take it as an indication that the VLM is able to capture the concept in principle.

Surprisingly, we find that for BP#16 (*cf.* Figure 3 top left), even though some images are classified correctly, none of the models can classify all images correctly. Instead, we can see a tendency to classify one of the directions rather than the other, e.g., GPT-4o and Claude almost always predict *counter-clockwise* as the turning direction for all of the diagrams.

For BP#55 (*cf.* Figure 3 top right) we also see that none of the models is able to classify the concepts of all images correctly. Here we can see very similar behaviors across the different models, e.g., all models categorize the first diagram of the BP falsely, but in most of the attempts, they are able to categorize the second diagram correctly. Interestingly, we found a strong correlation between the absolute position of the circle in the diagram (*left* vs. *right*) and the models responses. We take this as an indicator, that the models did not manage to reason spatially about the position of the circle in relation to the cavity.

For BP#29, the models were asked to determine whether there are more shapes inside or outside the bigger shape (*cf.* Figure 3 bottom left). Even though the final decisions of the models were primarily correct, we saw in the answers that except for Claude, none of the models was able to count the shapes correctly. GPT-4o, for example, was convinced that image 7 has as many shapes on the outside as on the inside.

For the last BP, BP#36 (*cf.* Figure 3 bottom right), the models can identify the concept better, with some even classifying all 12 images correctly (cf. Table 9). The ground truth concept is *triangle vs. circle on top* and for this the perception

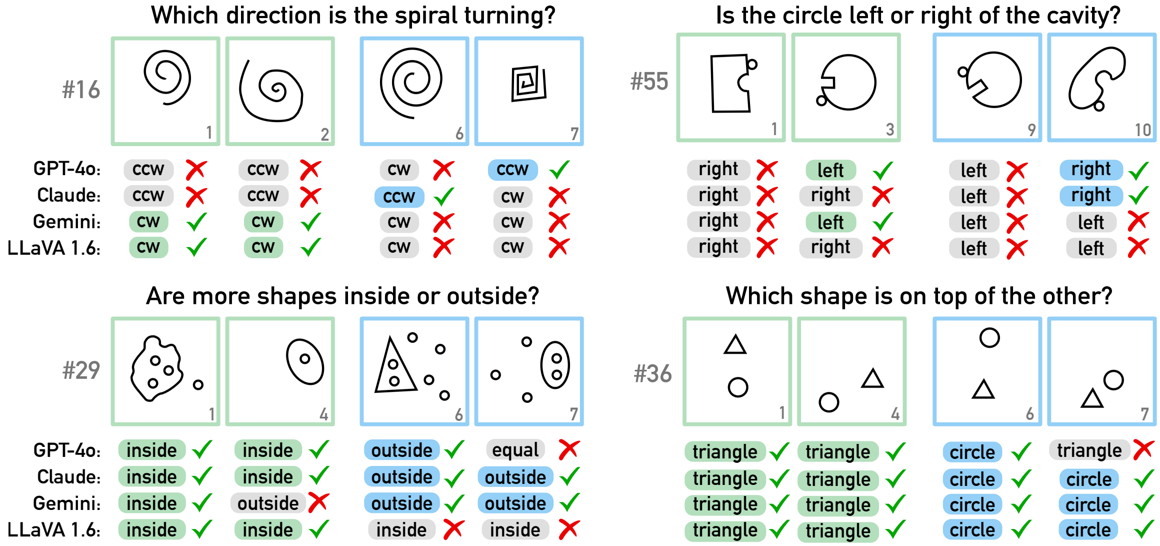


Figure 3: **VLMs fail to identify simple visual concepts.** VLMs challenged with identifying visual concepts in BPs. Although the VLM is able to recognize some of the concepts when specifically asked for (bottom), on the others, it continues to falter (top). Abbreviations used for *clockwise (cw)* and *counter-clockwise (ccw)*.

seems to work more reliably. This could be because the concept of typical shapes positioned above or below one another is generally more familiar compared to other BPs. In this case, the challenge of solving the BP from scratch likely lies more in pattern discovery or reasoning rather than perception.

Overall, the observed behavior is remarkable and seems to indicate that perception is the key issue for not identifying the correct rules of BP#16, BP#29, and BP#55. However, the results of BP#36 open the possibility of improving the performance of the VLMs through more sophisticated approaches, e.g., a multi-stage approach where first possible rules are discovered and then evaluated individually. Nonetheless, with regard to Q3 we have to conclude that the current perception capabilities of the evaluated VLMs are still insufficient to visually capture the concepts required for solving BPs.

**Limitations.** While BPs are valuable for assessing abstract reasoning, they also represent a narrow and highly specialized set of challenges that may not comprehensively reflect the broad range of problems VLMs encounter in real-world applications. Additionally, the reliance on our LLM-Judge introduces some uncertainty in the evaluation process. Future work should expand these evaluations to more diverse tasks and evaluate the judge's performance and additional model architectures to address these limitations.

## 5 Conclusion and Future Work

We presented a diagnostic evaluation of VLMs using the classical Bongard problems (BPs), providing valuable insights into their current capabilities of pattern recognition and abstract reasoning. Our experimental results highlight a significant gap between human-like visual reasoning and machine cognition. Specifically, we found that VLMs are still largely unable to solve the majority of BPs, with the best-performing model, GPT-4o, solving only 21 out of the 100 BPs. Moreover, our analysis suggests that the limitations of current VLMs extend beyond just visual reasoning; they also struggle to perceive and comprehend elementary visual concepts, such as simple spirals. A model that cannot recognize the direction in which a spiral is rotating cannot reason about whether multiple spirals are rotating in the same direction. Our findings raise several critical questions: Why do VLMs encounter difficulties with seemingly simple Bongard Problems, despite performing impressively across various established VLM benchmarks [11, 21]? How meaningful are these benchmarks in assessing true reasoning capabilities?

An intriguing direction for future research would be analyzing the visual and textual latent spaces of the VLMs for BPs [6]. Additionally, more recent approaches for image encoding in VLMs such as CLOC [7] could be investigated or fine-tuning of CLIP [32] could be considered. Further, it could be interesting to translate the visual concepts of BPs to real-world contexts, to evaluate to what extend the visual appearance or style of the images influences the perception. For higher transparency of the model's reasoning when solving BPs, concept-based approaches could be considered, such as first discovering relevant concepts of the BPs, e.g., via VLMs or approaches like the Neural Concept Binder [36] combined with explicit reasoning, e.g., via program synthesis [40].

