Bongard in Wonderland: Visual Puzzles that Still Make AI Go Mad?

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recently, newly developed Vision-Language Models (VLMs), such as OpenAI's GPT-4o, have emerged, claiming to take complex reasoning to a new level. 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 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 [1]. 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) [2, 3, 4]. 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 [5] and visual perception [6, 7, 8] or their combination [9, 10, 11]

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 [12]. 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 [13].

Unlike pattern recognition in classification tasks, BPs are not about finding visual patterns that match certain concepts but about finding concepts that allow for a pattern in the description of the diagrams

[1]Our code and evaluation framework will be publicly available following the publication of this work.

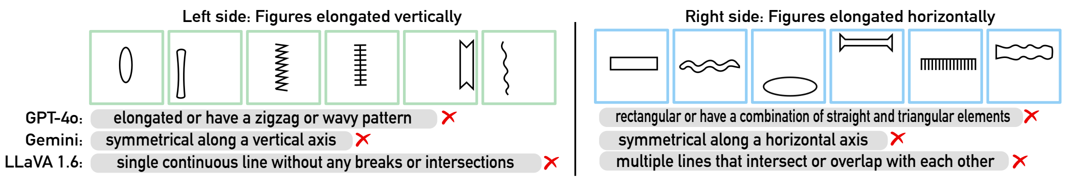

Figure 1: **VLMs struggle to solve BPs out of the box**. Although the concepts of *vertical* and *horizontal* may seem trivial to a human, the VLMs struggle to generate discriminative rules.

that matches the left-right separation. Thus, 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. An example BP is shown in Figure 1.

While traditional machine learning approaches have achieved some success with BPs [14, 15], the potential of VLMs remains largely unexplored. Since VLMs already struggle with recognizing rather simple visual patterns, as shown by [6] and [9], 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. 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. [15] 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. [15]. Raghuraman et al. [14] 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. [16] 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 [17, 18, 19, 20]. 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 [21, 22, 6, 23, 24, 25, 26, 27, 28, 29, 30]. Yet, most of these only require simple reasoning abilities. More recent benchmarks have been introduced to probe advanced reasoning skills, e.g., logical learning [31, 32], mathematical reasoning [33] or analogical visual reasoning [34]. Although this shift towards more cognitively demanding tasks is promising, comprehensive diagnostic evaluations of VLMs' reasoning 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 Method

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 1. The task is to find a linguistic expression of the underlying rule

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	24	2

that distinguishes the two groups. To analyze the VLM's capabilities, we evaluate two approaches: For the first approach, we provide the context to split the BP challenge into a *description* task and a *reasoning* task. This is done by step-by-step instructions inside a single prompt for the first approach (cf. 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). For the second approach, we investigate the limitations with visual descriptions via a *perception* task. Here, the relevant concept for the BP is provided and the task is to predict for each image of the BP whether it belongs to the concept or not. For this four specific prompts were implemented (Listings 4, 5, 6, 7).

4 Experiments

In our experiments, we investigate to what extent state-of-the-art VLMs can solve Bongard problems. At first, we evaluate the models quantitatively on all 100 puzzles and then investigate them qualitatively in more detail. For our evaluations, we consider the 100 original Bongard problems of [12]. We use the dataset variation of [15], which contains high-resolution images of the original diagrams. For the evaluations we use the models GPT-4o [35], LLaVA versions v1.6-34b [3] and v1.5-13b [4], Gemini 1.5 Pro 36, and Claude 3.5 Sonnet [37]. For the LLM-judge, we use GPT-4o.

Can VLMs solve Bongard problems? As a first step, we want to investigate to what extent current state-of-the-art VLMs can solve Bongard problems. For this, we ask our selection of VLMs to solve each BP three times. The answers are given to the LLM-judge, which decides for each answer whether it solves the BP or not. The results of this evaluation can be seen in the top row of Table 1. We can see that GPT-4o is by far the best-performing models with an average of 22 solved BPs. However, this performance is still surprisingly poor, especially considering human abilities [15, 38, 39]. In Table 2 we provide a more detailed overview of which BPs were solved. It shows that even rather simple BPs with concepts like "small vs. large shapes" (BP#2) and "left vs. right figures" (BP#8) are not solved correctly in most attempts.

In a further setting, we analyze how the results change when providing the models with all existing rule pairs of the BPs and ask them to select the correct one (multiple choice). 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 reduce the number of options to 10 possible rule pairs and repeat the procedure. Now, the models perform better, reaching up to 69 solved BPs. This is interesting since before the correct solution was present as well, but the models could not select it correctly. 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 want to investigate this next using the example of four selected BPs.