## Acknowledgements

The authors would like to thank Frank Jäkel for useful discussions. This work was supported by the Priority Program (SPP) 2422 in the subproject "Optimization of active surface design of high-speed progressive tools using machine and deep learning algorithms" funded by the German Research Foundation (DFG). We acknowledge support of the hessian.AISC Service Center (funded by the Federal Ministry of Education and Research, BMBF, grant No 01IS22091), the "Third Wave of AI", and "The Adaptive Mind". Further, this work was funded by the European Union (Grant Agreement no. 101120763 - TANGO). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them. The authors of the Eindhoven University of Technology received support from their Department of Mathematics and Computer Science and the Eindhoven Artificial Intelligence Systems Institute.

## References

[1] Open AI. Gpt-4o. `https://openai.com/index/hello-gpt-4o/`, 2024. Accessed: 2024-09-25.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In *International Conference on Computer Vision (ICCV)*, pages 2425–2433. IEEE Computer Society, 2015.

[3] Antropic. Claude 3.5 sonnet. `https://www.anthropic.com/news/claude-3-5-sonnet`, 2024. Accessed: 2024-09-25.

[4] M.M. Bongard and J.K. Hawkins. *Pattern Recognition*. Spartan Books, 1970.

[5] Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, et al. An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247*, 2024.

[6] Shivam Chandhok, Wan-Cyuan Fan, and Leonid Sigal. Response wide shut: Surprising observations in basic vision language model capabilities. *arXiv preprint arXiv:2408.06721*, 2024.

[7] Hong-You Chen, Zhengfeng Lai, Haotian Zhang, Xinze Wang, Marcin Eichner, Keen You, Meng Cao, Bowen Zhang, Yinfei Yang, and Zhe Gan. Contrastive localized language-image pre-training. *arXiv preprint arXiv:2410.02746*, 2024.

[8] François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223. IEEE Computer Society, 2016.

[10] Stefan Depeweg, Constantin A Rothkopf, and Frank Jäkel. Solving bongard problems with a visual language and pragmatic reasoning. *Cognitive Science*, 2024.

[11] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models, 2024. URL `https://arxiv.org/abs/2407.11691`.

[12] Gregor Geigle, Radu Timofte, and Goran Glavaš. African or european swallow? benchmarking large vision-language models for fine-grained object classification. *arXiv preprint arXiv:2406.14496*, 2024.

[13] Chenhui Gou, Abdulwahab Felemban, Faizan Farooq Khan, Deyao Zhu, Jianfei Cai, Hamid Rezatofighi, and Mohamed Elhoseiny. How well can vision language models see image details? *arXiv preprint arXiv:2408.03940*, 2024.

[14] Lukas Helff, Wolfgang Stammer, Hikaru Shindo, Devendra Singh Dhami, and Kristian Kersting. V-lol: A diagnostic dataset for visual logical learning, 2023.

[15] José Hernández-Orallo, Fernando Martínez-Plumed, Ute Schmid, Michael Siebers, and David L Dowe. Computer models solving intelligence test problems: Progress and implications. *Artificial Intelligence*, 230:74–107, 2016.

[16] Yining Hong, Li Yi, Josh Tenenbaum, Antonio Torralba, and Chuang Gan. PTR: A benchmark for part-based conceptual, relational, and physical reasoning. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 17427–17440, 2021.

[17] Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6700–6709. Computer Vision Foundation / IEEE, 2019.

[18] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1988–1997. IEEE Computer Society, 2017.

[19] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[20] Laurynas Karazija, Iro Laina, and Christian Rupprecht. Clevrtex: A texture-rich benchmark for unsupervised multi-object segmentation. In *Conference on Neural Information Processing Systems (NeurIPS), NeurIPS Datasets and Benchmarks*, 2021.

[21] Bo Li, Peiyuan Zhang, Kaichen Zhang, Fanyi Pu, Xinrun Du, Yuhao Dong, Haotian Liu, Yuanhan Zhang, Ge Zhang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Accelerating the development of large multimoal models, March 2024. URL `https://github.com/EvolvingLMMs-Lab/lmms-eval`.

[22] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *European Conference on Computer Vision (ECCV)*, volume 8693, pages 740–755. Springer, 2014.

[23] Alexandre Linhares. A glimpse at the metaphysics of bongard problems. *Artificial Intelligence*, 121(1-2):251–270, 2000.

[24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[25] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL `https://llava-vl.github.io/blog/2024-01-30-llava-next/`.

[26] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024.

[27] David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 2010.

[28] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning, 2022. URL `https://arxiv.org/abs/2203.10244`.

[29] Arsenii Kirillovich Moskvichev, Victor Vikram Odouard, and Melanie Mitchell. The conceptarc benchmark: Evaluating understanding and generalization in the arc domain. *Transactions on machine learning research*, 2023.

[30] Marianna Nezhurina, Lucia Cipolina-Kun, Mehdi Cherti, and Jenia Jitsev. Alice in wonderland: Simple tasks showing complete reasoning breakdown in state-of-the-art large language models. *arXiv preprint arXiv:2406.02061*, 2024.

[31] Weili Nie, Zhiding Yu, Lei Mao, Ankit B Patel, Yuke Zhu, and Anima Anandkumar. Bongard-logo: A new benchmark for human-level concept learning and reasoning. *Advances in Neural Information Processing Systems*, 33:16468–16480, 2020.

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[33] Nikhil Raghuraman, Adam W Harley, and Leonidas Guibas. Cross-image context matters for bongard problems. *arXiv preprint arXiv:2309.03468*, 2023.

[34] Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. *arXiv preprint arXiv:2407.06581*, 2024.

[35] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: an open large-scale dataset for training next generation image-text models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[36] Wolfgang Stammer, Antonia Wüst, David Steinmann, and Kristian Kersting. Neural concept binder. *Advances in Neural Information Processing Systems*, 2024.

[37] Gemini Team. Gemini: A family of highly capable multimodal models, 2024. URL `https://arxiv.org/abs/2312.11805`.

[38] Ramakrishna Vedantam, Arthur Szlam, Maximilian Nickel, Ari Morcos, and Brenden M. Lake. CURI: A benchmark for productive concept learning under uncertainty. In *International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 10519–10529. PMLR, 2021.

[39] Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805*, 2024.

[40] Antonia Wüst, Wolfgang Stammer, Quentin Delfosse, Devendra Singh Dhami, and Kristian Kersting. Pix2code: Learning to compose neural visual concepts as programs. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024.

[41] Salahedine Youssef Youssef, Matej Zečević, Devendra Singh Dhami, and Kristian Kersting. Towards a solution to bongard problems: A causal approach. *arXiv preprint arXiv:2206.07196*, 2022.

[42] Liu Yuan, Duan Haodong, Zhang Yuanhan, Li Bo, Zhang Songyang, Zhao Wangbo, Yuan Yike, Wang Jiaqi, He Conghui, Liu Ziwei, Chen Kai, and Lin Dahua. Mmbench: Is your multi-modal model an all-around player? *arXiv:2307.06281*, 2023.

[43] 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. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024.

[44] Qiao Yuxuan, Duan Haodong, Fang Xinyu, Yang Junming, Chen Lin, Zhang Songyang, Wang Jiaqi, Lin Dahua, and Chen Kai. Prism: A framework for decoupling and assessing the capabilities of vlms. *arXiv preprint arXiv:2406.14544*, 2024.

[45] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6720–6731. Computer Vision Foundation / IEEE, 2019.

[46] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. RAVEN: A dataset for relational and analogical visual reasoning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5317–5327. Computer Vision Foundation / IEEE, 2019.

[47] Yizhe Zhang, He Bai, Ruixiang Zhang, Jiatao Gu, Shuangfei Zhai, Josh Susskind, and Navdeep Jaitly. How far are we from intelligent visual deductive reasoning? *ICLR 2024 Workshop: How Far Are We From AGI*, 2024.

[48] Kankan Zhou, Eason Lai, Wei Bin Au Yeong, Kyriakos Mouratidis, and Jing Jiang. Rome: Evaluating pre-trained vision-language models on reasoning beyond visual common sense. *EMNLP*, 2023.

# A Experimental Details

In the following, the prompts used during the experiments are provided. The prompt for the main experiment is shown in Listing 1. The prompt for the multiple choice setting is in Listing 2 and the prompt for the LLM-judge is in Listing 3. The prompts for the second part of the experiment are provided in Listings 4, 5, 6, 7,.

````
You are provided with a black-and-white image consisting of 12 simple diagrams. Each diagram represents shapes with specific features, such as geometric properties or higher-level concepts.

- The 6 diagrams on the left side belong to Set A.
- The 6 diagrams on the right side belong to Set B.

## Task:

Your task is to determine two distinct rules:

1. Set A Rule: Identify a rule that applies to all diagrams in Set A.
2. Set B Rule: Identify a separate rule that applies to all diagrams in Set B.

Important: The rule for Set A must not apply to any diagram in Set B, and the rule for Set B must not apply to any diagram in Set A.

## Step-by-Step Process:

1. Diagram Analysis: Carefully describe each diagram in detail, noting any geometric properties, patterns, or conceptual features.
2. Rule Derivation: Based on your analysis, deduce the rule for Set A and the rule for Set B, ensuring that each rule is unique to its set.

## Final Answer Format:

Provide the final answer using the following format:

```python

answer = {
    'set A rule': '[LEFT RULE]',
    'set B rule': '[RIGHT RULE]'
}
```

Ensure that the rules are clearly defined, concise, and do not overlap between the sets.
````

Listing 1: Prompt used in first experiment. The model is asked to provide rules for the left side and the ride side images of the Bongard Problem.

````

You are provided with a black-and-white image consisting of 12 simple diagrams. Each diagram represents shapes with specific features, such as geometric properties or higher-level concepts.

- The 6 diagrams on the left side belong to Set A.
- The 6 diagrams on the right side belong to Set B.

Additionally, you are given a list of possible rule pairs, one of which is true for this image. Your goal is to identify the correct rule pair based on the features of the diagrams in Set A and Set B.

## Task:

Your task is to identify and select the correct rule pair that is true for the sets. The rule pair is structured as follows:

1. Rule part 1: This rule should apply to all diagrams in Set A
2. Rule part 2: This rule should apply to all diagrams in Set B

Important: The rule for Set A must not apply to any diagram in Set B, and the rule for Set B must not apply to any diagram in Set A.

## Step-by-Step Process:

1. Diagram Analysis: Carefully describe each diagram in detail, noting any geometric properties, patterns, or conceptual features.
2. Rule Derivation: Based on your analysis of the diagrams, select one rule from the provided list for Set A and a different rule for Set B.

## Available Rules
You can choose from the following rule pairs:

<SOLUTIONS>

## Final Answer Format:

Provide the final answer using the following format:

```python

answer = {
    'answer': <Solution ID>,
}
```
Where <Solution ID> is the number corresponding to the correct rule pair that fits the criteria.
````

Listing 2: Prompt used in multiple choice experiment for solving BPs with solution options provided. The model is asked to select the rules for the left side and the ride side images of the BP that fits best. <SOLUTION> is replaced by a dictionary of the possible solutions the model can select from (either all 100 or a subset of 10).

````
You will be given a correct answer that states a rule for the left side and a rule for the right side of a visual pattern or scenario. You will also be given an answer from a model that attempts to describe these rules. Your task is to evaluate whether the model’s answer accurately reflects the intent and essence of the correct answer.
# Evaluation Criteria:

1. Semantic Accuracy: Does the model’s answer convey the same underlying concept or rule as the correct answer, even if the wording differs?
2. Logical Consistency: Is the model’s answer logically consistent with the correct answer's rules?
3. Relevance: Does the model's answer directly address the rules provided in the correct answer?

# Response Instructions:

- Respond with "answer": 1 if the model's answer is correct according to the criteria above.
- Respond with "answer": 0 if the model's answer is incorrect.
- If the model's answer is only partially correct, consider whether the partial match sufficiently conveys the intended rule. If it does, respond with "answer": 1; otherwise, respond with "answer": 0.

# Examples:
## Example 1:

- Correct Answer:
    - Left: Round shapes
    - Right: Angular shapes
- Model Answer:
    - Left: Circles
    - Right: Squares
- Expected Response:

```python
{
    "answer": 1
}
```

## Example 2:

- Correct Answer:
    - Left: Large shapes
    - Right: Small shapes
- Model Answer:
    - Left: Circular shapes
    - Right: Irregular shapes
- Expected Response:

```python
{
    "answer": 0
}
```

Use the format above to judge the correctness of the model's answer based on the given correct answer.

# Task
- Correct Answer:
    - Left: LEFT_RULE_SOLUTION
    - Right: RIGHT_RULE_SOLUTION
- Model Answer:
    - Left: LEFT_RULE_ANSWER
    - Right: RIGHT_RULE_ANSWER
- Response:
````

Listing 3: Prompt for LLM-judge used across the experiments. There are two example judgements provided. The judge needs to decide whether the answer is correct based on the provided ground truth (1) or incorrect (0).

````
Your task is to determine the direction in which a spiral depicted in a 2D black and white diagram is turning.
The given diagram shows a spiral-like shape. In which direction is the spiral turning, starting from the center?

Please decide carefully whether the spiral is turning in clockwise or counterclockwise direction. Take a deep breath and think step-by-step. Give your answer in the following format:
```
answer = {
    "direction": <your answer>
}
```
where <your answer> can be either "counterclockwise" or "clockwise".
````

Listing 4: Prompt for concepts of BP#16.

````
Your task is to determine if the number of objects inside a big shape is bigger than the number of objects outside of it in a 2D black and white diagram.
The given diagram shows a big shape that can contain smaller shapes inside it. There can also be other small shapes outside of the big shape. Is the number of shapes inside or outside the big shape higher?

Please decide carefully. Take a deep breath and think step-by-step. Give your answer in the following format:
```
answer = {
    "more shapes": <your answer>
}
```
where <your answer> can be either "inside" if the number of shapes inside the big shape is higher or "outside" if the number of shapes outside the big shape is higher.
````

Listing 5: Prompt for concepts of BP#29

````

Your task is to determine the relative position of two objects in a 2D black and white diagram.
The given diagram shows a triangle and a circle. Which of the shapes is located above the other, i.e., has a higher y-value than the other?

Please decide carefully. Take a deep breath and think step-by-step. Give your answer in the following format:
```
answer = {
    "shape": <your answer>
}
```
where <your answer> can be either "triangle" if the triangle is above the circle or "circle" if the circle is above the triangle.
````

Listing 6: Prompt for concepts of BP#36.

````
Your task is to determine the position of a circle in relation to a cavity in a 2D black and white diagram.

The given diagram shows a shape with a cavity. From inside the figure, you need to decide if the circle is on the left or the right of the cavity. Carefully analyze the diagram step-by-step to identify the correct side.

Please decide carefully. Take a deep breath and think step-by-step. Give your answer in the following format:

```python
answer = {
    "position": <your answer>
}
```
where <your answer> can be either "left" if the circle is to the left of the cavity or "right" if the circle is to the right of the cavity.
````

Listing 7: Prompt for concepts of BP#55.

## B Additional Results

In the following the detailed results of the evaluations are presented. In Table 3, Table 4 and Table 5 the results for the single BPs for each model are reported. Please note that LLaVA-v1.6-34b could not be considered for Table 4 since the context size of the model was too small to consider all 100 options.

In Table 6 we compare the results of the VLMs on a subset of BPs against human results from [10]. Each BP has been categorized based on the ground truth rules, the possible categories are existence, size, concept, number and spatial. Existence means that the rules correspond to the existence or absence of something in the diagrams. For size, the rule is based on the size of one or multiple shapes in the BP. Concept means, that the BP tests for a specific, more abstract concept, e.g., *convex* and *concave*. Under number BPs are grouped that require some form of counting, e.g., *one* vs *two* shapes. And finally spatial takes into account BPs that require spatial reasoning, e.g., some *shape is on top of the other*.

Further, we report the classification results of the second part of the experiments for the concepts of BP#16 (Table 7), BP#29 (Table 8, BP#36 (Table 9) and BP#55 (Table 10).

Table 3: Results of each VLM on the individual Bongard Problems. Each model was prompted three times and the number of correct responses is reported (of 3).

| BP# | gpt-4o | claude | gemini | llava 1.6 | llava 1.5 |
|---|---|---|---|---|---|
| 1 | 2/3 | 3/3 | 3/3 | 0/3 | 0/3 |
| 2 | 3/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 3 | 3/3 | 3/3 | 3/3 | 0/3 | 0/3 |
| 4 | 0/3 | 1/3 | 0/3 | 0/3 | 0/3 |
| 5 | 2/3 | 3/3 | 3/3 | 2/3 | 0/3 |
| 6 | 3/3 | 2/3 | 0/3 | 0/3 | 2/3 |
| 7 | 1/3 | 1/3 | 0/3 | 0/3 | 0/3 |
| 8 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 9 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 10 | 2/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 11 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 12 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 13 | 1/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 14 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 15 | 3/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 16 | 1/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 17 | 0/3 | 1/3 | 0/3 | 0/3 | 0/3 |
| 18 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 19 | 0/3 | 0/3 | 0/3 | 0/3 | 2/3 |
| 20 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 21 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 22 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 23 | 3/3 | 3/3 | 1/3 | 0/3 | 0/3 |
| 24 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 25 | 3/3 | 1/3 | 0/3 | 0/3 | 0/3 |
| 26 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 27 | 1/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 28 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 29 | 0/3 | 1/3 | 0/3 | 0/3 | 0/3 |
| 30 | 3/3 | 1/3 | 1/3 | 0/3 | 0/3 |
| 31 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 32 | 3/3 | 2/3 | 0/3 | 0/3 | 0/3 |
| 33 | 2/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 34 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 35 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 36 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 37 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 38 | 0/3 | 2/3 | 0/3 | 0/3 | 0/3 |
| 39 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 40 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 41 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 42 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 43 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 44 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 45 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 46 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 47 | 3/3 | 3/3 | 0/3 | 0/3 | 0/3 |
| 48 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 49 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 50 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 51 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 52 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 53 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 54 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 55 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 56 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 57 | 1/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 58 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 59 | 3/3 | 1/3 | 0/3 | 0/3 | 0/3 |
| 60 | 0/3 | 1/3 | 0/3 | 0/3 | 0/3 |
| 61 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 62 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 63 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 64 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 65 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 66 | 1/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 67 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 68 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 69 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 70 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 71 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 72 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 73 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 74 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 75 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 76 | 0/3 | 2/3 | 0/3 | 0/3 | 0/3 |
| 77 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 78 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 79 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 80 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 81 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 82 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 83 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 84 | 3/3 | 0/3 | 1/3 | 0/3 | 0/3 |
| 85 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 86 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 87 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 88 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 89 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 90 | 0/3 | 0/3 | 0/3 | 1/3 | 0/3 |
| 91 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 92 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 93 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 94 | 2/3 | 1/3 | 0/3 | 0/3 | 0/3 |
| 95 | 3/3 | 3/3 | 0/3 | 0/3 | 0/3 |
| 96 | 3/3 | 0/3 | 0/3 | 1/3 | 0/3 |
| 97 | 3/3 | 3/3 | 0/3 | 2/3 | 0/3 |
| 98 | 3/3 | 1/3 | 3/3 | 0/3 | 0/3 |
| 99 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 100 | 3/3 | 3/3 | 1/3 | 0/3 | 0/3 |