Why do VLMs fail to solve Bongard problems? We saw that the VLMs show poor performance on the BP dataset. This can be due to issues with the perception of the diagrams of the BPs but also due to reasoning failures, i.e., when creating rules that apply to each side distinctively. None of the models was able to solve BP#16, BP#19, BP#29. and BP#36 correctly even though the conceptual complexity of the rules is rather small. We investigate this in more detail by providing individual images of the BPs to the models and asking them directly for the relevant concepts. An excerpt of the

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

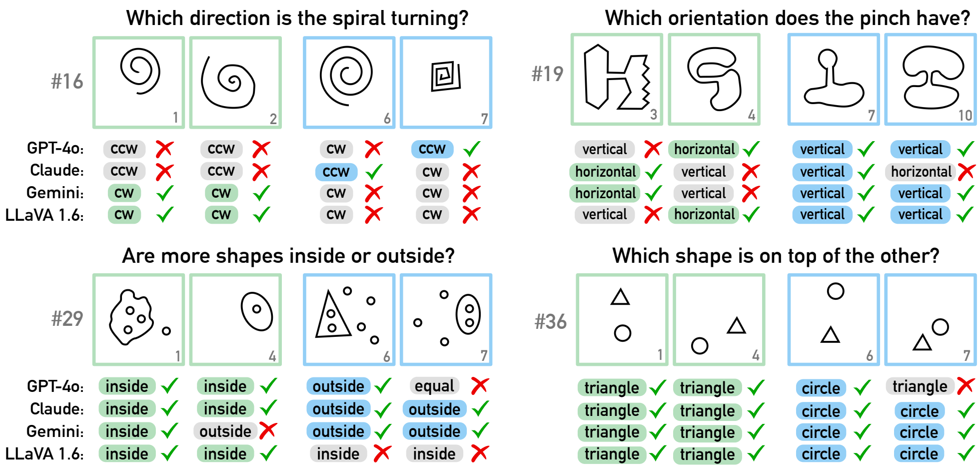

Figure 2: **VLMs fail to identify simple visual concepts.** VLMs challenged with identifying visual concepts in BPs. Although the VLM is able to recognize some of the concepts when specifically asked for (bottom), on the others, it continues to falter (top).

responses is displayed in Figure 2 (cf. Tab. 5, 6, 7, 8). When all images from the BP are correctly categorized, we take it as an indication that the VLM has likely captured the concept.

Surprisingly, we find that for BP#16, 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. For BP#19 there is a similar behavior where none of the models is able to classify the concepts of all images correctly. For BP#29, on the other hand, GPT-4o, for example, is convinced that image 7 has as many shapes on the outside as on the inside. Except for Claude, none of the models could count the shapes correctly, even though the final decision was primarily correct. The observed behavior is remarkable and shows that perception is the key issue for not identifying the correct rules. For the last BP, BP#36, the models can identify the concept better, with some even classifying all 12 images correctly (cf. Table 8). Here, the perception seems to work more reliably and the problem for solving the BP from scratch might be more on the pattern recognition or reasoning side.

Limitations. While useful for assessing abstract reasoning, BPs represent a small and highly specialized set of challenges, which might not fully capture the diverse challenges VLMs face 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 Discussion and Future Work

This work presented a diagnostic evaluation of VLMs using the classical Bongard problems, providing valuable insides into their current capabilities of pattern recognition and abstract reasoning. Our experiments 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 Bongard Problems, 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. E.g. concepts that appear trivial to humans, such as simple spirals, posed considerable challenges to these models. 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? 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. Such an analysis could help pinpoint the specific sources of error and identify whether these failures emerge from perceptual shortcomings, reasoning limitations, or both.

References

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

[2] 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.

[3] 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.

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

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

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

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images, 2016.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

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

[32] 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.

[33] 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.

[34] 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.

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

[36] Gemini Team. Gemini: A family of highly capable multimodal models, 2024.

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

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

[39] 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.

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 the orientation of a pinch or neck in a 2D black and white diagram.
The given diagram shows a shape that has a pinch in the middle. The pinch connects the both ends of the shapes and can be interpreted as a bridge as well. Which orientation does this pinch have, horizontal or vertical?

Please decide carefully. Take a deep breath and think step-by-step. Give your answer in the following format:
```
answer = {
    "orientation": <your answer>
}
```
where <your answer> can be either "horizontal" if the bridge is aligned horizontal or "vertical" if the bridge is aligned vertical.
````

Listing 5: Prompt for concepts of BP#19.

````
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 6: 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 7: Prompt for concepts of BP#36.

Table 2: 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

B Additional Results

In the following the detailed results of the evaluations are presented. In Table 2, Table 3 and Table 4 the results for the single BPs for each model are reported. Please note that LLaVA-v1.6-34b could not be considered for Table 3 since the context size of the model was too small to consider all 100 options.

Further, we report the classification results of the second part of the experiments for the concepts of BP#16 (Table 5), BP#19 (Table 6), BP#29 (Table 7 and BP#36 (Table 8).

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

Table 5: **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 6: **BP#19.** Correctly classified for concepts of BP#19. Models were asked whether the present pinch in the diagram is horizontal or vertical.

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

Table 7: **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 8: **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