Table 4: Results of each VLM on the individual Bongard Problems when provided with all possible solutions. Each model was prompted three times and the number of correct responses is reported (of 3).

| BP# | gpt-4o | claude | gemini | llava 1.5 |
|---|---|---|---|---|
| 1 | 3/3 | 3/3 | 3/3 | 0/3 |
| 2 | 1/3 | 3/3 | 0/3 | 0/3 |
| 3 | 3/3 | 3/3 | 3/3 | 0/3 |
| 4 | 0/3 | 2/3 | 2/3 | 0/3 |
| 5 | 3/3 | 3/3 | 3/3 | 0/3 |
| 6 | 3/3 | 3/3 | 3/3 | 0/3 |
| 7 | 3/3 | 3/3 | 0/3 | 0/3 |
| 8 | 0/3 | 0/3 | 0/3 | 0/3 |
| 9 | 3/3 | 3/3 | 3/3 | 0/3 |
| 10 | 3/3 | 3/3 | 3/3 | 2/3 |
| 11 | 0/3 | 1/3 | 0/3 | 0/3 |
| 12 | 0/3 | 0/3 | 0/3 | 0/3 |
| 13 | 0/3 | 2/3 | 3/3 | 0/3 |
| 14 | 0/3 | 0/3 | 0/3 | 0/3 |
| 15 | 0/3 | 0/3 | 0/3 | 0/3 |
| 16 | 2/3 | 0/3 | 0/3 | 0/3 |
| 17 | 0/3 | 0/3 | 0/3 | 0/3 |
| 18 | 0/3 | 0/3 | 0/3 | 0/3 |
| 19 | 0/3 | 0/3 | 0/3 | 0/3 |
| 20 | 0/3 | 1/3 | 1/3 | 0/3 |
| 21 | 0/3 | 0/3 | 0/3 | 0/3 |
| 22 | 0/3 | 0/3 | 0/3 | 0/3 |
| 23 | 3/3 | 3/3 | 0/3 | 0/3 |
| 24 | 1/3 | 3/3 | 3/3 | 0/3 |
| 25 | 2/3 | 0/3 | 1/3 | 0/3 |
| 26 | 0/3 | 1/3 | 2/3 | 0/3 |
| 27 | 1/3 | 2/3 | 0/3 | 0/3 |
| 28 | 0/3 | 1/3 | 0/3 | 0/3 |
| 29 | 1/3 | 2/3 | 3/3 | 0/3 |
| 30 | 3/3 | 2/3 | 3/3 | 0/3 |
| 31 | 0/3 | 2/3 | 0/3 | 0/3 |
| 32 | 0/3 | 0/3 | 0/3 | 0/3 |
| 33 | 0/3 | 1/3 | 0/3 | 0/3 |
| 34 | 0/3 | 2/3 | 0/3 | 0/3 |
| 35 | 0/3 | 1/3 | 0/3 | 0/3 |
| 36 | 2/3 | 3/3 | 0/3 | 0/3 |
| 37 | 1/3 | 0/3 | 0/3 | 0/3 |
| 38 | 0/3 | 1/3 | 0/3 | 0/3 |
| 39 | 2/3 | 2/3 | 3/3 | 0/3 |
| 40 | 0/3 | 1/3 | 0/3 | 0/3 |
| 41 | 0/3 | 0/3 | 0/3 | 0/3 |
| 42 | 0/3 | 0/3 | 0/3 | 0/3 |
| 43 | 0/3 | 1/3 | 0/3 | 0/3 |
| 44 | 0/3 | 0/3 | 0/3 | 0/3 |
| 45 | 0/3 | 1/3 | 0/3 | 0/3 |
| 46 | 0/3 | 0/3 | 0/3 | 0/3 |
| 47 | 3/3 | 2/3 | 0/3 | 0/3 |
| 48 | 0/3 | 0/3 | 0/3 | 0/3 |
| 49 | 0/3 | 0/3 | 0/3 | 0/3 |
| 50 | 0/3 | 0/3 | 0/3 | 0/3 |
| 51 | 0/3 | 0/3 | 0/3 | 0/3 |
| 52 | 0/3 | 0/3 | 0/3 | 0/3 |
| 53 | 0/3 | 3/3 | 0/3 | 0/3 |
| 54 | 0/3 | 1/3 | 0/3 | 0/3 |
| 55 | 0/3 | 1/3 | 0/3 | 0/3 |
| 56 | 0/3 | 0/3 | 0/3 | 0/3 |
| 57 | 0/3 | 2/3 | 0/3 | 0/3 |
| 58 | 0/3 | 0/3 | 0/3 | 0/3 |
| 59 | 0/3 | 0/3 | 0/3 | 0/3 |
| 60 | 0/3 | 0/3 | 0/3 | 0/3 |
| 61 | 0/3 | 1/3 | 0/3 | 0/3 |
| 62 | 0/3 | 0/3 | 0/3 | 0/3 |
| 63 | 0/3 | 0/3 | 0/3 | 0/3 |
| 64 | 0/3 | 0/3 | 0/3 | 0/3 |
| 65 | 0/3 | 0/3 | 0/3 | 0/3 |
| 66 | 0/3 | 0/3 | 0/3 | 0/3 |
| 67 | 1/3 | 1/3 | 2/3 | 0/3 |
| 68 | 0/3 | 0/3 | 0/3 | 0/3 |
| 69 | 0/3 | 1/3 | 3/3 | 0/3 |
| 70 | 0/3 | 1/3 | 0/3 | 0/3 |
| 71 | 0/3 | 0/3 | 0/3 | 0/3 |
| 72 | 0/3 | 0/3 | 0/3 | 0/3 |
| 73 | 0/3 | 0/3 | 0/3 | 0/3 |
| 74 | 0/3 | 0/3 | 0/3 | 0/3 |
| 75 | 0/3 | 0/3 | 0/3 | 0/3 |
| 76 | 0/3 | 0/3 | 0/3 | 0/3 |
| 77 | 0/3 | 0/3 | 0/3 | 0/3 |
| 78 | 0/3 | 0/3 | 0/3 | 0/3 |
| 79 | 0/3 | 0/3 | 0/3 | 0/3 |
| 80 | 0/3 | 0/3 | 0/3 | 0/3 |
| 81 | 0/3 | 0/3 | 0/3 | 0/3 |
| 82 | 0/3 | 0/3 | 0/3 | 0/3 |
| 83 | 0/3 | 0/3 | 0/3 | 0/3 |
| 84 | 0/3 | 0/3 | 0/3 | 0/3 |
| 85 | 2/3 | 0/3 | 0/3 | 0/3 |
| 86 | 3/3 | 0/3 | 0/3 | 0/3 |
| 87 | 0/3 | 0/3 | 0/3 | 0/3 |
| 88 | 0/3 | 0/3 | 0/3 | 0/3 |
| 89 | 3/3 | 0/3 | 0/3 | 0/3 |
| 90 | 0/3 | 0/3 | 0/3 | 0/3 |
| 91 | 0/3 | 0/3 | 0/3 | 0/3 |
| 92 | 0/3 | 0/3 | 0/3 | 0/3 |
| 93 | 0/3 | 0/3 | 0/3 | 0/3 |
| 94 | 2/3 | 3/3 | 2/3 | 0/3 |
| 95 | 3/3 | 1/3 | 0/3 | 0/3 |
| 96 | 3/3 | 0/3 | 0/3 | 2/3 |
| 97 | 3/3 | 1/3 | 0/3 | 0/3 |
| 98 | 3/3 | 3/3 | 3/3 | 1/3 |
| 99 | 0/3 | 0/3 | 0/3 | 0/3 |
| 100 | 3/3 | 3/3 | 0/3 | 0/3 |

Table 5: Results of each VLM on the individual Bongard Problems when provided with a selection of 10 possible solutions. Each model was prompted three times and the number of correct responses is reported (of 3).

| BP# | gpt-4o | claude | gemini | llava 1.6 | llava 1.5 |
|---|---|---|---|---|---|
| 1 | 3/3 | 3/3 | 3/3 | 0/3 | 1/3 |
| 2 | 2/3 | 1/3 | 0/3 | 1/3 | 0/3 |
| 3 | 3/3 | 3/3 | 2/3 | 0/3 | 0/3 |
| 4 | 3/3 | 3/3 | 1/3 | 0/3 | 0/3 |
| 5 | 3/3 | 3/3 | 3/3 | 0/3 | 0/3 |
| 6 | 3/3 | 3/3 | 3/3 | 0/3 | 0/3 |
| 7 | 3/3 | 3/3 | 3/3 | 0/3 | 1/3 |
| 8 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 9 | 3/3 | 3/3 | 3/3 | 0/3 | 0/3 |
| 10 | 3/3 | 3/3 | 3/3 | 0/3 | 0/3 |
| 11 | 3/3 | 2/3 | 1/3 | 0/3 | 0/3 |
| 12 | 3/3 | 2/3 | 1/3 | 1/3 | 0/3 |
| 13 | 3/3 | 3/3 | 3/3 | 2/3 | 0/3 |
| 14 | 3/3 | 0/3 | 0/3 | 1/3 | 0/3 |
| 15 | 2/3 | 2/3 | 0/3 | 0/3 | 0/3 |
| 16 | 3/3 | 3/3 | 1/3 | 0/3 | 3/3 |
| 17 | 2/3 | 2/3 | 2/3 | 1/3 | 0/3 |
| 18 | 3/3 | 2/3 | 0/3 | 0/3 | 0/3 |
| 19 | 3/3 | 2/3 | 0/3 | 1/3 | 0/3 |
| 20 | 3/3 | 2/3 | 2/3 | 0/3 | 0/3 |
| 21 | 3/3 | 3/3 | 3/3 | 0/3 | 0/3 |
| 22 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 23 | 3/3 | 3/3 | 3/3 | 0/3 | 0/3 |
| 24 | 3/3 | 3/3 | 2/3 | 0/3 | 0/3 |
| 25 | 3/3 | 0/3 | 3/3 | 1/3 | 0/3 |
| 26 | 2/3 | 3/3 | 3/3 | 0/3 | 0/3 |
| 27 | 2/3 | 2/3 | 1/3 | 1/3 | 0/3 |
| 28 | 2/3 | 2/3 | 1/3 | 2/3 | 0/3 |
| 29 | 3/3 | 3/3 | 2/3 | 2/3 | 0/3 |
| 30 | 3/3 | 2/3 | 3/3 | 1/3 | 0/3 |
| 31 | 1/3 | 1/3 | 1/3 | 0/3 | 0/3 |
| 32 | 3/3 | 2/3 | 3/3 | 1/3 | 0/3 |
| 33 | 2/3 | 3/3 | 2/3 | 0/3 | 0/3 |
| 34 | 1/3 | 2/3 | 3/3 | 0/3 | 0/3 |
| 35 | 2/3 | 3/3 | 3/3 | 1/3 | 0/3 |
| 36 | 2/3 | 3/3 | 2/3 | 2/3 | 0/3 |
| 37 | 2/3 | 0/3 | 2/3 | 1/3 | 0/3 |
| 38 | 2/3 | 3/3 | 0/3 | 1/3 | 0/3 |
| 39 | 3/3 | 3/3 | 3/3 | 3/3 | 0/3 |
| 40 | 3/3 | 3/3 | 3/3 | 3/3 | 0/3 |
| 41 | 0/3 | 2/3 | 0/3 | 2/3 | 0/3 |
| 42 | 1/3 | 2/3 | 0/3 | 0/3 | 0/3 |
| 43 | 3/3 | 2/3 | 0/3 | 3/3 | 0/3 |
| 44 | 3/3 | 2/3 | 3/3 | 0/3 | 0/3 |
| 45 | 2/3 | 3/3 | 2/3 | 0/3 | 0/3 |
| 46 | 1/3 | 2/3 | 2/3 | 0/3 | 0/3 |
| 47 | 3/3 | 3/3 | 3/3 | 2/3 | 1/3 |
| 48 | 1/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 49 | 1/3 | 3/3 | 1/3 | 1/3 | 0/3 |
| 50 | 0/3 | 1/3 | 3/3 | 1/3 | 0/3 |
| 51 | 2/3 | 1/3 | 2/3 | 0/3 | 0/3 |
| 52 | 3/3 | 3/3 | 2/3 | 3/3 | 0/3 |
| 53 | 2/3 | 3/3 | 3/3 | 1/3 | 0/3 |
| 54 | 2/3 | 3/3 | 0/3 | 1/3 | 0/3 |
| 55 | 0/3 | 2/3 | 1/3 | 0/3 | 0/3 |
| 56 | 3/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 57 | 2/3 | 3/3 | 0/3 | 1/3 | 0/3 |
| 58 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 59 | 2/3 | 2/3 | 0/3 | 0/3 | 0/3 |
| 60 | 1/3 | 0/3 | 1/3 | 0/3 | 0/3 |
| 61 | 3/3 | 3/3 | 3/3 | 1/3 | 0/3 |
| 62 | 2/3 | 3/3 | 3/3 | 0/3 | 0/3 |
| 63 | 1/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 64 | 2/3 | 1/3 | 2/3 | 0/3 | 1/3 |
| 65 | 1/3 | 0/3 | 0/3 | 0/3 | 0/3 |
| 66 | 3/3 | 1/3 | 0/3 | 0/3 | 0/3 |
| 67 | 3/3 | 3/3 | 3/3 | 2/3 | 0/3 |
| 68 | 3/3 | 2/3 | 3/3 | 2/3 | 0/3 |
| 69 | 2/3 | 3/3 | 2/3 | 3/3 | 0/3 |
| 70 | 3/3 | 3/3 | 3/3 | 3/3 | 0/3 |
| 71 | 1/3 | 2/3 | 3/3 | 0/3 | 0/3 |
| 72 | 3/3 | 0/3 | 2/3 | 0/3 | 0/3 |
| 73 | 1/3 | 1/3 | 0/3 | 0/3 | 0/3 |
| 74 | 1/3 | 3/3 | 0/3 | 2/3 | 0/3 |
| 75 | 1/3 | 3/3 | 3/3 | 0/3 | 0/3 |
| 76 | 2/3 | 3/3 | 1/3 | 0/3 | 0/3 |
| 77 | 1/3 | 2/3 | 1/3 | 0/3 | 0/3 |
| 78 | 1/3 | 1/3 | 3/3 | 0/3 | 0/3 |
| 79 | 1/3 | 3/3 | 2/3 | 0/3 | 0/3 |
| 80 | 1/3 | 1/3 | 2/3 | 0/3 | 0/3 |
| 81 | 2/3 | 0/3 | 3/3 | 1/3 | 0/3 |
| 82 | 0/3 | 1/3 | 2/3 | 2/3 | 0/3 |
| 83 | 3/3 | 3/3 | 3/3 | 0/3 | 0/3 |
| 84 | 3/3 | 3/3 | 3/3 | 0/3 | 0/3 |
| 85 | 3/3 | 3/3 | 3/3 | 3/3 | 0/3 |
| 86 | 3/3 | 3/3 | 3/3 | 3/3 | 0/3 |
| 87 | 0/3 | 1/3 | 0/3 | 1/3 | 0/3 |
| 88 | 0/3 | 0/3 | 3/3 | 3/3 | 0/3 |
| 89 | 3/3 | 1/3 | 3/3 | 3/3 | 0/3 |
| 90 | 0/3 | 2/3 | 0/3 | 0/3 | 0/3 |
| 91 | 0/3 | 1/3 | 0/3 | 0/3 | 0/3 |
| 92 | 0/3 | 2/3 | 3/3 | 0/3 | 0/3 |
| 93 | 1/3 | 1/3 | 1/3 | 0/3 | 0/3 |
| 94 | 3/3 | 3/3 | 1/3 | 0/3 | 0/3 |
| 95 | 3/3 | 3/3 | 2/3 | 0/3 | 0/3 |
| 96 | 3/3 | 3/3 | 3/3 | 0/3 | 0/3 |
| 97 | 3/3 | 3/3 | 3/3 | 0/3 | 0/3 |
| 98 | 3/3 | 3/3 | 3/3 | 0/3 | 0/3 |
| 99 | 2/3 | 3/3 | 2/3 | 0/3 | 0/3 |
| 100 | 3/3 | 3/3 | 2/3 | 3/3 | 0/3 |

Table 6: Results of each VLM on the individual Bongard Problems compared to the results of 13 humans from [10]. Each model was prompted three times and the number of correct responses is reported (of 3).

| BP# | Category | GPT-4o | Claude | Gemini | LLaVA 1.6 | LLaVA 1.5 | Humans |
|---|---|---|---|---|---|---|---|
| 1 | existence | 13/13 | 12/13 | 12/13 | 0/13 | 0/13 | 13/13 |
| 2 | size | 1/13 | 4/13 | 0/13 | 0/13 | 0/13 | 11/13 |
| 3 | concept | 13/13 | 13/13 | 5/13 | 0/13 | 0/13 | 13/13 |
| 4 | concept | 0/13 | 1/13 | 0/13 | 0/13 | 0/13 | 3/13 |
| 6 | concept | 13/13 | 5/13 | 0/13 | 1/13 | 4/13 | 12/13 |
| 7 | concept | 7/13 | 8/13 | 0/13 | 0/13 | 0/13 | 13/13 |
| 8 | concept | 0/13 | 0/13 | 0/13 | 0/13 | 0/13 | 13/13 |
| 9 | concept | 0/13 | 2/13 | 1/13 | 0/13 | 0/13 | 12/13 |
| 11 | concept | 0/13 | 0/13 | 0/13 | 0/13 | 0/13 | 8/13 |
| 12 | concept | 2/13 | 0/13 | 0/13 | 0/13 | 0/13 | 7/13 |
| 21 | size | 0/13 | 4/13 | 0/13 | 0/13 | 0/13 | 10/12 |
| 22 | size | 0/13 | 0/13 | 0/13 | 0/13 | 0/13 | 11/12 |
| 23 | number | 4/13 | 13/13 | 3/13 | 2/13 | 0/13 | 11/11 |
| 24 | existence | 0/13 | 3/13 | 0/13 | 0/13 | 0/13 | 10/11 |
| 25 | concept | 8/13 | 2/13 | 0/13 | 2/13 | 0/13 | 9/11 |
| 26 | existence | 0/13 | 0/13 | 0/13 | 0/13 | 0/13 | 9/10 |
| 27 | number | 0/13 | 0/13 | 0/13 | 0/13 | 0/13 | 8/10 |
| 28 | number | 0/13 | 0/13 | 0/13 | 0/13 | 0/13 | 4/10 |
| 29 | number | 0/13 | 0/13 | 0/13 | 0/13 | 0/13 | 7/10 |
| 34 | size | 0/13 | 0/13 | 0/13 | 0/13 | 0/13 | 9/9 |
| 35 | spatial | 0/13 | 1/13 | 1/13 | 0/13 | 0/13 | 9/9 |
| 36 | spatial | 0/13 | 2/13 | 0/13 | 0/13 | 0/13 | 9/9 |
| 37 | spatial | 0/13 | 0/13 | 0/13 | 0/13 | 0/13 | 3/8 |
| 38 | size | 0/13 | 6/13 | 0/13 | 0/13 | 0/13 | 6/8 |
| 39 | spatial | 0/13 | 2/13 | 0/13 | 0/13 | 0/13 | 8/8 |
| 40 | spatial | 0/13 | 0/13 | 0/13 | 0/13 | 0/13 | 8/8 |
| 41 | spatial | 0/13 | 0/13 | 0/13 | 0/13 | 0/13 | 8/8 |
| 42 | spatial | 0/13 | 0/13 | 0/13 | 0/13 | 0/13 | 7/7 |
| 47 | spatial | 13/13 | 6/13 | 9/13 | 0/13 | 0/13 | 6/6 |
| 48 | spatial | 0/13 | 0/13 | 0/13 | 0/13 | 0/13 | 6/6 |
| 49 | spatial | 0/13 | 0/13 | 0/13 | 0/13 | 0/13 | 6/6 |
| All solved | | 4 | 2 | 0 | 0 | 0 | **15** |
| 2/3 solved | | 4 | 3 | 2 | 0 | 0 | **26** |
| Once solved | | 9 | 16 | 6 | 3 | 1 | **31** |

Table 7: **BP#16.** Classification results when providing the single images of BP#16 and asking for clockwise or counterclockwise.

| | Clockwise | | | | | | Counter-Clockwise | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| GPT-4o | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 | 2/3 | 3/3 | 3/3 | 2/3 | 3/3 | 3/3 |
| Claude | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 | 0/3 | 3/3 | 3/3 | 2/3 | 2/3 | 0/3 | 3/3 |
| Gemini | 2/3 | 0/3 | 3/3 | 2/3 | 1/3 | 0/3 | 3/3 | 1/3 | 3/3 | 2/3 | 1/3 | 0/3 |
| LLaVA 1.6 | 2/3 | 2/3 | 2/3 | 1/3 | 1/3 | 2/3 | 1/3 | 2/3 | 2/3 | 3/3 | 1/3 | 2/3 |

Table 8: **BP#29.** Correctly classified concepts of BP#29. Models were asked wether there are more shapes inside or outside the big figure.

| | Inside | | | | | | Outside | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| GPT-4o | 3/3 | 3/3 | 3/3 | 3/3 | 3/3 | 3/3 | 3/3 | 3/3 | 1/3 | 3/3 | 0/3 | 3/3 |
| Claude | 3/3 | 3/3 | 3/3 | 3/3 | 3/3 | 3/3 | 3/3 | 3/3 | 3/3 | 3/3 | 3/3 | 3/3 |
| Gemini | 3/3 | 3/3 | 0/3 | 1/3 | 1/3 | 0/3 | 3/3 | 3/3 | 3/3 | 3/3 | 3/3 | 3/3 |
| LLaVA 1.6 | 3/3 | 3/3 | 3/3 | 3/3 | 3/3 | 3/3 | 1/3 | 1/3 | 0/3 | 0/3 | 0/3 | 0/3 |

Table 9: **BP#36.** Correctly classified concepts for BP#36. Models were asked to output whether triangle or circle is on top.

| | Triangle | | | | | | Circle | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| GPT-4o | 3/3 | 3/3 | 3/3 | 3/3 | 3/3 | 3/3 | 3/3 | 1/3 | 3/3 | 3/3 | 3/3 | 3/3 |
| Claude | 3/3 | 3/3 | 3/3 | 3/3 | 3/3 | 3/3 | 3/3 | 3/3 | 3/3 | 3/3 | 3/3 | 3/3 |
| Gemini | 3/3 | 3/3 | 2/3 | 3/3 | 3/3 | 2/3 | 3/3 | 3/3 | 3/3 | 3/3 | 3/3 | 3/3 |
| LLaVA 1.6 | 2/3 | 1/3 | 0/3 | 0/3 | 0/3 | 3/3 | 3/3 | 3/3 | 3/3 | 3/3 | 3/3 | 3/3 |

Table 10: **BP#55.** Correctly classified for concepts of BP#55. Models were asked whether the circle shape is located left or right from the cavity in the big shape (viewed from inside the shape).

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GPT-4o | 0/3 | 3/3 | 3/3 | 3/3 | 0/3 | 0/3 | 3/3 | 0/3 | 0/3 | 3/3 | 0/3 | 3/3 |
| Claude | 0/3 | 3/3 | 1/3 | 3/3 | 0/3 | 0/3 | 1/3 | 0/3 | 0/3 | 2/3 | 0/3 | 3/3 |
| Gemini | 0/3 | 2/3 | 3/3 | 2/3 | 0/3 | 0/3 | 3/3 | 0/3 | 0/3 | 0/3 | 0/3 | 3/3 |
| LLaVA 1.6 | 0/3 | 2/3 | 1/3 | 1/3 | 0/3 | 0/3 | 3/3 | 1/3 | 0/3 | 1/3 | 0/3 | 3/3 |